SESSION

RECOGNITION METHODS: FACE, GESTURE, OBJECT, CHARACTER

Chair(s)

TBA

Gesture Recognition with the Leap Motion Controller

R. McCartney[1], **J. Yuan**[1], **and H.-P. Bischof**[1,2]
[1]Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA
[2]Center for Computational Relativity and Gravitation, Rochester Institute for Technology, Rochester, NY, USA

Abstract— *The Leap Motion Controller is a small USB device that tracks hand and finger movements using infrared LEDs, allowing users to input gesture commands into an application in place of a mouse or keyboard. This creates the potential for developing a general gesture recognition system in 3D that can be easily set up by laypersons using a simple, commercially available device. To investigate the effectiveness of the Leap Motion controller for hand gesture recognition, we collected data from over 100 participants and then used this data to train a 3D recognition model based on convolutional neural networks, which can recognize 2D projections of the 3D space. This achieved an accuracy rate of 92.4% on held out data. We also describe preliminary work on incorporating time series gesture data using hidden Markov models, with the goal of detecting arbitrary start and stop points for gestures when continuously recording data.*

Keywords: Gesture recognition, CNN, HMM, deep learning

1. Introduction

There was a time when communication with programs like 'vi' [1] was done via keyboard only. The keyboard was used to input data and to change the execution behavior of the program. The keyboard was a sufficient input device for a one-dimensional system.

At the moment that operating systems moved to GUI's the use of a mouse became handy to switch between graphical applications and exert control over them. The first notable applications making use of a mouse came when Microsoft introduced a mouse-compatible Word version in 1983 and when Apple released Macintosh 128 with an updated version of the Lisa mouse in 1984 [2].

Visualizations and games moved parts of computing into 3 dimensional spaces. The visualizations of these 3D worlds are either projected onto the 2D space of your screen or experienced with stereoscopic viewing devices. The control of these 3D worlds is not easy, some would say extremely unnatural, with a 2D mouse. The availability of 3D input devices allows for better control of this 3D world, but requires gesture recognition algorithms in order to use such devices in a natural way. This paper evaluates different gesture recognition algorithms on a novel dataset collected for such purposes.

2. Problem Description

The inputs coming from a mouse or a keyboard are discrete and have limited interpretation. A mouse down event is a single event at a given position on the screen, and dependent on the environment carries information with it about the position of the mouse pointer, time of the click, and so on. A double or triple click is an event over time, and will only count as such if the click events happen within a predefined window. Apple's Magic Mouse [3] somewhat opened the door to 2D gestures, allowing users to swipe between pages or full screen applications and to double tap for access to mission control.

A keyboard or mouse sends an event only if a key or button is pressed or the mouse is moved. They do not start to send events as your hand approaches the device. In contrast, 3D input devices, like the Leap Motion controller[1], start to send frames as soon as they are turned on. These devices send a series of positions in space over time of whatever they detect in their views. The problem becomes to convert the output of these devices into something meaningful.

The output from motion sensing devices comes in two flavors: high-level and low-level. Low-level output is a series of frames where each frame contains information on what the device has sensed, such as the number of fingers, finger tip positions, palm position and direction, etc. The frame rate depends on the user settings and compute power, but 60 or more frames per second is typical. High-level output is the interpreted version of the raw frame data. This allows users or application developers to be informed when a particular predefined gesture is recognized.

We are interested in gesture recognition algorithms. Therefore, we are interested in the low-level information in order to interpret this into high-level information for others. The next section will describe the device we have used as our sensor, a relatively new and inexpensive motion sensing device. Then, we will discuss the gestures used and the dataset we captured for such purposes. Following that, we will discuss the particular form of dimensionality reduction and normalization we used on this data. The last sections will discuss the different gesture recognition algorithms we used as well as their results.

[1]https://www.leapmotion.com/

3. Leap Motion Device

There are many motion sensing devices available in the marketplace. The Leap Motion controller was chosen for this project because of its accuracy and low price. Unlike the Kinect, which is a full body sensing device, the Leap Motion controller specifically captures the movements of a human hand, albeit using similar IR camera technology.

The Leap Motion controller is a very small ($1.2 \times 3 \times 7.6cm$) USB device [4]. It tracks the position of objects in a space roughly the size of the top half of a 1m beach ball through the reflection of IR light from LEDs. The API allows access to the 'raw' data, which facilitates the implementation of gesture recognition algorithms. A summary of the specifications of the API: Language support for Java, Python, JavaScript, Objective C, C# and C++; data is captured from the device up to 215 frames per second; the precision of the sensor is up to 0.01mm in the perception range of 1 cubic feet, giving it the ability to identify 7×10^9 unique points in its viewing area.

The SDKv2 introduced a skeletal model for the human hand. It supports queries such as the five finger positions in 3 dimensional space, open hand rotation, hand grabbing, pinch strength, and so on. The SDK also gives access to the raw data it sees. Here we use this device to implement and analyze different gesture recognition algorithms from a dataset collected by this API.

4. Previous Work

One commonly used method of recognition involves analyzing the path traced by a gesture as a time series of discrete observations, and recognizing these time series in a hidden Markov model [5]. Typically, the discrete states are a set of unit vectors equally spaced in 2D or 3D, and the direction of movement of the recorded object between every two consecutive frames is matched to the closest of these state vectors, generating a sequence of discrete directions of movement for each gesture path [6], [7], [8]. Hidden Markov models have also been used to develop online recognition systems, which record information continuously and determine the start and stop point of a gesture as it collects data in real time [8], [9].

Another class of methods for recognition of dynamic gestures involves the use of finite state machines to represent gestures [10], [11]. Each gesture can be represented as a series of states that represent regions in space where the recorded object may be located. The features of these states, such their centroid and covariance, can be learned from training data using methods such as k-means clustering. When evaluating a new gesture, as the recorded object travels through the regions specified by these states, these sequences of states are fed into finite state machines representing each of the trained gestures. In this way, gestures whose models are consistent with the input state sequences are identified.

Neural networks have typically been used to recognize static gestures, but recurrent neural networks have also been used to model gestures over time [12], [13]. One of the main advantages of this type of model is that multiple inputs from different sources can be fed into a single network, such as positions for different fingers, as well as angles [13]. Additionally, convolutional neural networks and deep learning models have been used with great success to recognize offline handwriting characters [14], which can be considered analogous to hand gestures under certain representations as shown in this paper. A similar problem domain of gesture recognition, although in a lower dimensional space, is that of handwritten text recognition, where long-short term memory networks are the current state of the art [15], [16], [17].

5. Dataset

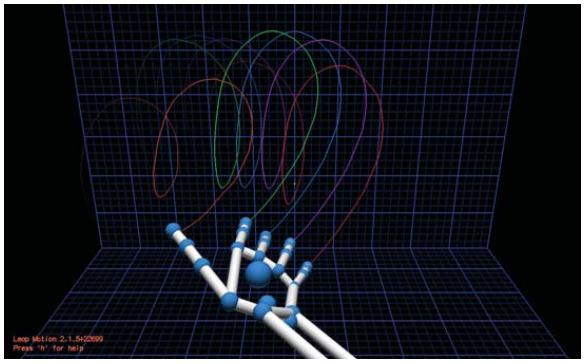


Fig. 1: Leap Motion Visualizer

In order to examine various machine learning algorithms on gestures generated through the Leap Motion controller, we needed to have a dataset that captured some prototypical gestures. To this end, a simple GUI was created that gave users instructions on how to perform each of a chosen set of 12 hand gestures and provided visual feedback to the participant when the system was in the recording stage. All gestures were performed by holding down the 's' key with the non-dominant hand to record and then using the primary hand to execute the gesture at a distance 6" to 12" above the top face of the controller. The code for this capture program is located online[2].

Students and staff on the RIT campus used the GUI to record their versions of each of 12 gesture types: one finger tap, two finger tap, swipe, wipe, grab, release, pinch, check mark, figure 8, lower case 'e', capital 'E', and capital 'F'. The one and two finger taps were vertical downward movements, performed as if tapping a single or set of keys on an imaginary keyboard. The swipe was a single left to right movement with the palm open and facing downwards, while the wipe was the same movement performed back and forth

[2]https://github.com/rcmccartney/DataCollector

several times. The grab motion went from a palm open to a closed fist position, while the release was performed in the opposite direction. Pinch was performed with the thumb and forefinger going from open and separated to touching. The check mark was performed by pointing just the index finger straight out parallel to the Z axis, then moving the hand in a check motion while traveling primarily in the X-Y plane. The figure 8, lower case 'e', capital 'E', and capital 'F' were all similarly performed by the index finger alone, in the visual pattern indicated by their name in the plane directly above the Leap Motion controller. The native Leap Motion Visualizer shown in Figure 1 was available for each subject to use alongside of our collection GUI while performing the gestures if so desired, providing detailed visual feedback of the user's hand during motion.

As each gesture was performed, the Leap Motion API was queried for detailed data that was then appended to the current gesture file. The data was captured at over 100 frames per second, and included information for the hand such as palm width, position, palm normal, pitch, roll, and yaw. Positions for the arm and wrist were also captured. For each finger 15 different features were collected, such as position, length, width, and direction. In all, we collected 116 features for each frame of the recording, with the typical gesture lasting around 100 to 200 frames, although this average varies greatly by gesture class. Files for each gesture are arranged in top-level folders by gesture type, inside which each participant in the study has an anonymous numbered folder that contains all of their gesture instances for that class. Typically, each user contributed 5 to 10 separate files per gesture class to the dataset, depending on the number of iterations each participant performed.

In all, approximately 9,600 gesture instances were collected from over 100 members of the RIT campus, with the full dataset totaling around 1.5 GB. The data is hosted online for public download[3]. Individual characteristics of each gesture vary widely, such as stroke lengths, angles, sizes, and positions within the controller's field of view. Some users had used the Leap Motion before or were comfortable performing gestures quickly after starting, while others struggled with the basic coordination required to execute the hand movements. Thus, there is considerable variation within a gesture class, and identifying a particular gesture performed given the features captured from the Leap Motion device is not a trivial pattern recognition task.

6. Image Creation

In its raw form, the varying temporal length of each gesture and large number of features make it difficult to apply traditional machine learning techniques to this dataset. Thus, a form of dimensionality reduction and normalization is needed for any learning technique to be effectively applied. For the convolutional neural network (CNN) that we employ in Section 7, this dimensionality reduction took the form of converting each instance of real-valued, variable length readings into a fixed-size image representation of the gesture.

Fig. 2: One instance example of each of the gestures used for the CNN experiment

Fig. 3: The mean image of the dataset on the left and the standard deviation on the right used for normalization

CNNs traditionally operate on image data, using alternating feature maps and pooling layers to capture equivariant activations in different locations of the input image. Due to the complex variations that are nevertheless recognizable to a human observer as a properly performed gesture, CNNs offer a way to allow for differences in translation, scaling, and skew in the path taken by an individual's unique version of the gesture. To transform each gesture into constant-sized input for the convolutional network, we created motion images on a black canvas using just the 3 dimensional position data of the index finger over the lifetime of the gesture. That is, for each frame we took the positions reported in the Leap coordinate axes, which varies approximately from -200 to 200 in X and Z and 0 to 400 in Y, and transformed those coordinates into pixel space varying from 0 to 200 in three different planes, XY, YZ, and XZ. For each reported position, the pixels in the $5x5$ surrounding region centered on the position were activated in a binary fashion. From this

[3]http://spiegel.cs.rit.edu/~hpb/LeapMotion/

point, each of the three coordinate planes are separately or jointly able to be used as image input data in the learning model. However, for this first experiment on the dataset we kept only the XY plane of index finger positions and concentrated on those gestures that mainly varied in that plane, as explained below.

Despite being equivariant across feature maps, CNNs still have some difficulty in classification over widely varying positions and orientations of input activations. Thus, we cropped each image to fit the minimum and maximum indices of nonzero activations, and then sampled the resulting pixels to resize each image to a constant $50x50$ input size. After resizing, the pixel activations were normalized by subtracting the mean and standard deviation for that pixel across the entire training set, rather than using the statistics within a single image. Note that using only the XY positions of the index figure is a significant simplification of the data contained in an instance of the Leap dataset, but it served to show the applicability of computer vision techniques to the task of gesture recognition. As a result of this, we kept only those gestures that varied in the XY planes for the CNN experiments, namely the check mark, lower-case 'e', capital 'E', capital 'F', and figure 8. Since this subset of the gestures are guided by the index finger in the XY plane they appear rather well-formed there, but appear as mostly noise in the other two planes as their appearance in those projections largely depends upon unconscious movements of the hand. Expanding this representation to all three planes of movement for all gesture classes should be sufficiently expressive to broaden the learning algorithm to the entire dataset, and will be explored in future work. An example of each of the gesture classes in this representation after preprocessing is shown in Figure 2. Figure 3 shows the normalization factors used for the dataset, with the mean image on the left and the standard deviation on the right.

Note that there are other possibilities for generation of the images here that we did not do, such as removing skew and including motion history into gray-scale representations. There are still present many forms of variation in the input activations that are inherent to the users, such as the left-handed version of the check mark shown in Figure 4. While such differences as this and other examples of allowable variance in gestures from a given class are easily and unconsciously accounted for by humans, for instance by two people conversing in American Sign Language, they pose a significant challenge to the classification models that we discuss further in Section 7 and must be accounted for when training such classifiers.

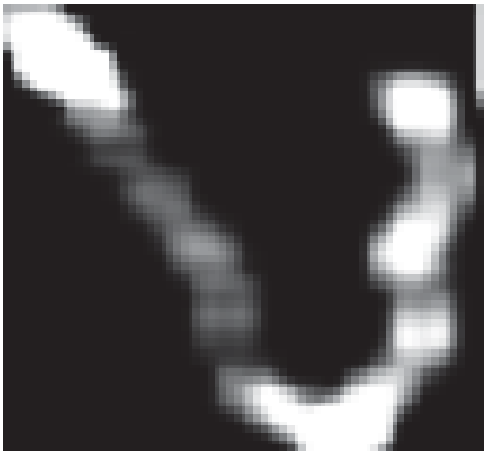


Fig. 4: A left-handed check mark after cropping, sampling, and normalization

7. Models

We have chosen our initial experiments on this dataset using two diverse models for classification of temporal sequences. The first is to convert the data into a fixed image representation as discussed above and use a CNN for classification. The second is to use a hidden Markov model to aid in a time series recognition task.

7.1 Convolutional neural network

Convolutional neural networks are powerful models in computer vision due to their ability to recognize patterns in input images despite differences in translation, skew, and perspective [14], [18], [19], [20]. They can be effective at finding highly complex and nonlinear associations in a dataset. They do so in the context of supervised learning, by allowing the model to update parameters dynamically so as to minimize a cost function between a target value and the observed output of the model. An advantage they have over traditional, fully-connected neural networks is that the learned feature maps are applied with the same parameters to an entire image, drastically reducing the number of parameters required to learn without seriously degrading the capacity of the model [14]. This allows for more complex and deeper architectures to be employed without as serious a risk of overfitting the training data.

Human gestures are highly complex, nonlinear, and context-dependent forms of communication with both considerable overlap and great divergence between gesture types. People often perform the same gesture class in highly unique and differing ways, yet to the human brain these are easily recognized as constituting the same meaning. Further, very subtle and small differences exist between gestures that impart greatly differing meanings to the separate classes, yet such differences are not easily defined or separated. Given this type of data, convolutional neural networks have the advantage of learning good features as part of the classification task itself. Thus, we do not need to handcraft features of each valid gesture but allow the model to learn them as a product of minimizing the loss function. The model can thus learn to classify gestures based off of the

		Truth				
		E	✓	e	F	8
Prediction	E	28	0	1	0	0
	✓	1	62	2	0	1
	e	2	0	30	1	4
	F	1	2	1	40	0
	8	1	0	1	0	34

Table 1: Confusion matrix for CNN without dropout

		Truth				
		E	✓	e	F	8
Prediction	E	28	0	0	0	1
	✓	0	63	1	0	1
	e	2	0	30	0	3
	F	2	1	1	41	0
	8	1	0	3	0	34

Table 2: Confusion matrix for CNN with dropout

complex interactions between learned features that may not otherwise be easily discerned or discovered.

The convolutional neural networks used in these experiments came from MatConvNet, a toolbox for Matlab developed in the Oxford Visual Geometry Group [21]. All experiments were run on a GeForce GTX 960 GPU, with 1024 CUDA cores and 2 GB memory. In addition, NVIDIA's CUDA Deep Neural Network library (cuDNN)[4] was installed as the convolution primitives inside the MatConvNet library. The network consisted of alternating convolutions and max pooling layers, as depicted in Figure 5, followed by two layers of a fully-connected neural network with a softmax output. All neurons were rectified linear units, as they can be trained faster than their sigmoid counterparts [18]. The model was trained both with and without dropout, following the techniques described in [22], [23], [24]. See Tables 1 and 2 for the results of training this network with the 5 input gesture classes. The code for this network and for image creation is hosted online[5]. Overall the network produced a 92.5% recognition rate on held-out data after training to perfectly fit the input data, with a very modest improvement seen from using dropout with a rate of 50% on the two fully connected layers. This modesty may be due to the fact that dropout was not applied to the convolutional layers, which in the future could lead to greater improvements in generalization. A few of the misclassified gesture image representations can be seen in Figure 6.

[4]https://developer.nvidia.com/cuDNN

[5]https://github.com/rcmccartney/LeapDeepLearning

7.2 Time series recognition with HMMs

Though the convolutional neural network performs well on images of the whole gesture, it does not take into account temporal information such as the order in which the strokes are performed. This can be addressed by modeling individual or groups of points as discrete states in a hidden Markov model. However, one of the principle challenges of definition and recognition of arbitrary gestures in 3D space is the high variability of gestures within the sequence space. For example, many traditional dynamic gesture recognition models have used translations between pairs of consecutive frames to generate a sequence of observations by fitting the translations to the closest of a set of evenly-distributed discrete vectors [6]. These methods work well in 2D, but suffer in 3D because 3D motions tend to be more varied and uncontrolled. Any portion of a single curved motion may be represented by slightly different vector sequences, and these sequences may result in highly distinct observations sequences even though they represented the same intended movement.

To solve this high-variability problem, we propose a method to process a sequence of frames of positional data and summarize them to a shorter and more generalized sequence of lines and curves, which are then fed into a hidden Markov model as discrete states. This method involves first identifying line segments in the sequence of frames by calculating average vectors of consecutive points âĂŞ from the sequence of average vectors those within a minimum angle distance are combined into one growing line segment. This line segment is then fit to one of 18 discrete observation states represented by vectors pointing away from the origin distributed equally in 3D space. Next those sequences of points that do not satisfy the criteria above but are of some minimum length of frames are likely curved segments. These sequences of points are fit to a sphere using a least squares approximation method [25]. The sphere then defines the centroid of the curve's rotation. To discretize the curve, the normal vector of the rotation is found by taking the cross product of the vectors emanating from the discovered centroid to the two end points of the curve. This normal vector is then fit to a set of six state vectors (clockwise and counterclockwise for each of roll, pitch, and yaw). The sequence of discovered lines and points then serves as the observation sequence, which is much shorter and more invariant to individual differences between training examples.

The performance of this model was relatively poor, at around 50% recognition for the specified gesture set. We believe this time series model is less robust to sources of error in the data, specifically the combination of very small and very large drawn gesture examples, as well as examples containing large disjointed spaces between consecutive segments of points due to sampling or user error. We hope to address these errors in the future by experimenting with

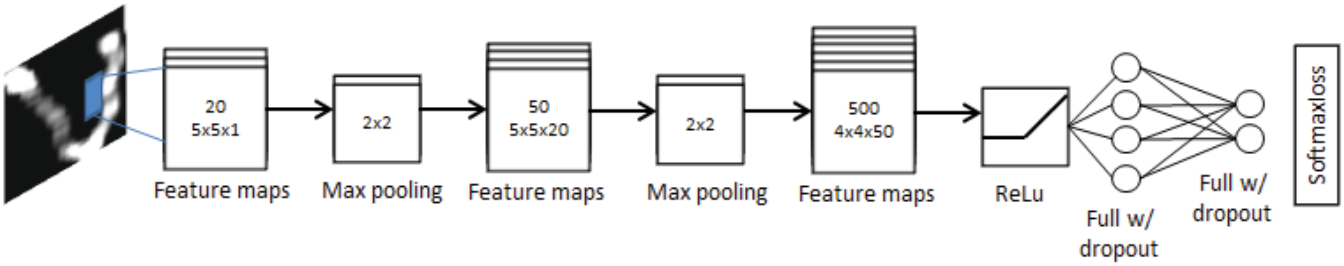


Fig. 5: A depiction of the CNN topology used

Fig. 6: Examples of misclassified gestures

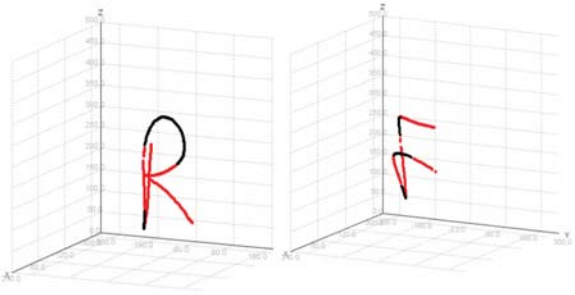


Fig. 8: Example gestures described as sequences of lines (black) and curves (red)

rescaling and re-sampling training gesture paths.

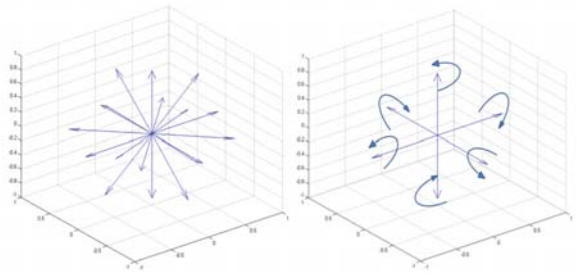


Fig. 7: The set of discretized states describing the motion of consecutive groups of 3D points, including 18 line directions and 6 curve directions

8. Future Work

This experiment represents the first to use the novel dataset collected from the Leap Motion controller. There is still much to be explored with this dataset as well as with applying other forms of learning algorithms to our representations of the gestures. Different forms of dimensionality reduction, such as PCA or gradient-based methods, could be used to help deal with the large amount of features available per gesture instance. Recurrent neural networks-long short term memory models in particular-could prove effective at dealing with the varying temporal nature of human gestures. Future work will also expand the scope to encompass the segmentation task as well as the classification task. One particularly interesting avenue of research is in combining the models discussed in Section 7 into a single online recognition engine. The HMM could specialize in segmenting gestures as they occur, using the two hidden states of “in-gesture" and “between-gesture" to distinguish between when a human hand is trying to semantically communicate or just resting. Once segmented, the frames of data from the “in-gesture" state could then be sent to the CNN model for classification. Note that the requirement to segment actual communication from idling is not an issue when using other input devices such as a mouse, and arises here due to the inability to set these 3D devices into non-recording states.

9. Conclusions

The Leap Motion controller is a promising device for enabling user-friendly gesture recognition services. Based on our results, the data generated by this device can be accurately classified by representing its 3D gesture paths as sets of 2D image projections, which can then be classified by convolutional neural networks. Here we limited the classification results to gestures performed in the XY plane,

but the model can be extended to give equal consideration to all 3 planes of 2D projections, allowing for a wide variety of gesture representations. Despite its good performance, one of the limitations of this model is that it cannot provide online recognition of gestures in real time. As future work we look to incorporate an alternative model, such as a hidden Markov model, as a segmentation method to determine likely start and stop points for each gesture, and then input the identified frames of data into the CNN model for gesture classification.

References

[1] W. Joy and M. Horton. (1977) An introduction to display editing with vi. [Online]. Available: http://www.ele.uri.edu/faculty/vetter/Other-stuff/vi/vi-intro.pdf

[2] A. S.-K. Pang, "The making of the mouse," *American Heritage of Invention and Technology*, vol. 17, no. 3, pp. 48–54, 2002.

[3] R. Loyola, "Apple's magic mouse offers multitouch features," p. 65, 01 2010. [Online]. Available: http://search.proquest.com.ezproxy.rit.edu/docview/231461266?accountid=108

[4] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013. [Online]. Available: http://www.mdpi.com/1424-8220/13/5/6380

[5] L. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[6] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec 2008, pp. 1–4.

[7] T. Schlömer, B. Poppinga, N. Henze, and S. Boll, "Gesture recognition with a wii controller," in *Proceedings of the 2nd international conference on Tangible and embedded interaction*. ACM, 2008, pp. 11–14.

[8] H.-K. Lee and J.-H. Kim, "An hmm-based threshold model approach for gesture recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 961–973, 1999.

[9] S. Eickeler, A. Kosmala, and G. Rigoll, "Hidden markov model based continuous online gesture recognition," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 2. IEEE, 1998, pp. 1206–1208.

[10] P. Hong, M. Turk, and T. S. Huang, "Gesture modeling and recognition using finite state machines," in *Automatic face and gesture recognition, 2000. proceedings. fourth ieee international conference on*. IEEE, 2000, pp. 410–415.

[11] R. Verma and A. Dev, "Vision based hand gesture recognition using finite state machines and fuzzy logic," in *Ultra Modern Telecommunications & Workshops, 2009. ICUMT'09. International Conference on*. IEEE, 2009, pp. 1–6.

[12] H. Hasan and S. Abdul-Kareem, "Static hand gesture recognition using neural networks," *Artificial Intelligence Review*, vol. 41, no. 2, pp. 147–181, 2014.

[13] K. Murakami and H. Taguchi, "Gesture recognition using recurrent neural networks," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1991, pp. 237–242.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 5, pp. 855–868, 2009.

[16] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 545–552.

[17] A. Graves, "Offline arabic handwriting recognition with multidimensional recurrent neural networks," in *Guide to OCR for Arabic Scripts*. Springer, 2012, pp. 297–313.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[21] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.

[22] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[23] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8609–8613.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[25] D. Eberly. (2015) Least squares fitting of data. Geometric Tools, LLC.

Designing a Lightweight Gesture Recognizer Based on the Kinect Version 2

Leonidas Deligiannidis
Wentworth Institute of Technology
Dept. of Computer Science and Networking
550 Huntington Av.
Boston, MA 02115
USA
deligiannidisl@wit.edu

Hamid R. Arabnia
University of Georgia
Dept. of Computer Science
415 GSRC
Athens, GA 30602
USA
hra@cs.uga.edu

Abstract - *We present a lightweight gesture recognizer utilizing Microsoft's Kinect version 2 sensor. Our recognizer seems to be robust enough for many applications and does not require any training. Because this version of the Kinect sensor is equipped with a higher resolution than its predecessor depth camera, it can track some of the user's fingers, and the Kinect SDK is able to provide state information of the user's hands. Armed with this capability, we were able to design our gesture recognizer. New gestures can be specified programmatically at the moment, but we are also working on a graphical user interface (GUI) that would allow a user to define new gestures with it. We show how we built the recognizer and demonstrate its usage via two applications we designed. The first application is a simple picture manipulation application. For the second application, we designed a 3DOF robotic arm that can be controlled using gestures.*

Keywords: Kinect 2, Gesture Recognition.

1. Introduction

In late 2010, Microsoft Corporation introduced the first version of a gaming device, the Kinect, which could be used along with its Xbox gaming console. Recently, Microsoft released the second version of the Kinect [1], for their new Xbox One console, which is faster and provides higher resolution video and depth feeds. The Kinect is a motion sensing input device and can connect to a PC via a Universal Serial Bus (USB) adapter. The Kinect sensor consists of a video camera and a depth camera. The depth camera provides depth information for each pixel using an infrared-IR projector and an IR camera. It also has a multi-array microphone that can detect the direction of where spoken commands are issued. The primary purpose of the Kinect sensor is to enable game-players interact and play games without the need of holding a physical game controller. This innovation changed the way with which we play and interact with games. Players can now use natural command such as tilting to the left / right, raising their hands, jumping, etc. to issue commands. The Kinect sensor enables the players to do this by continuously tracking their body movements, tracking their gestures, as well as observing for verbal commands.

These capabilities that it offers, along with its low and affordable price, made the Kinect sensor an attractive device to researchers. Using the freely available SDK [1] for the Kinect, we can now design programs that incorporate the functionalities of the Kinect sensor in our research. The fact that the device does not need to be trained or calibrated to be used, make it easy and simple to be used in many environments other than the environments it was originally designed for. For example, in [2] the functionality of the sensor has been extended to detect and recognize objects and obstacles so that visual impaired people can avoid them. Because of its contactless nature of interaction [3], the Kinect found its way into operating rooms where non-sterilize-able devices cannot be used [4]. Its depth camera can be used to scan and construct 3D maps and object [5]. An effective techniques to control applications such as Google Earth and Bing Maps 3D utilizing a small, yet easy to remember, set of hand gestures is illustrated in [6]. The robotics community [7] adopted the Kinect sensor so that users can interact with robots in a more natural way. Some applications require the tracking of the fingers [8], which was not supported by the original sensor [9][10][11] but it is now (with the second generation of the sensor), in a limited way however.

There are several methods for detecting and tracking fingers, but this is a hard problem mainly because of the resolution of the depth sensor; this is also true for the second generation of the camera. Some methods work well such as [12], as long as the orientation of the hands does not vary. Other techniques to work require specialized instruments and arrangements such as infrared camera [13], stereo camera [14], a fixed background [15], and track-able markers on hands and fingers [16]. Other systems need a training phase [17] to recognize gestures such as "clapping", "waving", "shaking head", etc.

2. Gestures

Interacting in an application utilizing the Kinect sensor requires the sensor to actively track the motion of the user. Even though playing a game could require only large movements of the user's body, limbs, or hands, to interpret the movements as commands such as jumping, leaning, waving, ducking, etc. other applications could require more precise input. Specifically, an application should be able to classify a posture as an event as well as a gesture and should be able to differentiate between the two. A gesture normally has a beginning and an end. A posture is a static positioning of the user and her arms, legs, etc. where as a gesture is dynamic by nature. A user should be able to indicate the beginning and possibly the ending of a gesture. A gesture, such as waving, is a dynamic motion of one's hand(s) but doesn't have to be precise. Other gestures need to be more precise as the user's arm, for example, is being tracked to control the movement of a remote robotic arm, but more importantly the initiation and the termination of the gesture could be of greater importance. A system that actively tracks the movements of a user and miss-interprets actions of a user's intent as actual commands will soon be abandoned by the user as it confuses her. Initiating the termination of a gesture is very valuable too, as it can be used to cancel or stop the current tracking state. As it is reported in [18] the main difficulties in designing a gesture recognizer has to address the issues of "temporal segmentation ambiguity" which deals with the beginning and the ending of a gesture, and "spatial-temporal variability" which deals with the tolerance of the initiation and termination of a gesture since each person performs the same gestures differently.

The Kinect's version 1 depth sensor was low resolution to a point where finger detection was difficult to make, and impossible to make from a distance. The Kinect 2 has a higher resolution depth camera, and finger detection is provided by the SDK. Finger detection is still limited but at least the SDK reports postures of the hand based on the fingers' arrangement. For example, the Kinect 2 can distinguish, in any orientation, if the hand is Open, Closed, or Lasso. Lasso is defined by closing the hand and extending the index finger (like pointing to an object). However, because of the still-low resolution of the depth sensor, it is recommended that the user extends both the index and the middle fingers while touching each other to indicate the Lasso posture. If the user is close to the camera, extending the index finger alone is enough. Having two hands, where each hand can perform 3 different postures, we have 9 different combinations we can use to indicate the beginning and ending of a gesture. Additional postures can be defined by, for example, hiding your hand behind your back while performing a posture with your other hand; the Kinect SDK provides tracking state information for each joint, in addition to its location and orientation in space. Depending on the posture and the gesture, the user must be aware of the position of the Kinect sensor. For example, if the user performs the Lasso posture and points at the Kinect, it is possible that the Kinect will report false posture as the Lasso gesture seen from the front looks very similar to the Closed hand-posture.

3. Gesture Recognizer

The Kinect SDK provides an API where the skeleton information of up to 6 people can be reported. It tracks a human body, even partially when some joints are hidden, and reports the position and orientation of each of the 25 joints. It also reports if a joint information is accurate or not; it is being tracked or it is not visible in the current frame and its value is inferred.

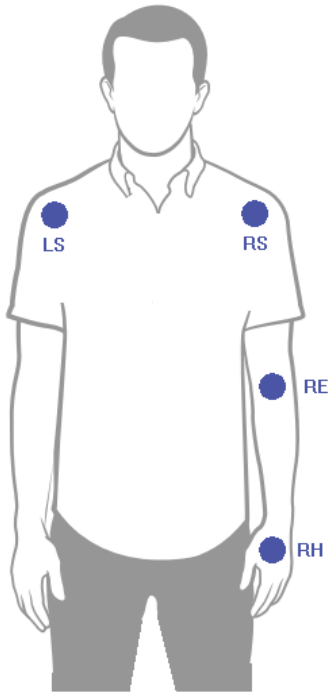


Figure 1. The four joints needed by the gesture recognizer to define the gestures involving the right hand.

In Kinect 2, which has a higher resolution depth sensor, the hands states are also reported; the main states are Open, Closed and Lasso – open palm, fist, fist with index finger extending. Based on the upper body joint positions and the state of the hands we designed a gesture recognizer engine. The engine takes as input the joint information from the Kinect and determines which gesture / posture is being performed. The main advantages of this recognizer are that it is lightweight, rotation invariant, does not require any training, and in addition, it can be configured for many different gestures. The configuration, however, is done programmatically at this time but we are developing a graphical user interface tool where no programming will be needed to define new gestures.

Based on only a few joints, we divide the user space into five areas. Figure 1 shows the four joints needed to recognize gestures for the right hand: the left and right shoulder, the right elbow and the right hand. The left elbow and the left hand are needed for the gestures involving the left hand, but for simplicity reasons we only show here the joints involved in the gesture recognizer for the right hand.

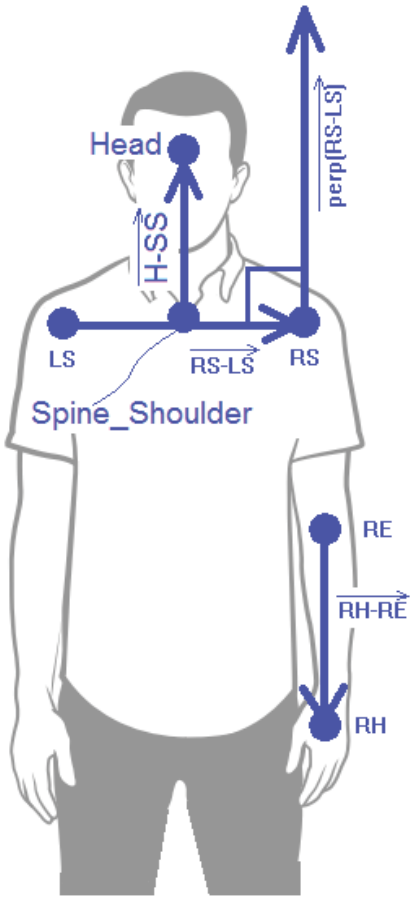


Figure 2. Calculating the 3 vectors needed for the recognizer. A vector defining the shoulder line, another vector that is perpendicular to the vector that defines the shoulder line, and another vector that defines the elbow-hand. The H-SS vector is only used in the robotic arm application we will discuss later.

If we treat the positions of these joints as vectors, we can define a vector RS-LS as shown in figure 2. Then we define a vector that is perpendicular to RS-LS, shown as perp(RS-LS). We can also calculate the vector RH-RE which is the vector defined be the right elbow and right hand. The "Head" and "Spine_Shoulder" joints are only used to control the roll of the robotic arm application that we will discuss later.

Using the dot product operation of vectors, we can calculate in which area the right hand is located, as shown in figure 3. The hand can be in 3 different areas: a) above the shoulder line and to the left, b) above the shoulder line and to the right, and c) below the shoulder line and to the right.

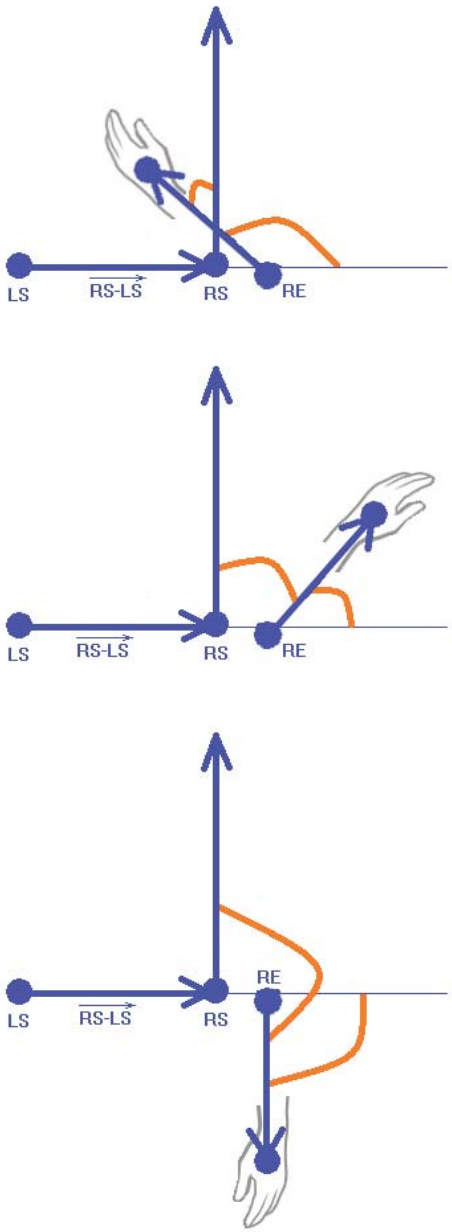


Figure 3. Using the dot product of vectors, we can calculate in which area the right hand is.

Two vector subtractions are needed to calculate the shoulder line and the elbow-hand vectors. Based on the shoulder line vector, we can easily construct a perpendicular to it vector as well. Then, two dot product operations are needed to determine in which area the hand is. Figure 4 shown the five areas we define for both hands. Even though a user could move his right hand to the area

defined for the left hand, we don't consider motions like these as valid, as these movements obstruct the view of the user and they are anatomically awkward to perform. Using this technique, one can define other areas for tracking hands, such as using the waist line, or the spine line, etc. depending on the applications needs.

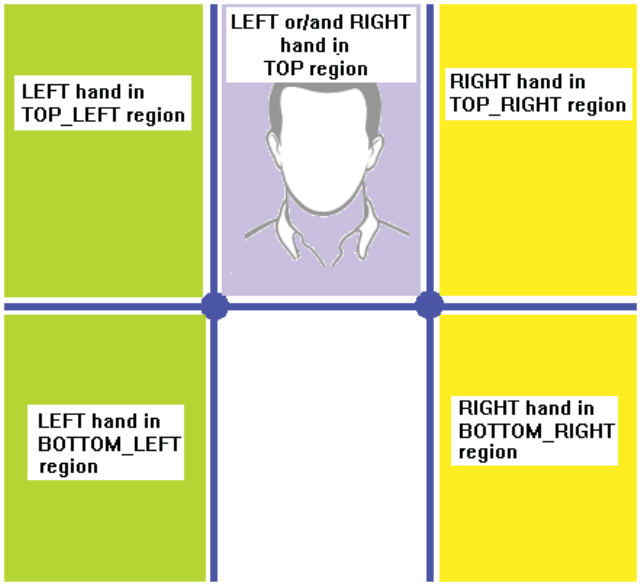


Figure 4. The five areas defined by the shoulder line and the two hands.

4. Picture Control Application

The first application we designed based on our gesture recognizer, was a picture manipulation application. The application is designed in Java. Using the Java Native Interface (JNI), we call our C++ compiled functions that communicate with the Kinect and deliver the joint information and the hands states to our java gesture recognizer. As shown in figure 5, the user moves his hand near his head and closes his hands to grab a picture, and then pulls apart his hands to increase the size of the picture. Opening his hands, stops the current operation. Similarly, if the user grabs the picture with his hands apart (by closing his hands) and moves them close to each other, the size of the picture decreases; this operation is similar to what most users are familiar with on mobile devices but instead of using their hands, they use two finger. The last gesture is used to rotate an image. To activate and control the rotation of the picture, the user's left hand moves close to the body in the open posture, and the right hand performs the lasso posture. The orientation of the picture is controlled by the right hand's continues rotations.

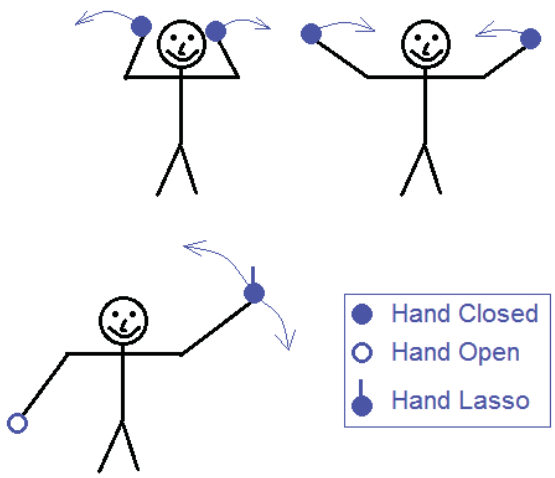


Figure 5. The first gesture is used to increase the size of the picture, the second gesture is used to decrease the size of the picture, and the last gesture is used to rotate the selected picture.

5. Robotic Arm Control

We designed a second application that uses our gesture recognizer which controls a robotic arm. Figure 6 shows a top view of the robot arm which consists of three heavy duty servo motors, a USB servo controller from Phidgets.com and a 5V / 5A power supply to power the servo motors. The servo motors are physically connected to each other to provide a three-degrees-of-freedom of the arm, shown in figure 7. Figure 7 also shows how the robotic arm is connected to a PC and the Kinect sensor. The sensor is connected to a PC via a proprietary Kinect-to-USB adapter. Via another USB port, the PC is connected to the servo-controller of the robotic arm. The application receives joint and hand state information from the Kinect, the gesture recognizer component interprets these commands and instructs the servo-controller to rotate the appropriate servo motors by a specified amount.

Figure 8 shows the gestures implemented to control the robotic arm. The top two figures (in figure 8), are used to disengage and engage the servo motors respectively. The second set of gestures are used to control the bottom servos to make the arm rotate right and left respectively. The third set of gesture, are used to control the top servo motor and move the arm up and down. The last gesture is used to instruct the arm to follow the user's right hand. As the user moves his arm left-right and up-down, the robotic arm mimics these movements by controlling the three servo motors simultaneously and in real time. By leaning the head left and right, we change the "roll" of the arm ±10 degrees. As shown in figure 2, we construct the H-SS vector. By taking the dot product of the H-SS and the RS-LS vectors, we can determine the direction and the amount

of the roll (which is implemented by rotating the middle servo motor).

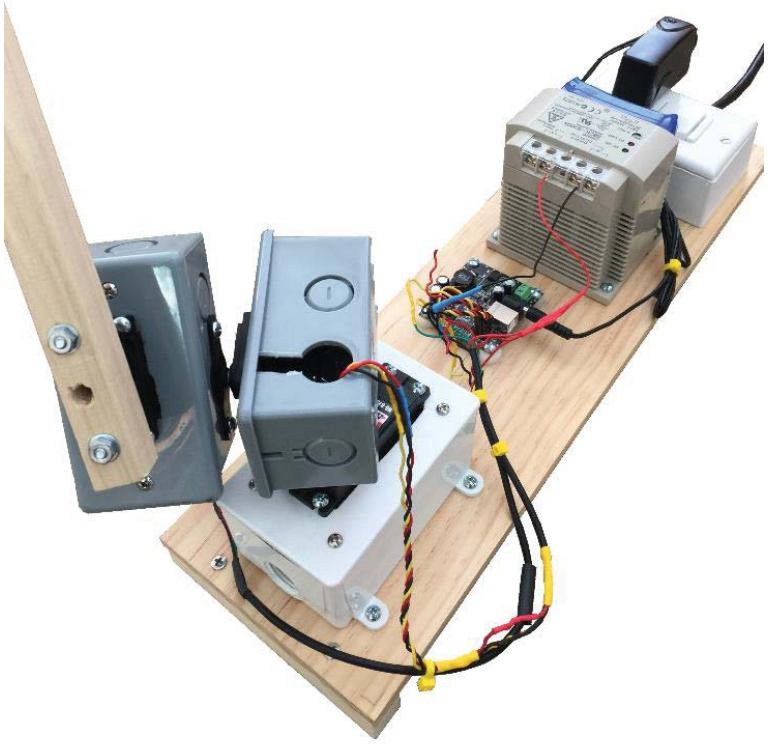


Figure 6. The robotic arm (top view), showing the three servo motors attach to each other to provide 3-degrees-of- freedom. Next to the servo assembly is the Phidgets servo controller which receives its commands via its USB port. At the other end, the 5V / 5A power supply is shown which provides power to the three servo motors.

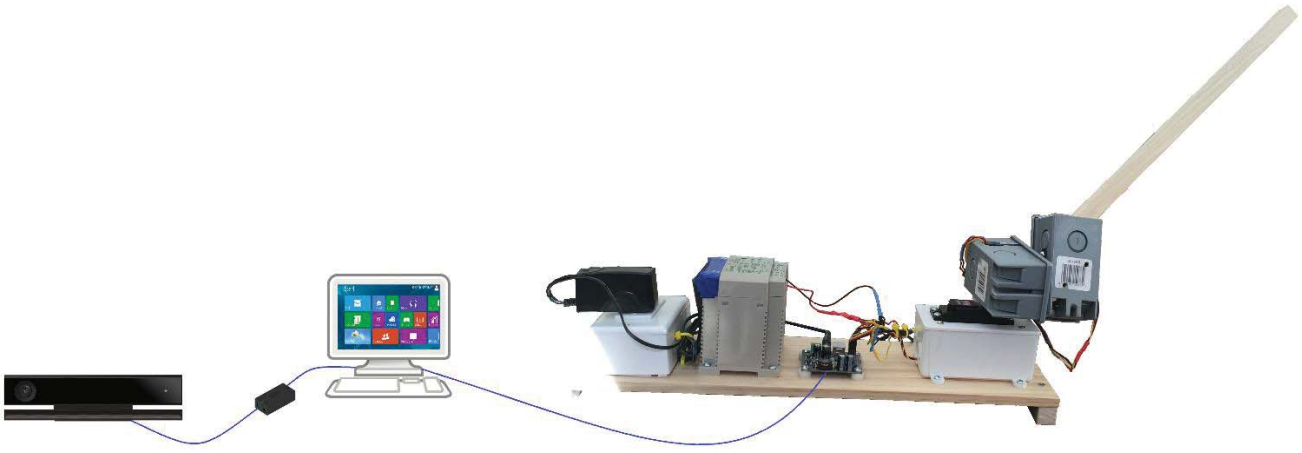


Figure 7. The robotic arm (side view), and how it is connected to the controlling PC and the Kinect camera. The Kinect camera is connected to the PC via a proprietary adapter. There is also a USB connection between the PC and the robotic arm's servo controller.

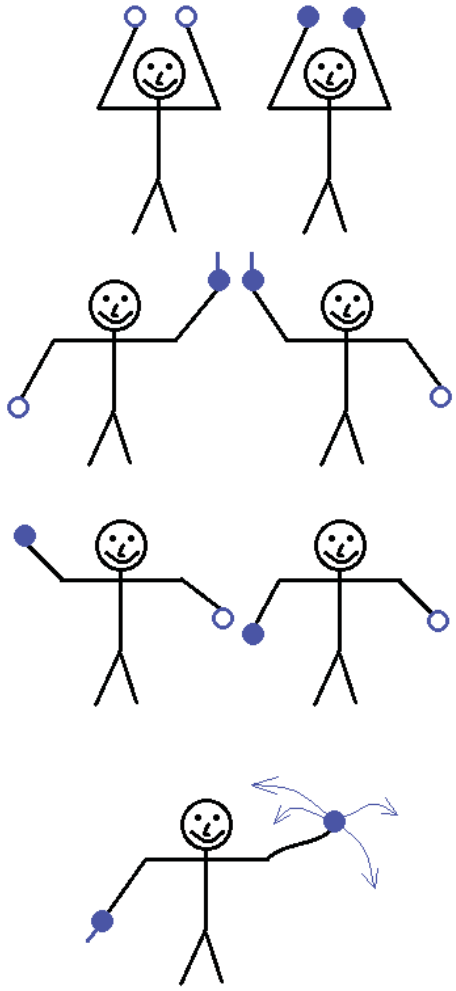


Figure 8. The gestures used to control the robotic arm. The top two gestures are used to disengage and engage the servo motors, respectively. The next set of gestures are used to rotate the arm left-right. The next set of gestures are used to move the tip of the arm up-down. The last gesture at the bottom, is used to allow the robotic arm to follow the user's right hand. As the user moves his hand, the robotic arm mimics these movements in real time.

6. Conclusion

Skeleton tracking with joint position and hand state information from the Kinect version 2 sensor can be very useful input to a gesture recognizer. Having a gesture recognizer, we can interact with software applications and other hardware devices without using a tangible controller. We illustrated our gesture recognizer in this paper by presenting a couple of applications utilizing it. Because this new version of Kinect reports hand state information, we can design many different applications that require gestures. We wish to develop a graphical interface where one would be able to define gestures and associated actions via a GUI instead of doing the same programmatically.

7. References

[1] Microsoft Corporation's Kinect version 2 home page. "http://www.microsoft.com/en-us/kinectforwindows/" Retrieved March 2015.

[2] Atif Khan, Febin Moideen, Juan Lopez, Wai L. Khoo and Zhigang Zhu. "KinDectect: Kinect Detecting Objects". K. Miesenberger et al. (Eds.): Computers Helping People with Special Needs, Lecture Notes in Computer Science (LNCS) Volume 7383, 2012, pp 588-595, Springer-Verlag Berlin Heidelberg 2012.

[3] K. Montgomery, M. Stephanides, S. Schendel, and M. Ross. User interface paradigms for patient-specific surgical planning: lessons learned over a decade of research. Computerized Medical Imaging and Graphics, 29(5):203–222, 2005.

[4] Luigi Gallo, Alessio Pierluigi Placitelli, Mario Ciampi "Controller-free exploration of medical image data: experiencing the Kinect". 24th International Symposium on Computer-Based Medical Systems (CBMS), June 30-27 2011 p1-6.

[5] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, Dieter Fox. "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments". The International Journal of Robotics Research 0(0) 1–17, March 14 2012.

[6] Maged N Kamel Boulos, Bryan J Blanchard, Cory Walker, Julio Montero, Aalap Tripathy, Ricardo Gutierrez-Osuna. "Web GIS in practice X: a Microsoft Kinect natural user interface for Google Earth navigation". International Journal of Health Geographics 2011, 10:45.

[7] Wei-Chen Chiu, Ulf Blanke, Mario Fritz, "Improving the Kinect by Cross-Modal Stereo", In Jesse Hoey, Stephen McKenna and Emanuele Trucco, Proceedings of the British Machine Vision Conference, pages 116.1-116.10. BMVA Press, September 2011.

[8] Guanglong Du, Ping Zhang, Jianhua Mai and Zeling Li. "Markerless Kinect-Based Hand Tracking for Robot Teleoperation". International Journal of Advanced Robotic Systems Vol 9(36) May 2012.

[9] Zhou Ren, Junsong Yuan, Jingjing Meng, Zhengyou Zhang. "Robust Part-Based Hand Gesture Recognition Using Kinect Sensor". IEEE Transactions on Multimedia, Vol. 15, No. 5, pp.1-11, Aug. 2013.

[10] Jagdish L. Raheja, Ankit Chaudhary, Kunal Singal, "Tracking of Fingertips and Centre of Palm using KINECT", In proceedings of the 3rd IEEE International Conference on Computational Intelligence, Modelling and Simulation, Malaysia, 20-22 Sep, 2011, pp.248-252.

[11] Valentino Frati, Domenico Prattichizzo, "Using Kinect for hand tracking and rendering in wearable haptics".

IEEE World Haptics Conference 2011 21-24 June, Istanbul, Turkey, pp317-321.

[12] Yang, D., Jin, L.W., Yin, J. and Others, An effective robust fingertip detection method for finger writing character recognition system, Proceedings of the Fourth International Conference On Machine Learning And Cybernetics, Guangzhou, China, 2005, pp. 4191–4196.

[13] Oka, K., Sato, Y., Koike, H., Real time Tracking of Multiple Fingertips and Gesture Recognition for Augmented Desk Interface Systems, Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR.02), Washington, D.C., USA, May, 2002, pp. 411–416.

[14] Ying H., Song, J., Renand, X., Wang, W., Fingertip Detection and Tracking Using 2D and 3D Information, Proceedings of the seventh World Congress on Intelligent Control and Automation, Chongqing, China, 2008, pp. 1149-1152.

[15] Crowley, J. L., Berardand F., Coutaz, J., Finger Tacking As an Input Device for Augmented Reality, Proceedings of International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, 1995, pp. 195-200.

[16] Raheja, J. L., Das, K., Chaudhary, A., An Efficient Real Time Method of Fingertip Detection, Proceedings of 7th International Conference on Trends in Industrial Measurements and Automation (TIMA 2011), CSIR Complex, Chennai, India, 6-8 Jan, 2011, pp. 447-450.

[17] K. K. Biswas, Saurav Kumar Basu. "Gesture Recognition using Microsoft Kinect". Proceedings of the 5th International Conference on Automation, Robotics and Applications, Dec 6-8, 2011, Wellington, New Zealand, pp100-103.

[18] Caifeng Shan. "Gesture Control for Consumer Electronics". In Ling. Shao et. al., editors, Multimedia Interaction and Intelligent User Interfaces, Advances in Pattern Recognition, pages 107–128. Springer London, 2010.

Face Recognition and Using Ratios of Face Features in Gender Identification

Yufang Bao[1], **Yijun Yin**[2*], **and Lauren Musa**[3]
[1] Department of Mathematics and Computer Science and Center of Defense and Homeland Security
Fayetteville State University, NC, USA
[2]Department of Computer Science, Rutgers University, NJ, USA
[3] Department of Mathematics and Computer Science, Fayetteville State University, NC, USA

Abstract - *In this paper, we have developed a system of algorithms for human face identification and for gender classification. The DRLSE level set method is used for identifying the face location, in which we propose to use a reinitialization step to accelerate the speed of finding the face contours. Gabor wavelet transformation is also used to extract the eye and eyebrow regions of a face, from which a set of triplet parameters are created as ratio values in term of the eye and eyebrow features. A three dimensional linear discrimination algorithm is applied to this set of triplet parameters. This gender identification method takes advantage of the invariant ratio of feature distances to build a criterion that is robust and avoids potential problems caused by the change of the field of view (FOV). The criterion is further applied to a set of testing face images to identify the gender of each individual human face, and improved accuracy rate is achieved.*

Keywords: Level set function, Gabor wavelet transformation, human face recognition, gender identification, 3D linear discriminant method.

1 Introduction

Human gender is an important feature used in a computer security system when identifying a person of interest, such as biometric authentication. It is a well-known fact that humans naturally perceive features, including identifying the gender, of a person quickly while it not an easy task for a computer program to do so because it involves complicated information to be processed through various facial appearances. Typically, computerized gender recognition technique is preceded by a face recognition technique.

Face recognition is mostly based upon detecting invariant features of faces regardless of different poses, skin tone, lighting conditions and background. Even though numeral face recognition techniques have existed in literature [1, 2], there are still many challenges as each algorithm fits a specific setting. A particular algorithm may work well in finding faces in a certain setting, but it may fail in a different setting due to the face image qualities. For example, hair, hat, or eye glasses can introduce problems in recognizing a face. One difficulty in face recognition is to establish well-defined rules for identifying a face. Extracting features of a frontal face appeared as a round object with two symmetric eyes, a nose and a mouth can become a complicated process. It involves techniques such as segmentation, morphological operations, and circle fitting, etc.. Various algorithms have been developed for identifying human faces. The approach differs when identifying the face outline first and then the eyes/nose/mouth, or in the reverse order. Some researchers also proposed using a template of human faces. However, each algorithm has its limitations. When locating a head boundary as a closed contour of round shape, the difficulty lies upon integrating detected components together as a face outline because classical edge detection algorithms mostly extracts the edge of a supposedly continuous face outline as disconnected components. Some researchers even suggest using votes of the occurrence of hair and skin textures to find the face in an image [3]. Lam and Yan [4] proposed using snake (active contour) method to locate head boundaries with a greedy algorithm in minimizing an energy function. The snake method is indeed equivalent to a level set method [5]. The common problem with this kind of methods is the expense of the evolution; thus the snake curve (level set surface) needs to be reinitialized in order to efficiently drive the contour to the boundary.

In this paper, we proposed to identify face location in a gray scale frontal face image using Distance Regularized Level Set Evolution (DRLSE) method [6]. DRLSE is the most recent improvement of the level set method that has devised an intricate adjustment of the level set function during its evolution course. With a built-in function, it automatically controls the forward and backward diffusion of the level set function. Although it seems no need to reinitialize the level set function, we found that a re-initialization step to accompany the DRLSE method will improve face outline identification. Our result shows that re-initialization periodically after the DRLSE method was applied have actually accelerated the speed of the algorithm in searching for the face contour. This is effective when a single face is presented as a frontal view gray scale image.

We have also proposed a gender identification algorithm. For gender identification, statistical method has been used from neurological research point of view. Cellerino et al [7] has used a statistical approach together with two modalities of spatial filtration methods to study the minimum information

required for correct gender recognition. Lower accuracy rate and more instability is reported for recognizing female faces than recognizing male faces.

Geometrical based methods are further used to incorporate shape features of human faces. Lian and Lu [8] used the local binary pattern (LBP) method to extract texture information such as edges and corners by labeling image pixels. LBP histograms of separated small regions in a face, namely, the histogram of the labels, are extracted and concatenated into a single vector to represent a face image. Support vector machine (SVM) is then used to perform the gender classification on all the vectors collected from face images. Dong and Woodard [9] improved the gender identification rate by extracting three global geometric features from each eyebrow. Minimum distance (MD) classifier, linear discrimination analysis classifier and support vector machine classifier are then used for identifying human genders.

In this paper, we propose to extract three parameters of different ratio values for identifying genders. This eliminates problems caused by the various FOV of faces in an image and is relatively robust in identifying. We then use a three dimensional linear discrimination algorithm to establish a criterion for classifying the gender of a human face. Our result shows that this method is robust, and an improved accuracy rate is achieved for recognizing the gender of a human face.

2 DRLSE Level Set Method

The advantage of using a level set function for identifying the face outline is that an enclosed continuous contour is ready for use to determine if it belongs to a face. Once a face is located, a circle fitting algorithm can be applied to determine the center of the face.

DRLSE [6] is a geometric active contour model that is implemented using level set gradient flow to minimize designed energy functional consisting of a distance regularization and an external energy. The contour is obtained as a zero level set of an auxiliary function $\phi(x,y)$ called a level set function. The gradient flow drives its zero level set towards desired boundary locations in an image. The evolution can be described as the following partial differential equation:

$$\frac{\partial\phi}{\partial t} = \mu div\left(d_p\left(|\nabla\phi|\right)\nabla\phi\right) + \lambda\delta_\varepsilon(\phi)div\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \alpha g\delta_\varepsilon(\phi) \quad (2.1)$$

where μ, λ, α are constants and function $d_p(x)$ is a double-well potential function defined as

$$d_p(x) = \begin{cases} \frac{1}{2\pi s}\sin(2\pi s) & \text{if } x \le 1 \\ \frac{s-1}{s} & \text{if } x > 1 \end{cases}$$

The function $\delta_\varepsilon(x)$ is defined as

$$\delta_\varepsilon(x) = \begin{cases} \frac{1}{2\varepsilon}[1+\cos(\frac{\pi x}{\varepsilon})] & \text{if } |x| \le \varepsilon \\ 0 & \text{if } |x| > \varepsilon \end{cases}$$

The function g is defined as

$$g(I) = \frac{1}{1+|\nabla(G_\sigma * I)|}$$

The function g is indeed defined as a function of the gradient of the image, I , after a Gaussian smooth operator is applied. It takes smaller values at object boundaries than at the smooth locations. Eqn. (2.1) uses the initial level set function ϕ_0 selected as the following

$$\phi_0(x) = \begin{cases} -c_0, & \text{if } x \in R \\ c_0, & \text{otherwise} \end{cases}$$

Where R is a rectangle region usually selected to enclose the object of interest so that the initial contour of the zero level set will be placed outside the object. Eqn. (2.1) drives the zero level set of function ϕ towards the boundary presented inside an image.

Selection of the initial level set position is crucial for the zero level set to evolve to the desired object boundary. When using DRLSE level set function to detect a face in an image, it is best to place the initial level set close to the outside of the head area. Typically, a user has to input this information. An appropriate level set function can be determined quickly using inward evolution and small iterations. In this paper, instead of manually selecting the initial LSF, we placed the LSF in a broad rectangle region that covers most of the image region. We then applied an adaptive algorithm to periodically determine a new initial level set function based on the local properties of the pixels. The sides of the rectangle R will be adjust inward by 3 units accordingly when $A_i < \varepsilon$, where i=left, right, top, bottom, ε , is the threshold to determine if the left, right, top, and bottom extreme positions are too far out. In our algorithm, we used ε = 4 and A_i is defined as

A_i = Difference in the most left (right) x-coordinate of the zero and -1.9 level sets; for i=left, right;

A_i = Difference in the highest (lowest) y-coordinate of the zero and -1.9 level sets; for i=highest, lowest;

We also take into consideration the relative sharp shape of the chin by adjusting the lowest side of rectangle to the lowest vertical center point position when the difference between the average vertical position of the 5 bottom center points and the bottom extreme points is greater than 2.75.

The proposed re-initialization step combining DRLSE was applied to a low resolution image, with size 50x50, which is a resized copy of the original face image, for locating the face outline. In Figure 1, the initial rectangle level set is shown in (a). It is reinitiated into a new level set (shown in (c)) based

on applying the above criterion to (b). Our result shows that the re-initialization has reduced the number of iterations needed for the original DRLSE method to find the face outline.

(a) (b) (c)

Figure 1. (a) the initial rectangle region LSF. (b) the evolution of the level set. (c). the reinitialized rectangle LSF that is closer to the face.

3 Proposed Gender Identification Method

3.1 Gabor Wavelet Transformation

Gabor wavelets [10-13] are known for its capability of capturing edge information in the shape of a curve in relatively large coefficients. Gabor wavelets play an important role for facial representation, especially in representing round face features, such as face outlines, eyes, eyebrows, and lips. Typically local features can be obtained using a set of wavelet coefficients obtained from a sequence of dilating and rotating a selected mother wavelet. These locally estimated wavelet coefficients are robust to illumination change, translation, distortion, rotation, and scaling [1], therefore, the local features obtained are also robust.

We utilize Gabor wavelets as part of our algorithm in recognizing the left eye and eyebrow in a face image. The Gabor wavelets are also called Gabor filters in applications. Gabor wavelets are self-similar: all filters can be generated by dilating and rotating one selected mother wavelet. The frequency and orientation of resulting Gabor filters are similar to those of the human visual system [4]. After Gabor filters are applied to an image, the significant coefficient values obtained typically indicated the transients from one feature to another, while small coefficients indicated smooth textures within each object.

The Gabor wavelets are defined based on a complex function, called the Gabor function, and is defined as:

$$g(x',y';\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\exp\left[i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right] \quad (3.1)$$

where the first exponential function is a 2-D Gaussian-shaped function, known as the envelope, and the second exponential function is a complex sinusoid. The parameter λ is the wavelength of the sinusoidal factor; ψ is the phase offset; σ is the standard deviation of the Gaussian envelope that decides the size of the support for the Gaussian envelop; γ is the spatial aspect ratio, and specifies that the support of the Gabor function is an ellipse shape when γ≠1. The coordinate (x',y') is obtained from rotating the coordinate (x,y) by an angle θ, and can be written as:

$$\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} \quad (3.2)$$

where θ specifies the orientation of the rotated major axis of the elliptical Gaussian shape function in eqn. (3.1).

The real and imaginary parts of the Gabor function can be written as:

$$\text{Real}(x',y';\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\cos\left(\frac{2\pi x'}{\lambda}+\psi\right) \quad (3.3)$$

$$\text{Imag}(x',y';\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\sin\left(\frac{2\pi x'}{\lambda}+\psi\right) \quad (3.4)$$

The rotation of the Gabor function resulted in Gabor filters along N number of orientations, which are further applied to an image. They are corresponding to N angles used in Gabor functions with the angles equidistantly distributed between 0 and 2π radians and are increased at an interval of 2π/N. The value of N is selected based on the computation time used and the completeness of image representation. In our study, we chose N=16. Thus, N convolutions will be computed, and is defined as:

$$r(x,y)=g*I=\iint_{(\xi,\eta)\in\Omega}I(\xi,\eta)g(x-\xi,y-\eta;\lambda,\theta,\psi,\sigma,\lambda)d\xi d\eta \quad (3.5)$$

where Ω is the collection of image pixels. The Gabor wavelet representation of an image is the combination of the convolutions at the N different orientations. Examples of the image output using 2-D Gabor filter are given Figure 2.

Figure 2. Magnitude output of Gabor filter when N=16.

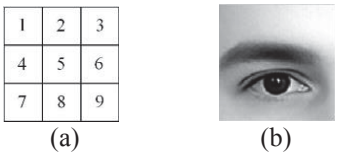
Typically, the preferred spatial frequency and the wave size are not completely independent when selecting the Gabor wavelets. The Gabor wavelets are used as a set of basis that best represents an image and typically the bandwidth and σ should satisfy the following equation [11]:

$$\sigma = \frac{1}{\omega_0}\sqrt{2\ln 2}\left(\frac{2^{bw}+1}{2^{bw}-1}\right)$$
$$\lambda \cdot \omega_0 = 2\pi \qquad (3.6)$$

where ω_0 is the radial frequency in radians per unit length and *bw* is the bandwidth. In this paper, we select $\sigma = 4.4974$ and $\omega_0 = 0.7854$.

3.2 Extracting the Eyebrow and Eyeball Features

Gender information is embedded all over human faces, but is more significant in salient features such as eyes, eyebrows and lips [14]. In this paper, because of the symmetry of the two eyes and eyebrows, we extract three ratio parameters from the left eye region only. In order to locate the desired area, the located face area is divided into 3×3 square sub-regions following the common face pattern that is shown in Figure 3(a). The left eye of a human face is typically located in the first region. To ensure that the left eye and eyebrow are fully selected, this region is extended to include a larger region , see Figure 3(b).



(a) (b)

Figure 3. The main face is divided into (a) 3×3 squares and (b) Extended image of the left eye area.

The selected left eye area is further separated into an eyebrow and an eye. The geometrical properties of these two objects are studied to exact their unique features separately. Here, we use the adaptive threshold algorithm by Niblack [15], which was originally used for segmenting document images. This method finds threshold value within a local window by calculating pixel wise threshold using local mean, $\mu(x, y)$, and local standard deviation, $\sigma^2(x, y)$ for a pixel (x, y) [16,17]. Let the local area of interested pixels be of size $k \times k$, the threshold for each pixel, $T(x, y)$, is calculated by using the following equation:

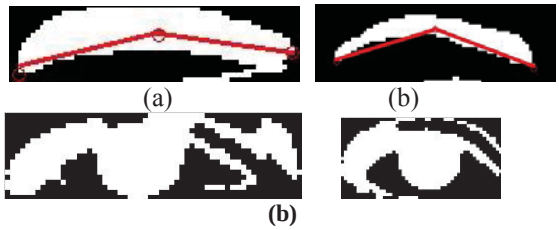
$$T(x, y) = \mu(x, y) + k\sigma^2(x, y) \qquad (3.7)$$

3.3 Ratios for Gender Representation

The eye in this paper is referred to the visible part of a human eyeball from a frontal face image; therefore, we also call it an eyeball in the rest of this paper. The length and height of the left eyeball and eyebrow of a human face are measured, and three parameters are defined as the ratios of the measured values:

$$ratio_1 = \frac{EyebrowLength}{EyebrowHeight},$$
$$ratio_2 = \frac{EyeballLength}{EyeballHeight}, \qquad (3.8)$$
$$ratio_3 = \frac{EyebrowHight}{EyeballHeight},$$

The length and height of the left eyeball are measured as the size of a rectangle that contains the extracted visible eye area see Figure 4(c)(d). The height of the eyebrow is measured as the height at the middle point location of the eyebrow. The length of the eyebrow is calculated by taking into account the curvature of the eyebrow and is approximated as the sum of two line segments, see Figure 4(a)(b). Because the eyes and eyebrows are most significant features of human face, the ratio parameters selected from these locations provide sufficient information in measuring the size of the eyes in relation to the size of a face; therefore, it is invariant to the FOV of the face.



(a) (b)

(b)

Figure 4. Extracted binary images of left eyebrow and left eye. (a) male eyebrow (b) female eyebrow (c) male eye (d) female eye.

3.4 Linear Discriminant Method

Fisher's linear discriminant (FLD) is a well-known method widely used in statistics, pattern recognition and machine learning to characterize or separate two or more classes of objects or events from a linear combination of features [18]. The resulting combination of features may be used as a linear classifier, or more commonly, for dimensionality reduction to serve as a criterion for classification.

Let $X_1 \in \Re^{n \times m_1}$ and $X_2 \in \Re^{n \times m_2}$ represent observations from two classes, in our case, female and male classes. X_1 has *n* properties and m_1 observations. X_2 has *n* properties and m_2 observations. The linear discriminant method first finds a vector $w \in \Re^n$ so that, after the observed data being projected onto vector w, the distances of means in the projected space for different classes will be maximized while the data scattered in the projected space will be minimized. The projection of a data X onto w can be defined as an operation of inner product:

$$y = w^T \cdot X \tag{3.9}$$

where '*T*' stands for the transpose of a vector. To maximize the distance of means in the projected space and minimize all of the scatters in the projected space, a criterion can be defined for the degree of discrimination, which is the Fisher discriminant ratio defined as

$$f\left(w, \mu_{X_1}, \mu_{X_2}, S_{X_1}, S_{X_2}\right) = \frac{\left|\mu'_{X_1} - \mu'_{X_2}\right|^2}{S'^{\,2}_{X_1} + S'^{\,2}_{X_2}} \tag{3.10}$$

where μ'_{X_i}, S'_{X_i} are the corresponding projected value of the mean, μ_{X_i}, the within each class scatter, S_{X_i}, of X_i (i=1, 2), and are calculated as: $\mu'_{X_i} = w^T \mu_{X_i}$ and $S'_{X_i} = w^T S_{X_i}$ where S_{X_i} is defined as:

$$S_{X_i} = \sum_{i=1,2} (X_i - \mu_{X_i})(X_i - \mu_{X_i})^T$$

It can be verified that $S'^{\,2}_{X_i} = w^T S_{X_i} w$. Therefore we have:

$$S'^{\,2}_{X1} + S'^{\,2}_{X2} = w^T S_w w$$

The scatter within the classes, S_w, and the scatter between class , S_B, of X_1 and X_2 can be defined as:

$$S_w = \sum S_{Xi} = S_{X1} + S_{X2}$$

$$S_B = (\mu_{X1} - \mu_{X2})(\mu_{X1} - \mu_{X2})^T$$

Hence, the Fisher discriminant ratio eqn (3.10) can be further written as:

$$f(w) = \frac{w^T((\mu_{X1} - \mu_{X2})(\mu_{X1} - \mu_{X2})^T)w}{w^T(S_{X1} + S_{X2})w} \tag{3.11}$$

The maximum of the Fisher discriminant ratio is reached at

$$w = (S_{X1} - S_{X2})^{-1}(\mu_{X1} - \mu_{X2}) \tag{3.12}$$

which gives the maximum Fisher discriminant ratio as

$$\max_{w \neq 0} f(w) = (\mu_{X1} - \mu_{X2})^T (S_{X1} + S_{X2})^{-1} (\mu_{X1} - \mu_{X2})^T \tag{3.13}$$

The calculated w is used in eqn. (3.9) to obtain the projected value, y(X_1) and y(X_2). A criterion is then established to determine the class of the sample:

$$y_0 = \frac{m_1 y(X_1) + m_2 y(X_2)}{m_1 + m_2} \tag{3.14}$$

where m_1, m_2 are the number of observations of X_1, X_2 separately.

Once the criterion of eqn. (3.14) is calculated, it can be used to determine the gender class of a new observation Z into the following two cases:

1) For y_0 < y(X_1), if y(Z)> y_0, Z belongs to X_1 class; otherwise, Z belongs to X_2 class.

2) For y_0 < y(X_2), if y(Z)> y_0, Z belongs to X_2 class; otherwise, Z belongs to X_1 class.

The triplet parameters defined in eqn. (3.8) is calculated on all the images in our human face library. There are 21 men faces and 21 women faces collected in the library. The Fisher linear discriminant is then applied to the data array of 42 entities to build a criterion. The data array consists of triplets of ratio values. This results in a criterion that is a three dimensional plane associated with the discriminant function. The equation of the discrimination surface can be written as:

$$y_0 = c_1 x + c_2 y + c_3 z \tag{3.15}$$

where $w = (c_1, c_2, c_3)$ is calculated from eqn. (3.12) and y_0 is calculated from eqn. (3.14). For our data collected from 21 men faces and 21 women faces, the coefficients are:

y_0 = - 0.2691; c_1= - 0.0457; c_2= - 0.0069; c_3=0.2510

This gives the function of discrimination surface as

$$-0.0457x - 0.0069y + 0.2510z + 0.2691 = 0$$

This criterion is then applied to test new images to classify whether the person on each frontal face image is a male or female. The image of discrimination surface is shown in Figure 5. It can be seen that the triplet data points from male face images fell inside the blue dot area and the triplet data points from female face images fell inside the red dot area for most of the images in our library.

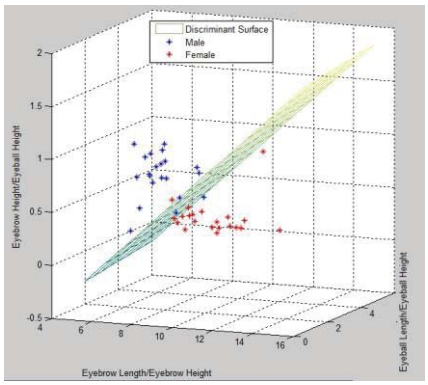


Figure 5 Discriminant surface and data used to build the discrimination function

4 Experimental Results

To test our criterion using the discriminant plane obtained from the linear discriminant method, we apply the criterion to a new set of 20 men and 20 women's images, and thus total 40 frontal face images for gender recognition. The testing data are shown in Figure 6, from which we can see that the data distributed in the three-dimension feature space, with majority male and female face data fell in the two sides of the discrimination surface. The figures shown in this paper is of an individual face from the database in [19]. The identification result is provided in a table shown in Table 1, in which the accuracy of this proposed gender classification method is calculated.

Table1. The experiment result of discrimination function

	Men	Women
Total Number	20	20
Correct Recognized	19	18
Success Rate	95.00%	90.00%

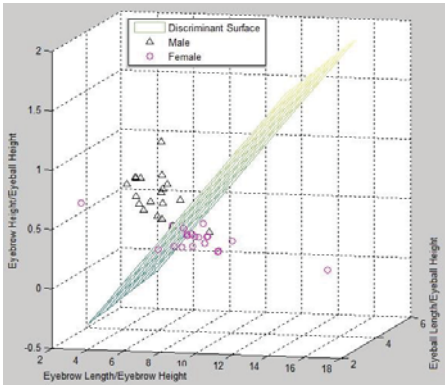


Figure 6. Discrimination surface and testing data distribution.

Compared with Dong and Woodard's study[9], in which a geometrical-based feature extraction method is used to extract three global features from eyebrow only, their average success rate, include the left eye and the right eye, is 79.25% and the average success rate for linear discrimination analysis classifier is 83.5%. For four features, the success rate of male identification is 73.3% and the success rate for female is 84%. Our proposed gender recognition algorithm improves the accuracy of gender identification rate by using the robust features of both the eyebrow and eye. It achieves a success rate of 95% for male and 90% for female on human frontal face images, which is a significant improvement compared to Dong and Woodard's study. This shows that the ratio parameters we selected provided more significant information for gender identification. The calculated three ratio measurements of human eyebrow and eyeball have taken advantage of the common features of every frontal human face image. The three measurements also take into account the common sense that the size of significant feature is proportional to the change of other landmark feature in a human face, which is indeed related to FOV that determined the face size in a face image. Therefore, the parameters chosen automatically eliminate the possible error caused by the different FOV of face images.

5 Conclusions

In this paper we have presented our algorithms to improve the accuracy of recognizing a human face and identifying the gender of an individual from a human frontal face image. This study shows the significant advantage of combining features of both eyebrow and eye in human gender differentiation. In addition, using the related ratios of the acquired parameters frees the users from worrying about the FOV of the images, thus increases the accuracy for gender identification.

6 Acknowledgments

This research is Partially funded by National Science Foundation, ISAS HBCU-UP #1036257.

* Part of the research in this paper was done while Mr. Yin was a student at Fayetteville State University.

7 References

[1] W. Zhao, R. Chellappa, P. J. Phillips, et al. "Face recognition: A literature survey". ACM Computing Surveys (CSUR), 2003, 35(4): 399-458.

[2] M. Yang, D. Kriegman, N. Ahuja, "Detecting Faces in Images: A Survey". IEEE Transactions On Pattern Analysis And Machine Intelligence, January 2002, Vol. 24(1):34-58.

[3] S. Fahlman and C. Lebiere, "The Cascade-Correlation Learning Architecture," Advances in Neural Information Processing Systems 2, D.S. Touretsky, ed., pp. 524-532, 1990.

[4] C. Xu, A. Yezzi, and J. Prince, "On the relationship between parametric and geometric active contours," in Proc. 34th Asilomar Conf. Signals Syst., Comput., Pacific Grove, CA, Oct. 2000, pp. 483–489.

[5] K. Lam and H. Yan, "Fast Algorithm for Locating Head Boundaries," J. Electronic Imaging, vol. 3, no. 4, pp. 351-359, 1994.

[6] C. Li, C. Xu, C. Gui, and M. Fox, "Distance Regularized Level Set Evolution and Its Application to Image Segmentation", IEEE Trans. Image Processing, vol. 19 (12), pp. 3243-3254, 2010.

[7] A. Cellerino, D. Borghetti, and F. Sartucci, "Sex differences in face gender recognition" in humans J. Brain research bulletin, 2004, 63(6): 443-449.

[8] H. C. Lian and B. L. Lu. "Multi-view gender classification using local binary patterns and support vector machines" Advances in Neural Networks-ISNN 2006. Springer Berlin Heidelberg, 2006: 202-209.

[9] Y. Dong and D. L. Woodard. "Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study", Biometrics (IJCB), 2011 International Joint Conference on. IEEE, 2011: 1-8.

[10] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition". Pattern analysis and applications, 2006, 9(2-3): 273-292.

[11] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition", Image processing, IEEE Transactions on, 2002, 11(4): 467-476.

[12] T. S. Lee, "Image representation using 2D Gabor wavelets" Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1996, 18(10): 959-971.

[13] F. Tang and H. Tao, "Non-orthogonal binary expansion of Gabor filters with applications in object tracking", Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on. IEEE, 2007: 24-24.

[14] J. Sadrô, I. Jarudi, and P. Sinhaô, "The role of eyebrows in face recognition" J. Perception, 2003, 32: 285-293.

[15] W. Niblack, "An Introduction to Digital Image Processing", Pretice-Hall, Englewood Cliffs, NJ, 1986

[16] Y. T. Pai, Y. F. Chang, and S. J. Ruan, "Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images Pattern Recognition, 2010, 43(9): 3177-3187.

[17] A. Majumder, M. Singh, and L. Behera, "Automatic eyebrow features detection and realization of avatar for real time eyebrow movement", Industrial and Information Systems (ICIIS), 2012 7th IEEE International Conference on. IEEE, 2012: 1-6.

[18] S. J. Kim, A. Magnani, and S. Boyd, "Robust Fisher discriminant analysis", Advances in Neural Information Processing Systems, 2006, 18: 659.

[19] http://facedetection.com/datasets/

Texture Modelling for Age Invariant Face Recognition

Fahad Bashir Alvi and Associate Professor Russel Pears
Knowledge Engineering & Discovery Research Institute
Auckland University of Technology, New Zealand
Private Bag 92006, Auckland 1142
AUT Tower, Level 7, Auckland 1010
falvi@aut.ac.nz,rpears@aut.ac.nz

Abstract— *This Research study proposes a novel method for face recognition based on Texture boundaries or edges by using Canny and Sobel Edge detection that make use of global and personalized models. The system is aimed to recognize faces and identify their similarity across ages. A Personalized model covers the individual aging patterns while a Global model captures general aging patterns in the population. We introduced a de-aging factor that de-ages each individual in the image gallery. We used the k nearest neighbor approach for building a personalized model. Regression analysis was applied to build the models. During the test phase, we built a similarity matrix and determined the rank 1 identification by using a Leave One Person Out strategy. We used FG-Net database for validating our technique and achieved 62 percent Rank 1 identification rate.*

Keywords- Edges; K Nearest Neighbor; Personalised Model; Regression;

1 INTRODUCTION

Face recognition is a complex area of research [1-6]. It has found many a useful application in real life. There are five major challenges which need to be addressed in face recognition systems: pose, expression, illumination, occlusions and aging. It has been observed that for solving one challenge we have to make a compromise on other challenges [7-8].

We base our study on the texture of the image. The basic theme is that with the age, the lines (edges) on the face grow in length and in number. Based on this theory, we took an image of a person across different ages and studied the edges on that person's face.

Face recognition across ages remains an open research question. This study is aimed at a solution to this problem. It has many useful real life applications. For example in law enforcement where an image of a suspect is obtained but no match is obtained for the suspect in the crime database. The law enforcement authorities need information on the suspect such as last known address and other details in order to apprehend the suspect. If the suspect does actually appear in the crime database at a previous age band then face recognition software would enable a match to be made.

In terms of face recognition, there are three main challenges:

1. We should be able to recognize a face of a 50 year old person when we only have his/her image taken at 25 years of age. Fig 1 shows the effects of aging on a given person over time.

2. Estimation of age from a given image.

3. Age Simulation of images to build age progressed or age retarded images.

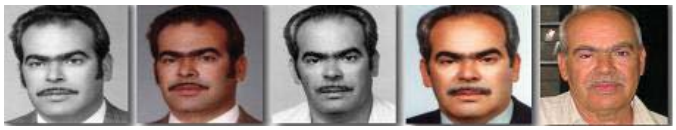


Figure 1 – Same Individual at different ages from FG-NET Database [17].

This study focuses on designing a novel framework for meeting challenge 1 on face recognition. We assume that this is a closed set identification task and that we know the age of the probe image. In real life it would be possible to obtain an estimate of the age of the probe image from a human expert.

In our approach we divide the face into five slices. Starting from top the first slice coves the area above and including eyebrows. The second slice covers the eyes. The third slice covers the nose, while the fourth one covers the mouth and the fifth one covers the area of chin. They are called five features (or indexes). Then edges are determined on each of the slice, using canny edge algorithm. The frequency of these edges in one slice is considered as an image feature. Thus we have five features for each face.

We divide the age range into ten age bands, each spanning five years, except for the last age band which spans 10 years, because the numbers of images in this age band for the FG-net database is lesser in number. This data is used to construct a

global model, where centroid of each band for a given feature forms the trajectory of that feature across the 10 age bands.

Later, we construct a personalized model, in which we use a deviation factor to de-age the values of features and thus develop the model.

2 LITERATURE REVIEW

Yan et al. [1] discussed a method that developed the concept of coordinate patches and GMMs . Facial ages were estimated by these patches. In their method, the face image of an individual is encoded as a group of overlapped spatially flexible patches (SFPs). Local features are extracted by a 2D discrete cosine transform (DCT). Coordinate information together with the local features is integrated with the help of these patches. These extracted SFPs are then modeled with GMMs. This model is is used to estimate the age of a person in the input facial image, by comparing the sum of likelihoods from total SFPs of the hypothetic age.

Guo et al. [2] developed a local based regression classifier to learn the aging function. He introduced a method to estimate the age by using a manifold learning scheme. It was then used to predict the age from a given image.

Fu et al. [3] also proposed a manifold learning scheme. In this technique, a low dimensional manifold is learnt from a set of age separated face images. Linear and quadratic regression functions were applied on the low dimensional feature vectors from the respective manifolds. These were used to estimate the face age.

Lanitis et al. [4] used the active appearance models. It is a statistical face model, which has been used in age estimation problems. In their approach, after AAM parameters were extracted from face images landmarked with 68 points, an aging function was developed. It was then optimized using Genetic Algorithms. Their results are shown in Table 1 given in subsequent text.

Sethuram et al. [5] also built a face aging model based on AAMs, support vector machines (SVMs) and Monte-Carlo simulation. In their paper, two experiments were setup. In experiment 1, they proved that the accuracy of face recognition decreases when probe faces age with time. In experiment 2, the probe faces are first artificially aged (using age simulation functions) to the same age of the gallery by using the face aging model. Then, the face recognition algorithm is applied. Thus they achieved greater accuracy than the ones in experiment 1.

Geng et al. [6] introduced a subspace called Aging pattern Subspace (AGES). The assumption was that similar faces age in similar ways for all individuals. Their basic idea is to model the aging pattern. This model is defined as a sequence of an individual's face images sorted in time order, by constructing a representative subspace. The proper aging pattern for a previously unseen face image is determined by the projection in the subspace that can reconstruct the face image with a minimum reconstruction error. Then the position of the face image in that aging pattern will indicate its age.

Suo et al. proposed a statistical model for face age simulation and age estimation. A hierarchical AND–OR graph representation is constructed . In this case faces are decomposed into different parts and organized in the graph structures. The face aging process is modelled as a Markov chain. In particular, hair styles are processed as one kind of aging effects [7]. Thus the complete hierarchy is built.

3 Methodology

3.1 Preprocessing

Each image that is used for model making and probing is first normalized. The images are of different sizes, so we first register all the images so that they have same size, taking eye distance as the basis. The final size chosen was 129x99. Secondly, we make five slices of the face as shown in Figure 2. In order to get edges from each slice of face image we use canny edge detector. We use frequency of edges as simple statistical parameter in each slice of image and consider it as feature and build the global and personalized function on that. Whole the population was divided into 10 age bands, each band spanning a period of 5 years. The last age band spans 10 years. It was necessitated because of the fact that number of images in last bands was small, so last two age bands have been made into one.

Figure 2 – Five features selected from one Image

3.2 Edge Detection

We used the Canny and Sobel edge detectors to find edges. The Canny operator was designed to be an optimal edge detector (according to particular criteria --- there are other detectors around that also claim to be optimal with respect to slightly different criteria). It takes as input a gray scale image,

and produces as output an image showing the positions of tracked intensity discontinuities

Figure 3 – Canny edge detection of one image from FG-NET[17].

The Canny operator works in a multi-stage process. First of all the image is smoothed by Gaussian convolution. Then a simple 2-D first derivative operator is applied to the smoothed image to highlight regions of the image with high first spatial derivatives. Edges give rise to ridges in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets to zero all pixels that are not actually on the ridge top so as to give a thin line in the output. This process is known as non-maximal suppression. The tracking process exhibits hysteresis is controlled by two thresholds: t1 and t2, with t1 > t2. Tracking can only begin at a point on a ridge higher than t1. Tracking then continues in both directions out from that point until the height of the ridge falls below t2. This hysteresis helps to ensure that noisy edges are not broken up into multiple edge fragments [11].

Sobel operator [12] is another edge detector. It is a discrete differentiation operator used to compute an approximation of the gradient of image intensity function for edge detection. At each pixel of an image the Sobel operator gives the corresponding gradient vector .It convolutes the input image with a kernel and computes the gradient magnitude and direction. It uses the following 3x3 two kernels

$$A_j = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad A_j = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}$$

3.3 Global Model

A global model is induced over the entire population capturing global trends. Such models capture useful general trends across the population over the spectrum of age bands used. They capture those trends in the data that are valid for the whole problem space.

After grouping images into their respective age bands, we determine the discriminative features and calculate the 5 unique features (or indices) (i.e. frequency of edge pixels for a given feature) for each image. The centroid of each age band is determined and an n^{th} order polynomial (we experiment with values of n in the range [1..3]) function is developed that spans all the age bands using methods of least squares and nonlinear regression. The Global model thus created can then be used for all modelling and probe images. Equation 1 represents the Global model where yi is the value of each coefficient and xi represents its age band index.

$$y_i(x_i)^{(global)} = y_1 x^n + y_2 x^{n-1} + \ldots + y_n x + y_{n+1} \quad (1)$$

3.4 Personalised Model

The transductive or personalized approach, in contrast to the inductive approach models each point in the problem space. It was defined by Vapnik in [14] and used by Kasabov in [15]. Recent research by Pears et al. in [16] shows that personalized modelling was responsible for obtaining an order of magnitude increase in predictive accuracy in a stock market application.

We observe that the aging process differs from person to person and hence the personalized approach is very useful for age invariant face recognition when change in features tend to be specific from one time period to another. This approach is based on the premise that to solve a given problem one should avoid first developing a generalized problem and should instead construct a solution for the particular problem at hand. K-NN (K-nearest neighbor) is one of the well-known transductive techniques and is most widely used for personalized modeling. For every new sample, the nearest K samples are extracted from a data set using a distance metric which defines similarity between the elements of sample. It is very common to use Euclidean distance for this purpose. In this method if a vector X is given, the output value y is calculated as the average of the K nearest neighbors to X over the given data set.

A personalized model needs to be prepared each time a new sample (image) is made available. The model is constructed as follows. We take an image, determine its age band (in closed set evaluation mode the age of the probe image is known; in

real-world applications, a human expert can be used to estimate the age) and obtain its K nearest neighbors (the optimal value of K can be found by experimentation). When we get a sample of a probe image we de-age that sample as follows:

$$DF_i = E(X_i) - Y_{p(i)} \qquad (2)$$

We have introduced a deviation factor (DF) as defined in equation (2) above. In equation 2 E(Xi) represents the mean of feature i across age band i and Yp(i) is the feature value of the probe image in that age band.

$$P_i = G_i - DF_i \qquad (3)$$

The de-aging process is done in equation (3) by subtracting the deviation factor DFi of the probe image in the given age band from the value Gi returned by the global model at that age band i. The resultant value Pi is then incorporated into all remaining age bands. We then find the K nearest neighbors to Pi and determine the centroid Ci of these neighbors. This process is repeated for each age band to complete the construction of the personalized model for the probe image. We use these centroids to develop a nonlinear function across the age bands; once again we fit an nth degree polynomial curve using the method of least squares.
Equation 4 represents the Personalized model where yi is value at age band i; xi = Ci represents the centroid value at age band of the probe image at band i and n is the degree of the polynomial.

$$y_i(x_i)^{(Personalized)} = y_1x^n + y_2x^{n-1} + \ldots\ldots + y_nx + y_{n+1} \qquad (4)$$

4. EXPERIMENT

4.1 Preparation of Data

Experiments were performed on the publicly available FG-NET database [17] which is one of the most widely used databases in image processing for benchmarking new methods. It contains 1002 color and gray face images of 82 persons across a range of different ethnicities. There is a large variation in lighting, expression and pose across the different images. The image size is 300×400 in pixel units, on the average. The ages vary from 0 to 69 years. There are on the average 12 images per person across different ages. The database was divided into ten different age bands (0-5, 6-10....etc). Last age band was taken to span 10 years (45-55) because of paucity of images as shown in Table 1 below.

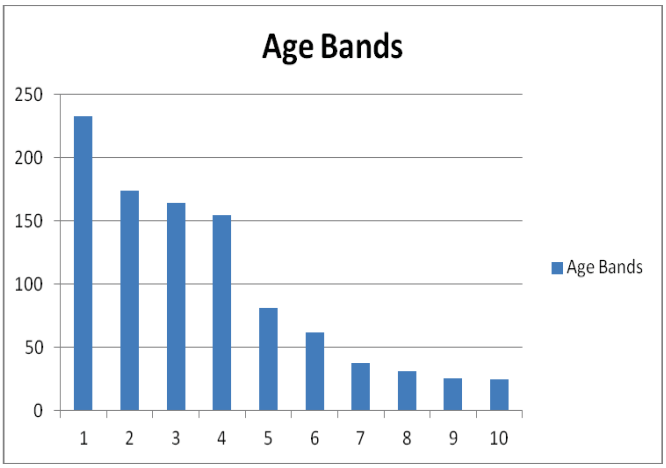


Figure 4 – Age band construction from the FG-NET Database [17].

It was observed that distribution of images in various age bands is random. There were certain age bands in which a person had no image. So, there was a need to fill in the missing data for such age bands. We used normal distribution for this task given a minimum and maximum range of the values of other images in that age band. Mean and Standard deviation for each band were calculated and these were used for calculating the missing values using a Confidence Interval of 95%. In this process, a considerable number of images were added making a total of 1362 images. These were used for creating global model.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < x < \infty \qquad (5)$$

4.2 Training

During training the major tasks were to construct the global and personalized models. Global model was created once for all images. Later, for every probe image, a personalized model was prepared. This model was used for finding specific trends for the given probe image.

4.3 Testing

During the testing phase we build a similarity matrix and use Leave One Person Out strategy (LOPO) for validation. We chose 82 different images from 82 individuals for testing. We took each image from the test set and constructed its personalized model. With the help of the personalized model, and the age band of the probe image, we arrived at a value, called predicted distance. We computed the 5 nearest neighbors to the predicted distance and attempted to match the neighbors obtained with images from the dataset for all of the n fiducial features. If the probe image was found to be nearest neighbor it was termed as Rank 1 identification.

5. EVALUATION

The results in Table 1 clearly indicate that the Personalized approach significantly outperformed the model proposed by Liu et al. In fact, the personalized model on its own was capable of improving on their approach on its own. This is to be expected as aging is a personalized process and its trajectory is very different for different individuals, depending on a range of factors such as lifestyle, genetic disposition and others.

Table 1 – Rank1-Identification Rate

Models	***Database(#subjects,# images) in probe***	***Results***
Liu at al[9]	(82,82)	48.5%
Usang [8]	(82,82)	37.5%
Proposed Personalised Model (Sobel)	(82,82)	24.0%
Proposed Personalised Model (Canny)	(82,82)	62.0%

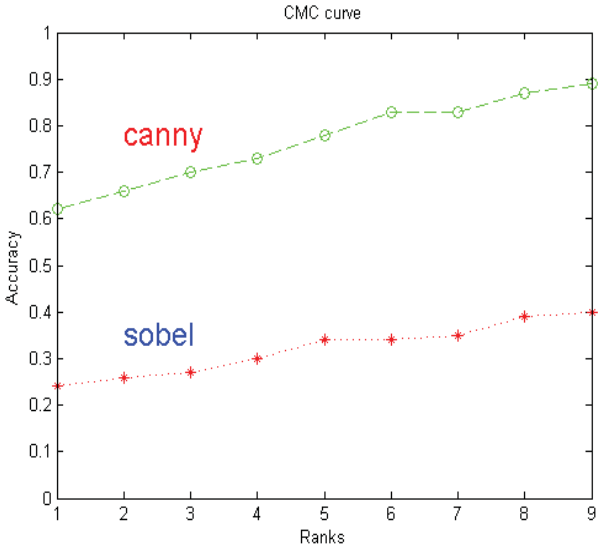


Figure 5 – Cumulative matching characteristics (CMC) curves.

6. CONCLUSION AND FUTURE WORK

We have presented a Personalized modeling approach to the problem of age invariant face recognition. Our experimentation indicated the superiority of the approach over the standard method of using a single global aging trajectory that is built on the entire population of individuals in the training dataset. In order to arrive at more accurate individualized aging profiles the scheme becomes costly in terms of computational complexity and time. The construction of the personalized model is time consuming as a Regressor needs to be constructed for each image and adjusted for aging across different age bands. Our future work will concentrate on making the individualized profiling more efficient. We are exploring different directions. One of them will be to explore constructing models for similar groups of individuals, rather than for every single individual. This will represent sharp reduction in the computation time as the size of group can be expected to be a small fraction of the global size of individuals. For the grouping process we are investigating the use of the Denfis proposed by Kasabov and Qun [18] clustering algorithm which uses an adaptive approach and is ideally suited to our problem context as new images are arriving on a continuous basis and a grouping (clustering) needs to be found dynamically without the need to completely reorganize existing groupings.

Another research direction that we are pursuing is the use of a non uniform voting process. At the moment all fiducial features receive an equal importance as they are used as a vector. However it may be beneficial to use a weighted voting scheme whereby more discriminative features receive a higher weight in the voting process than other less discriminative ones.

7. References

[1] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang. Regression from patch-kernel. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2008.7

[2] G. Guo, Y. Fu, C. Dyer, and T. .Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. In IEEE Trans. on Image Processing, pages 1178–1188, 2008.

[3] Y. Fu and T. Huang. Human age estimation with regression on discriminative aging manifold. In IEEE Trans. on Multimedia, volume 10, pages 578–584, 2008.

[4] A. Lanitis, C. Taylor, and T. Cootes. Modeling the process of ageing in face images. In the Seventh IEEE Int'l Conf. on Computer Vision (ICCV), Kerkyra, volume 1, pages 131–136, 1999.

[5] A. Sethuram, E. Patterson, K. Ricanek, and A. Rawls. Improvements and performance evaluation concerning synthetic age progression and face recognition affected by adult aging. In the 3rd IEEE Int'l Conf. on Biometrics, 2009.

[6] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. IEEE TPAMI, 29:2234–2240, 2007.

[7] Jinli, S., et al. (2010). "A Compositional and Dynamic Model for Face Aging." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(3): 385-401.

[8] Unsang, P.,etal. (2010). "Age-Invariant Face Recognition." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(5): 947-954.

[9] Z. Li, U. Park, and A. Jain. A discriminative model for age invariant face recognition. In IEEE Trans. on Information Forensics and Security, 2011.

[10] Jafri, R. and H. Arabnia (2008). Fusion of Face and Gait for Automatic Human Recognition. Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on.

[11] Canny, John, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*Vol. PAMI-8, No. 6, 1986, pp. 679-698.

[12] James Clerk Maxwell,1868 DIGITAL IMAGE PROCESSING Mathematical and Computational Methods.

[13] Illumination Invariant Face Recognition: A Survey. Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on.

[14]Vapnik, V.N.: Statistical Learning Theory. Wiley Inter-Science, Chichester (1998).

[15] Kasabov, N.(2007)"Global, local and personalised modeling and pattern discovery in bioinformatics": An integrated approach Pattern Recognition Letters Volume 28, Issue 6, 15 April 2007, Pages 673–685 Pattern Recognition in Cultural Heritage and Medical Applications.

[16] Pears, R., Widiputra, H., and Kasabov, N. "Evolving integrated multi-model framework for on line multiple times series prediction", Evolving Systems, Springer, 4(2): 99-117.

[17] FG-Net aging database. [Online]. Available: http://sting.cycollege.ac.cy/ _alanitis/fgnetaging/.

[18] Kasabov, N. K. and S. Qun (2002). "DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction." Fuzzy Systems, IEEE Transactions on 10(2):144-154.

Spectral Collaborative Representation Based Classification by Circulants and its Application to Hand Gesture and Posture Recognition from Electromyography Signals

Ali BOYALI, Naohisa HASHIMOTO, and Osamu MATSUMOTO
National Institute of Advanced Industrial Science and Technology
Robot Innovation Research Center - Smart Mobility Research Team / Tsukuba-Japan

Abstract— *In this study we introduce and demystify a novel signal pattern recognition method, Spectral Collaborative Representation based Classification (SCRC) and demonstrate its application for recognition of hand gestures and postures using Electromyography sensors. A recently released Thalmic Labs MYO armband is used to gather muscle electromyography signals. Along with the new signal pattern classification algorithm, we also introduce a training approach which implicitly embeds the gesture boundaries in a training dictionary that allows continous gesture and posture recognition. The worst recognition accuracy we obtained for a set of experiments is over 97% which is the highest recognition results in the literature where bio-signals are used.*

Keywords: EMG gesture, continous gesture recognition, spectral representation, gesture training matrix, myo armband

1. Introduction

Recognition of the patterns in bio-signal applications has been a challenging engineering problem due to the stochastic nature of the biological processes. The bio-signal classification applications have been utilized to diagnose the metabolic anomalies and diseases, for rehabilitation and monitoring in medicine. The proliferation of the mobile computing platforms such as smart phones and tablets made the bio-signal pattern recognition an appealing tool for creating intuitive Human Computer Interface (HCI) applications via gesture and posture recognition and detecting the physiological conditions such as heart rate on the human body by wearable gadgets. In line with the development of mobile computing platforms, the advanced sensors have been introduced to the market that can communicate with these platforms. Recently Thalmic Labs' MYO armband which can measure and recognize the hand gestures using Electromyography (EMG) signals has been released to the developers. The Software Development Kit (SDK) of the device allows the developers to capture raw EMG signals from the eight EMG sensors which is wearable on the arm in form of an armband.

In this study, we detail a novel method in signal pattern classification which yields far better gesture recognition accuracy then the armband reports. The method gives over 97% accuracy which is the worst result for our experiment set but the best among the EMG and other bio-signal classification studies in the related literature. The methods we propose for the training phase allow the user to obtain a gesture dictionary on the spot. The classification method proposed in this paper is a spectral variant of the Collaborative Representation based Classification (CRC) which has been successfully used in face [1] and signal pattern recognition applications [2] with an overwhelming classification accuracy.

A subspace clustering method is used to build a training dictionary which enables the system to recognize signal patterns in a streaming manner. This study brings about the following contributions in the bio-signal pattern classification literature.

- The spectral features of the observed signals are obtained using circulant and diagonalization matrices. This approach remarkably reduces the computational complexity and computation time.

- The subspace methods which are used to build a training matrix implicitly embed the start and end position of the hand gestures and postures, thus, they lead to continuous posture and gesture recognition eliminating spotting the patterns in the signals.

- The training phase is easy to implement, accordingly the end users can build a training dictionary on the spot with regards to their requirements. This flexibility of the training phase and the high recognition accuracy open the way for the use of proposed method in many research areas such as for manipulation of the robotic prosthesis.

- The dynamic hand gestures and static postures are captured during the training phase, hence in the real time application of the study, 10 hand gestures are recognized and mapped to the five hand postures with an overwhelming accuracy and low computation times. The number of gestures is the highest than the numbers

reported in the related literature.

The rest of the paper is organized as follows. In Section 2, we give the brief review of the CRC method and elaborate the SCRC. The sensor MYO armband information and the training procedures are given in Section 3. The simulation results and discussions are given in Section 4. The paper concludes with Section 5.

2. Collaborative Representation based Classification in Spectral Domain

The CRC method was first introduced in the literature to compare the classification mechanisms of the Sparse Representation based Classification (SRC) [3] and CRC [1, 4]. The authors prove that the SRC which makes use of ℓ_1 norm in the objective function is a special variant of the CRC methods in which $\ell_1 \quad and \quad \ell_2$ norms are utilized depending of the requirements in the problem such as noise on the measurements or occlusion on a face image. Both of the methods rely on representing the observed signal by a the linear combination of the representatives that are stacked into a training dictionary as a column vectors. The SRC methods yield high accuracies using an over-complete dictionary or training matrix, on the contrary, the recognition accuracy does not depend on an over-complete dictionary matrix in the CRC methods in which the coefficients of the linear combination are computed by collaboratively making use of the other class representative samples.

In the Regularized Least Square version of the CRC method (CRC_RLS), given the training matrix $A = [A_1, A_2, \ldots, A_n] \in \mathbb{R}^{mxn}$ and the observed signal y, the solution vector x which contains the linear representation coefficients for the systems of equations $y = Ax$ are obtained using the ridge regression. In this problem setup, the optimum coefficients for the objective function given in Eq(1) is calculated as $\hat{x} = Py$ where $P = (A^T A + \sigma I)^{-1} A^T$ and σ is the regularization parameter.

$$\min_x \quad \hat{x} = \|y - Ax\|_2 + \sigma\|x\|_2 \tag{1}$$

Once the solution vector $\hat{x}$ is obtained, the observed signal is labeled evaluating the minimum representation residuals r_i given in Eq. (2) where $\delta_i : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the selection operator that selects the coefficients of i^{th} class while keeping other coefficients zero in the solution vector $\hat{x}$.

$$\min_i \quad r_i(y) = \|y - A\delta_i(\hat{x})\|_2 \tag{2}$$

The ℓ_2 solution gives efficient results when there is data fidelity [1]. However, biological signals do not exhibit data fidelity due to the stochastic nature of biochemical processes. For example, the measured EMG signals are the superposition of Motor Unit Action Potentials (MUAP) of the individual muscle fibers and magnitude and shape of the signals which are random only depends on the duration of the muscle contraction or stretch and the force that the muscles produce [5].

In this case, the ordinary CRC_RLS method yields poor accuracy results. We overcome this difficulty and dramatically improve the recognition accuracy by using the spectral features of the observed signal. The eigenvalues and eigenvectors of a signal reveal important information about the characteristics of the system. In order to find the eigenvalues of a linear system of equations a square matrix is needed. The EMG signal is a continuous time series and we observe the signals by a sliding time window as 1D data. The conventional methods cannot be utilized directly to capture the spectrum of an 1D vector. Therefore, we employ the circulant matrix approach to obtain the spectrum by creating a circular matrix from the observed 1D signal. A circulant matrix is a square matrix with the circularly shifted columns. In this study, the resultant matrix is a Hankel matrix which consists of skew constant diagonals, however other diagonal direction can be used as a circulant. Formally expressing, let's assume the observed signal vector is $y = [y_1, y_2, \ldots, y_n]$, then the trajectory matrix C becomes a square matrix with the skew diagonal entries given in Eq. (3), the first row of which is the observed signal itself.

$$C_y = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ y_2 & y_3 & \cdots & y_1 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & y_1 & \cdots & y_{n-1} \end{pmatrix} \tag{3}$$

A unitary diagonalization matrix F can be used to obtain the eigenvalues of the composed circulant matrix C by relatively lesser number of matrix operation than that of the conventional eigenvalue decomposition. A Fourier matrix can be used for spectral decomposition of any circulant matrix. Assuming a Fourier matrix F, the eigenvalues and eigenvectors are computed by the following equations [6].

$$F = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W & W^2 & \cdots & W^{N-1} \\ 1 & W^2 & W^4 & \cdots & W^2N-2 \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & W^{N-1} & W^{2(N-2)} & \cdots & W^{(N-1)(N-1)} \end{pmatrix} \tag{4}$$

with the entries of $W = e^{\frac{-2\pi i}{N}}$ which are the n^{th} roots of unity. The eigenvalues are computed by the following matrix operations;

$$C = U^* \Omega U \tag{5}$$

where Ω is a diagonal matrix, the entries of which are the eigenvalues of the circulant matrix and $U = \frac{1}{\sqrt{N}}$F is the matrix of column eigenvectors. The matrix operator $()^H$ represents the conjugate or Hermitian transpose operation

which is important in the CRC method as the resultant eigenvalues consist of complex conjugate eigenvalue pairs.

The spectral CRC method operates on the complex eigenvalues which are taken as the features of the observed and training samples. The regression operator is given in Eq. (6) on the complex plane.

$$P = (A^H A + \sigma I)^{-1} A^H \tag{6}$$

It is important to note that, the eigenvectors of circulants captured by each time window then transformed are the same, therefore, the eigenvalues of the circulant matrix are the only features that increase the discriminative power of the method.

3. Sensor Description and Implementation

3.1 MYO Armband

The Thalmic Labs' MYO armband is a fairly new technology which is equipped with eight EMG sensors Fig. 1. The armband also reports the linear and angular acceleration of the device as well as the orientation angles. The affordable armband enables the researchers and developers to access the raw surface EMG signals of an arm which was previously possible with the expensive EMG sensors or laboratory equipments. The EMG sensor array of the armband reports the raw EMG measurements at a frequency of 200 Hz. The sample codes are provided by the official SDK [7]. In order to have access to the raw sensor measurements in a computation platform, we developed a Matlab library; MatMYO which is available online [8].

Fig. 1: MYO Armband Kit with Bluetooth Dongle

The MYO armband has a built-in gesture recognition software which can recognize six different gestures. These gestures are the Fist, Hand Relax (Free), Finger Spread, Wave In, Wave Out and Double Tap hand gestures Fig. 2.

Fig. 2: MYO Hand Gestures, 1- Fist, 2- Hand Relax, 3- Finger Spread, 4- Wave In, 5- Wave Out, 6- Double Tap (Thumb and middle finger tap each other two times)

We used the same gesture and hand posture set for comparison the accuracy.

3.2 Training

The recognition accuracy of the classification algorithms the SRC and CRC highly depends on the reliability of the representative samples in the training dictionary. Unlike the face recognition applications using these state of art classification methods, in the gesture recognition, obtaining the representative samples for the dynamical movements requires spotting the gestures, in other words, the start and end position of the gestures must be known in advance. Spotting gesture boundaries can be performed either by analyzing the signal if the boundaries are distinguishable or employing a switch to mark the boundaries while collecting gesture data. In this study and our previous studies [2], we employ a subspace clustering method to cluster the dictionary matrix its respective classes. The recent subspace clustering methods which are based on the self-representation property are capable of clustering the representative columns with a high clustering accuracy.

At least two successive hand gestures are necessary for clustering the data. For this reason, we collect training data by performing two successive gestures repeatedly in the training phase such as in the case for Wave In dictionary. The hand performs Wave In and Hand Relax gestures repeatedly for a short time Fig. 3.

In our experiments, we perform two gestures that takes one second 0.5 seconds for each of the hand states. Each gesture pair are repeated for only 10 seconds, but we use the first five seconds data to build a training dictionary.

A sliding window with a length of 100 data samples is used to capture the raw EMG sensor signal from the eight sensors and put in the dictionary matrix which is to be clustered as an 1D column vector. This procedures result

Fig. 3: Wave In -> Hand Relax -> Wave In Repetition

in a block circulant Hankel matrix which is the input of the subspace clustering method.

We use the Ordered Subspace Clustering (OSC) method [9] which meets the requirements of the continuous gesture recognition. The OSC method is based on the Sparse Subspace Clustering (SSC) [10] approach with an additional penalty term in its objective function for the sequential data. The additional penalty term in the objective function enforce the neighboring columns to be similar or close vectors. In our training matrices, the neighboring columns are time shifted and close to each other.

The training phase is easy to implement and take a short time to obtain a training dictionary for each of the gesture class. These classes are extracted from a gesture pair. In fact, when two gestures are clustered, two posture and two gesture sets are implicitly clustered into two class dictionaries. The hand switching between two gestures visits four hand states. In the case of the Wave In and Hand Relax training pair, the hand stands still at the relaxed position for a relatively short time, then performs Wave In gesture, stays at the Wave In position for a relatively short time then returns to the beginning and the whole cycle starts again. The representatives of the two postures are included in each dynamic gesture sets.

The continuous gesture recognition algorithm recognizes 10 dynamic hand gestures. However returning the original position from each of the hand gestures given in the Fig. (2) are mapped to hand relax posture. The results of the subspace clustering phase for a single gesture pair Double Tap and Hand Relax is given in Fig. (4) which demonstrates a successful clustering of the pair into the its respective EMG clusters.

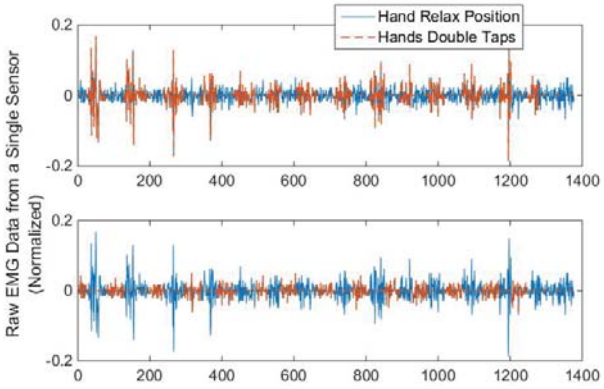


Fig. 4: Clustered Double Tap - Hand Relax Gestures on EMG data

4. Simulation Results

We collect five experimental data sets for each gesture pairs and use one set to build the class training dictionary. The remaining ones are used for testing. In addition to the gesture pairs, we collected data for a hand gesture sequences in which the hand visits all hand gesture and posture states arbitrarily. The results are given in the Figs. (5) - (10) from for each gesture pairs including one of the arbitrary hand gesture sequence experiment.

The recognition accuracy for the Fist and Hand Relax gesture states is 100% Fig. (5) for 1200 labelling computations. The algorithm runs at every sampling time in the experiments. There is no single error for the Fist and Hand Relax gestures in the Spectral CRC results.

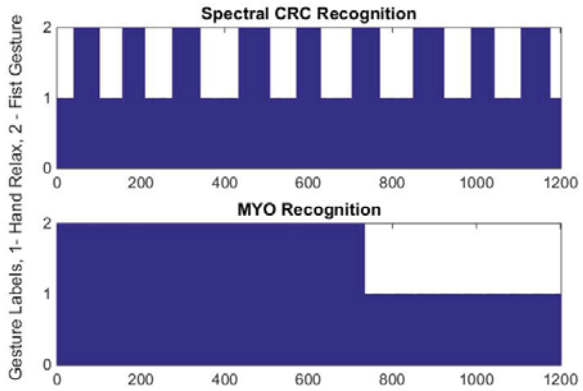


Fig. 5: Recognition Results for Fist and Hand Relax Experiment, Spectral CRC and MYO

The worst recognition accuracy result for the Wave Out - Hand Relax experiment is 98.34% out of 1274 classification computation. We give the results of this experiment without assigning the return gestures to the Hand Relax position from the Wave Out gesture in Fig. (6).

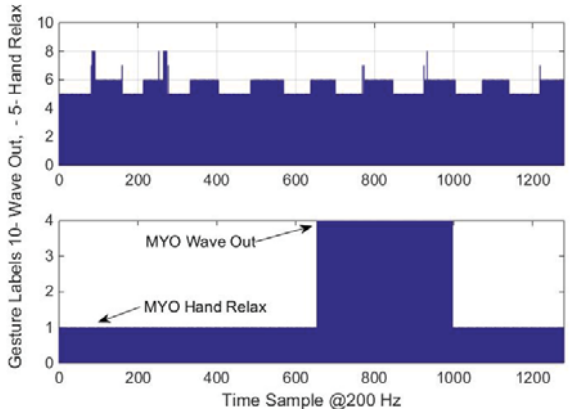


Fig. 6: Recognition Results for Wave Out and Hand Relax, Spectral CRC and MYO

Fig. (6) shows that, the algorithm can classify the return gestures from Wave Out gestures to Hand Relax position. In addition, the misclassifications occur on the change borders where hand switches between the gesture pairs. The method gives the Hand Spread labels at these change points where the hand involuntarily might perform Fingers Spread.

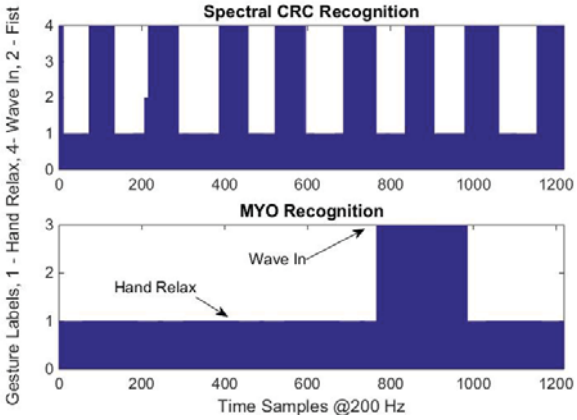


Fig. 7: Recognition Results for Wave In and Hand Relax, Spectral CRC and MYO

The results of a Wave In - Hand Relax experiments are given in Fig. (7). The recognition accuracy for this experiment is 99.34% while the recognition accuracy is 100% for Double Tap experiments Fig. (8).

The recognition accuracy for the Hand Spread-Hand Relax experiment is 97.3%. As shown in Fig. (9), the Spectral CRC method mis-classify the gesture as Wave Out while it is performing hand spread. This is due to the overlapping states of the Wave Out and Hand Spread gestures on which the same muscle groups activated and Wave Out hand gesture trajectory encompasses the Hand Spread wrist trajectory. We give the result of an arbitrary hand gesture sequence experiment in Fig. (10). The recognition accuracy is 98.47% for this experiment. In these experiments, unlike the others

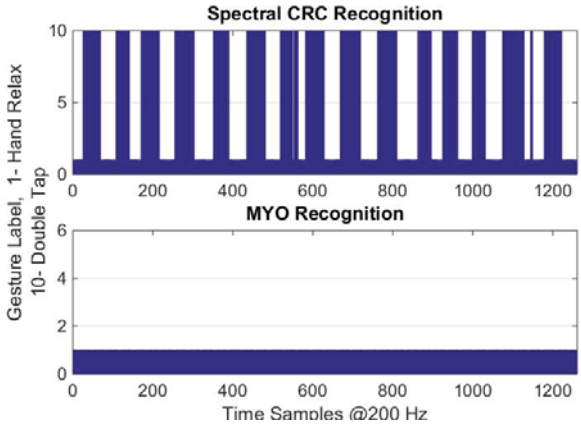


Fig. 8: Recognition Results for Double Tap and Hand Relax, Spectral CRC and MYO

in which only two gestures are performed repeatedly, the hand performs each gesture arbitrarily.

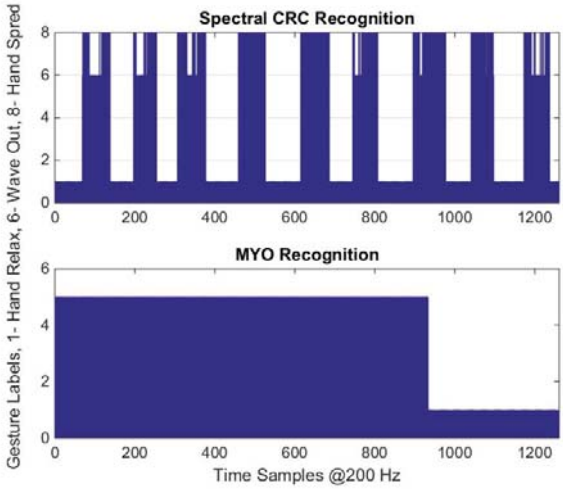


Fig. 9: Recognition Results for Hand Spread and Hand Relax, Spectral CRC and MYO

5. Conclusion

In this study we introduced a new signal pattern classification algorithm and methodologies to recognize a set of hand gestures and postures continuously on a multi-channel streaming signal using the raw EMG data. The initial results are promising in yielding high accuracy for a fairly rich hand gesture sets. It is worth to note that no signal pre-processing has been used in the study and the simulations are completed using only the raw data.

The methods and experiments explained in this paper is a small portion of our main project [11, 12] by which we have been developing multi-modal HCIs for elderly people to enable them to command a robotic wheelchair with their available resources. In our previous studies such

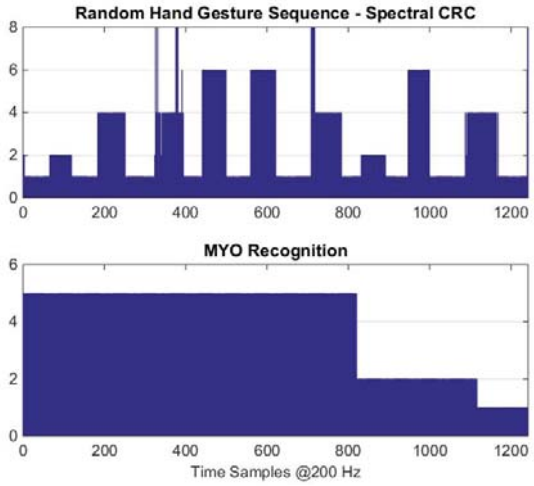


Fig. 10: Recognition Results for Random Hand Gesture Sequence, Spectral CRC and MYO, 1- Hand Relax, 2- Fist, 4- Wave In, 6- Wave Out, 8- Hand Spread

as the braking state classification of a mobility robot [2], Segway using a tablet PC and inertial sensors, we showed that, the CRC method is capable of labeling braking states of the robot with a high accuracy real-time and fast. We also verified the training approach and the CRC method for continuous hand gesture and posture recognition using the Leap Motion sensor which tracks the hand with optical cameras at sub-millimeter levels and achieved very high recognition accuracies more then reported in this paper with the same gesture sets. We will finalize the project with the developed gesture and posture recognition methods devising experiments to which elderly people will participate. In addition, with the developed interfaces, we will prepare a virtual reality training and rehabilitation environment for the elderly and the individuals with severe disabilities who are prescribed power wheelchairs. This is due to the fact that most people who are prescribed a power wheelchair experience difficulty to use them and it takes time to adapt to the new technologies.

6. Acknowledgments

The study is supported by the Japan Society for the Promotion of Science (JSPS) fellowship program and the KAKENHI Grant (Grant Number 15F13739).

References

[1] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," *arXiv preprint arXiv:1204.2358*, 2012.

[2] A. Boyali, N. Hashimoto, and O. Matsumoto, "A signal pattern recognition approach for mobile devices and its application to braking state classification on robotic mobility devices," *Robotics and Autonomous Systems*, 2015.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[4] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 471–478.

[5] S. Shahid, J. Walker, G. M. Lyons, C. A. Byrne, and A. V. Nene, "Application of higher order statistics techniques to emg signals to characterize the motor unit action potential," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 7, pp. 1195–1209, 2005.

[6] D. S. G. Pollock, "Circulant matrices and time-series analysis," *International Journal of Mathematical Education in Science and Technology*, vol. 33, no. 2, pp. 213–230, 2002.

[7] Thalmic Labs. (2015) Myo sdk. [Online]. Available: https://developer.thalmic.com/docs/api_reference/platform/the-sdk.html

[8] A. Boyali. (2015) Matmyo. [Online]. Available: https://github.com/boyali/matMYO

[9] S. Tierney, J. Gao, and Y. Guo, "Subspace clustering for sequential data," in *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1019–1026.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[11] N. Hashimoto, Y. Takinami, and O. Matsumoto, "An experimental study on vehicle behavior to wheel chairs and standing-type vehicles at intersection," in *ITS Telecommunications (ITST), 2013 13th International Conference on*. IEEE, 2013, pp. 350–355.

[12] A. Boyali and N. Hashimoto, "Block-sparse representation classification based gesture recognition approach for a robotic wheelchair," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, 2014, pp. 1133–1138.

Alert System for Securing Lost items

Mohammed Alsaadi, Khalid Babutain, and Sreela Sasi
Department of Computer and Information Science
Gannon University
Erie, PA 16541, USA

Abstract— Currently, most institutions, such as colleges and universities, have installed cameras in major critical areas for surveillance. However, these cameras lack an automated system or software that can support them for detection and recognition. Therefore, the Alert System for Securing Lost items is developed to monitor the hallways continually to detect incidences of lost or forgotten bags. It not only automatically detects a lost or forgotten bag but also will identify the owner of the bag, and will inform him/her through an e-mail. If the owner is not a university employee or student, it will report to the security personnel about the lost or forgotten bags with its owner's image in an e-mail. Additionally, in the case of a theft situation, the system can detect the person who stole the bag and inform the security personnel by sending an e-mail with that person's image attached for necessary action.

Index Terms— object detection and recognition; surveillance; hallway monitoring; lost or forgotten bag; face detection and recognition

I. Introduction

Nowadays, one of the most essential aspects of private and public security is the use of automatic surveillance systems. These systems process images and videos. Thus, face recognition, object recognition, and object detection are some of the systems used [1], [2]. However, some features from the aspect of image and video processing are fundamentally problematic. Object detection and recognition are required to find the existence of an object within an image [3], besides the need to provide its location [4].

On the other hand, face recognition includes identifying and locating a human face within an image. In some systems, the algorithms used for face recognition extract some specific features of the detected face. Most of such systems have a database of predefined facial features, which are then compared with the detected features from the new image to find the best match [5], [6].

At this time, there may or may not be any surveillance cameras installed in a university hallway. Assuming that one is set up in a hallway, the camera would not be installed with a software program that would automatically detect any lost or forgotten bag in that hallway. Such a camera's purpose would only be to stream the video of the hallway to the surveillance or control room where a person must be available to monitor the video. The security system might be recording this video, but any automatic searching or recognition functions would be unavailable on it. If the person in the control room would see an unattended bag through the video stream, the person leaving the bag behind in the hallway would be checked by watching the recorded video from the start, which would take a lot of time. Such an arrangement would also not ensure that the person leaving the bag would be identified. This manual system would highly depend on the memory of the observer in the control room who would be unable to recognize all the students.

This paper presents the '*Alert System for Securing Lost items (ASSL)*' as an effective solution to such issues. This system depends on image and video processing, which includes detecting and recognizing bags, as well as identifying and locating the faces of people present on the surveillance scene. Image and video processing aims to identify the presence of any specific bag by using a prebuilt bag classifier [7] that also provides the location of that bag. The algorithms used for face recognition extract the features of the face being detected. These features are compared with the ones that are already present in a predefined database containing the facial features of the university students to find the most matched face.

Thus, the ASSL automatically scans the hallway, which reveals the identity of the person leaving a bag there and immediately alerts that person about his or her forgotten bag. It also promptly notifies the staff in the control room if the person is not recognized as a university student.

Section II describes the existing surveillance systems, which also provides a clear background for this research. Section III proposes the system's architecture, including the functionality of its components. Section IV provides the details of the implementation of the system's prototype model, along with the experimental results.

II. Background Research

Surveillance systems are designed to provide detailed information from the environment they monitor, which may include moving objects or people. In one of these systems [8], a software architecture was proposed that could detect and track persons in motion and in a video sequence. It could specify their profiles as they entered a specific region from another controlled area. Therefore, by using this system, an alarm could be generated. For example, if a person who was already being monitored would enter a prohibited area or stay there longer than a predefined length of time, then the alarm would be activated.

A surveillance system for ramp operations in airports was proposed by [9]. Two algorithms were formulated and

implemented to localize and detect an aircraft. The basis of the very first algorithm was background subtraction concept, so any changes in video streams could be detected. For the second algorithm, a supervised-learning technique was used, so by using a database of images, the system could learn a model. These algorithms were implemented in a sample version with a realistic 1:400 scale model of an airport ramp area. The aircraft localization's accuracy was in a range of 30 ft.

Another surveillance framework was proposed by [10] for detecting motorbike theft actions, which combined object detection technology and recognition of human activities. To reduce the number of objects needed to be processed, [10] estimated a specific region of interest in the input video. Additionally, to detect the theft actions, a predefined dataset was built by analyzing the activity sequences of thieves. According to [10], they had positive outcomes in their experiments: "Our proposed framework works well on the reality dataset; it proves to be a feasible and applicable solution."

One of the other systems was designed to track an object after being moved or placed in a new position within the same room [11]. This system includes several models, such as 3D room model initialization, object detection, object monitoring and management, foreground extraction, background modeling, and human detection and removal. Besides a responsive time delay, this application was successfully completed.

III. System Architecture

The main goal of the ASSL is to build a surveillance application that immediately notifies any recognized student about his or her forgotten bag, as well as the security department in case the person is not recognized as a university student.

The structure of the data in the system will mainly utilize three major types of data. This data is designed for the system with the aim to operate for achieving its objective. The video scenarios taken as inputs for the system testing will have a resolution of 640 x 480 pixels. These videos are in the Audio Video Interleaved (AVI) type of file format.

Another data type used by the system will be the Extensible Markup Language (XML). This type of file can be generated by the Training a Cascade Object Detector [7] System, which will take a large number of negative and positive images that will be processed by using the Viola-Jones algorithm [12], [7]. The positive images received are those containing the object that is required to be detected, which is the bag in this case, whereas the negative images are those that do not contain any of the desired objects. The Training a Cascade Object Detector System will be able to generate the corresponding XML file for the detection of bags that will be carried out through the processing of the system.

The third data type to be used comprises the Portable Gray Map (PGM) images. This involves the most critical part of face recognition, which is based on the Eigenfaces technique. It utilizes a highly efficient encoding technique that is followed by the comparison of each face with a database of faces that are also encoded with the same technique [5], [6].

The ASSL contains a number of folders, each including 10 PGM images comprising different characteristics of facial features for the identification of a student. During the processing of the system, these folders are being used as its database.

Whenever the system begins its operation, and the targeted video is loaded in it, the video will be processed in a frame-by-frame manner. Every frame will be loaded in either a two- or three-dimensional array, depending on the processing functionality. The grayscale images will require data storage in a two-dimensional array, whereas the colored images will require a three-dimensional array.

To this end, the software designed will work on two major principles. First, it will send an email [13] to the student who has left his or her bag, along with the picture and location of the bag. Second, in case of a detection of a lost or forgotten bag whose owner is not recognized as a university student, the software will notify and send an email to the security department, along with a picture of the bag and that of the person who left it. Figure 1 shows the ASSL architecture.

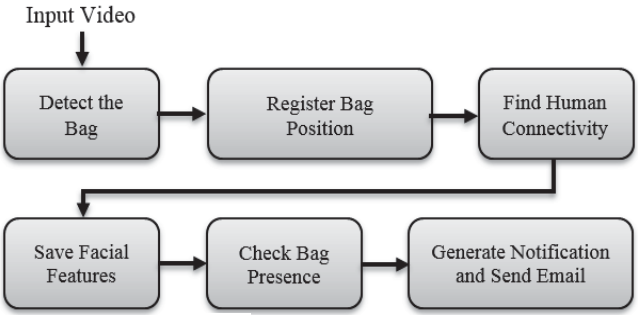


Fig. 1. ASSL System Architecture

A. Detect the Bag

The ASSL will perform automatic surveillance of a university hallway, using a trained bag classifier. This classifier has been trained for three different bags, which are used for the system testing. The system detects these bags by using the classifier used by the Cascade Object Detector System [12], which is also based on the Viola-Jones algorithm. By applying the custom-trained bag classifier with the Cascade Object Detector System on the input video, the coordinates and location of the detected bag are provided in a two-dimensional array containing the x and y axes of the upper-left corner of the detected bag, along with its width and height. This process is applied on every frame of the tested video in which this two-dimensional array is either not provided when there is no detection of a bag or changed based on the location of the bag on the surveillance scene.

B.Register Bag Position

When there is continuous detection of a bag for a predefined time, another two-dimensional array with the same structure as that of the two-dimensional array being provided by the Cascade Object Detector System is generated. This array also has x and y coordinates, along with a width and height, but it is stored after some calculations are performed, using the coordinates provided from the bag detection. This new array's purpose is to register the bag position as a bounding box that contains the detected bag inside it. This two-dimensional array is stored after subtracting 10 pixels from both of the provided x and y coordinates and adding 20 pixels to both of the provided width and height. A fixed bounding box is then drawn around the detected bag.

Figure 2 shows a detected bag with a yellow bounding box associated with the detection provided by the Cascade Object Detector System. The green, fixed bounding box represents a visible bag after the predefined time duration of a continuous detection.

Fig. 2. Green, Fixed Bounding Box

C. Find Human Connectivity

By the time the fixed bounding box has been registered and established, the system goes backward through the video for a predefined number of frames, 30 frames in this case, to find the connectivity between the bag and its owner. This process is also based on using the Cascade Object Detector System to detect the bag owner's face. However, rather than using the trained bag classifier, the Cascade Object Detector System is employed with its default value, which applies a built-in classifier that detects any visible human face on the scene and provides the coordinates and location of the detected face. Then by having the coordinates and locations of both the bag and the human, the connectivity can be established by calculating the distance between them. Thus, the nearest face to the detected bag is identified as the bag's owner.

Figure 3 shows the detection of the bag and its owner, with an illustration of the connectivity establishment.

Fig. 3. Establishing Human Connectivity

D. Save Facial Features

After the detected bag's owner is identified, the person's facial features are saved temporarily. The system continues to monitor the presence of the bag's owner and compares this person's temporarily saved facial features with those of any detected face, including those of the bag owner. In this case, a match is found, the owner is still on the scene, and the system continues monitoring the hallway. Figure 4 illustrates this process.

Fig. 4. Saving Facial Features and Checking Presence of Bag's Owner

E.Check Bag Presence

In case the system does not find a match of the temporarily saved facial features on the surveillance scene, it then checks for the bag's presence. This is done by checking the Cascade Object Detector System that uses the bag classifier to determine whether or not it is still providing coordinates for a bag. If it does not provide any data, this means that the bag is no longer present, and the owner has taken his or her bag and left the scene. The green, fixed bounding box is then discarded. Otherwise, if the system provides the data, it recognizes the bag's presence without its owner (which is done by calculating the center of the detected bag, using the provided coordinates) and compares it with the coordinates of the fixed bounding box. Figure 5 shows the owner taking his bag and leaving, so the fixed

bounding box is discarded. On the other hand, Figure 6 shows him leaving without his bag.

Fig. 5. Owner Takes his Bag and Leaves

Fig. 6. Owner Leaves without his Bag

F. Generate Notification and Send Email

In case the when the system recognizes the bag's presence without its owner, the system immediately changes the green, fixed bounding box to a red color. It then uses the temporarily saved facial features of the bag's owner to compare these with the predefined, student facial database. If a match is found, the system takes a picture of this bag in the hallway and immediately sends an email, which includes the picture, to the student as a notification about his or her forgotten bag. Figure 7 shows that the owner has left the scene without his bag, and the fixed bounding box has turned red.

Fig. 7. Owner Leaves without his Bag

In case there is no match, this means that the bag owner is not a university student. Thus, the system goes backward through the video and takes a picture of that bag with its owner and immediately sends an email, including the picture, to the university's security staff to notify them about this lost or forgotten bag. Such a scenario is depicted in the simulation section of this paper.

IV. Simulation Results

The ASSL was simulated with seven different recorded scenarios, which were included in the testing phase to verify the system behavior. Table 3 describes each of the recorded scenarios.

TABLE 1: SCENARIOS AND DESCRIPTIONS

Scenario	Scenario Description
1	A student sits in the hallway and puts his laptop bag next to him. Then he takes his bag and leaves the hallway.
2	A student sits in the hallway and puts his backpack next to him. Then he takes his backpack and also leaves the hallway.
3	A student sits in the hallway, puts his backpack next to him, and then leaves the scene without his bag.
4	A student sits in the hallway next to another student and puts his bag next to him. Then the student who owns the bag leaves the hallway without his bag.
5	Two students sit in the hallway, and then one of them leaves without his bag. After that, the other student takes the forgotten bag.
6	Three students are on the scene including the bag owner in which in this case is a different student. The student leaves his bag, while another takes it and exits the scene.
7	A nonstudent sits in the hallway, puts his bag next to him, and then leaves the scene without his bag.

During the testing of the scenarios, the system responded exactly as expected, indicating the fulfillment of the

system's development objectives. The responses obtained from each of the seven scenarios, as well as the screenshots of the system testing, are described below.

In the testing of Scenario 1 and Scenario 2, the student enters the scene where he sits and puts his bag next to him—a laptop bag in the first scenario and a backpack in the second. In both scenarios, the system detects the bags and draws the green, fixed bounding box around them. Then the student takes his bag and leaves the scene. The system does not take any action other than discarding the fixed bounding box and continuing to monitor the hallway until the tested scenarios end. Figures 8.1 and 8.2 show screenshots of Scenario 1; Figures 9.1 and 9.2 depict those of Scenario 2.

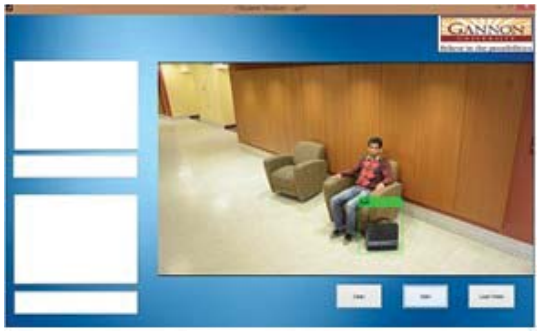


Fig. 8.1. Scenario 1

Fig. 8.2. Scenario 1

Fig. 9.1. Scenario 2

Fig. 9.2. Scenario 2

In Scenario 3 and Scenario 4, the system takes a picture of the backpack in the hallway and immediately sends an email, including the picture, to the student concerned as a notification about his forgotten bag. Although Scenario 4 includes more than one person, the system's behavior was not affected. Figures 10.1 and 10.2 illustrate Scenario 3, while Figures 11.1 and 11.2 present Scenario 4.

Fig. 10.1. Scenario 3

Fig. 10.2. Scenario 3

Fig. 11.1. Scenario 4

Fig. 11.2. Scenario 4

In Scenario 5, the system recognizes the student who forgets his bag and sends him an email, with a picture of his bag in the hallway, to notify him about his forgotten bag. Another person is also present on the scene where he takes that bag right after its owner leaves. In this instance, the system sends another email to the security department, including the picture of that person who took the bag, to notify them about his action of taking the bag that might not belong to him. Figure 12 shows the system response.

Fig. 12. Scenario 5

The sixth scenario is no different than the fifth unless the owner of the bag is a different student, and the system recognizes him when he forgets his bag, and it sends him a notification email right after he leaves the scene. Also, the number of persons within the scene other than the bag owner is 2 in which one of those persons takes that bag right after everyone leaves the area. At this time, the system also sends another email to the security department, including the picture of that person who took the bag, to notify them about his action of taking the bag that might not belong to him. Figure 13 shows the system response.

Fig. 13. Scenario 6

In Scenario 7, a person who is not a student of the university forgets his bag and leaves. The system sends an email to the security department, including the picture of that nonstudent who left his bag in the hallway. Figure 14 shows the system response.

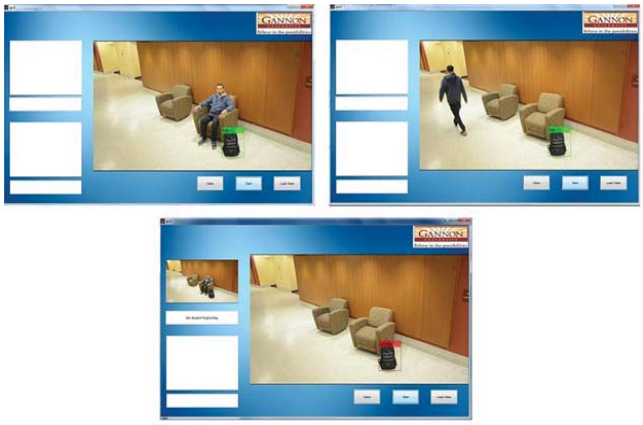


Fig. 14. Scenario 7

As witnessed in the seven given scenarios, the system performs its function by identifying a lost bag or continuing to display the video. In case of a forgotten or lost bag, the system recognizes and sends the necessary information to the responsible authorities or individuals, in this case, the security personnel and the students, respectively. Table 2 presents a summary of the scenarios. The tested scenarios are answered with *yes* or *no*, depending on whether or not the bag is detected, as well as the recipients of the notification emails.

TABLE 2: RESULTS

Sce-nario	Bag detec-tion	Identifi-cation of lost bag	Detection of stranger taking the bag	Notifica-tion email to student	Notifi-cation email to security
1	Yes	No	No	No	No
2	Yes	No	No	No	No
3	Yes	Yes	No	Yes	No
4	Yes	Yes	No	Yes	No
5	Yes	Yes	Yes	Yes	Yes
6	Yes	Yes	Yes	Yes	Yes
7	Yes	Yes	No	No	Yes

V. Conclusion and Future Work

The design and implementation of the *'Alert System for Securing Lost items (ASSL)'* are done successfully. The features included are detecting a bag, establishing human connectivity to the bag, and identifying a forgotten bag if applicable. It will also notify either the student or the security department based on the owner's actions in relation to his or her bag. All these system features have been tested using seven different scenarios and found to be promising. The system will fail if the bag is behind another object or invisible to the camera or if the face of the owner of the bag is not within the vicinity of the camera. This system could be used in any university hallway after extensive testing, and could be extended for other situations as well. For future work, this system could be further developed to be used for real-time video streaming.

References

[1] A. Ben Hamida et al., "Toward scalable application-oriented video surveillance systems," in *Proc. Science and Information Conf., 2014*, pp. 385–388. Red Hook, NY: Curran Associates.

[2] S. Sahra and S. Neogy, "A case study on smart surveillance application system using WSN and IP webcam," in *Proc. Applications and Innovations Mobile Computing,* 2014, pp. 36–41. Institute of Electrical and Electronics Engineers.

[3] J. Li et al., "Learning SURF cascade for fast and accurate object detection," in *Proc. 4th Nat. Conf. Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013, pp. 3468–3475. Institute of Electrical and Electronics Engineers.

[4] Y. Li et al., "Learning cascaded shared-boost classifiers for part-based object detection," *IEEE Trans. Image Process.,* vol. 23, pp.1858–1871, 2014.

[5] L. An et al., "Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks," *IEEE J. Emerging and Sel. Topics Circuits and Syst.*, vol. 3, no. 2, pp. 155–164, 2013.

[6] A. Chowdhury and S.S. Tripathy, "Human skin detection and face recognition using fuzzy logic and eigenface," in *Proc. Green Computing Communication and Electrical Engineering,* 2014, pp. 1–4. Institute of Electrical and Electronics Engineers.

[7] *Train a Cascade Object Detector* [Online]. Available: http://www.mathworks.com/help/vision/ug/train-a-cascade-object-detector.html

[8] A. Ezzahout and R.O.H. Thami, "Conception and development of a video surveillance system for detecting, tracking and profile analysis of a person," in *Proc. ISKO-Maghreb, 3rd Int. Symp.*, 2013, pp. 1–5.

[9] S. Vaddi et al., "Computer vision based surveillance concept for airport ramp operations," in *Proc. Digital Avionics Systems Conf., IEEE/AIAA 32nd*, 2013, pp. 3D2-1–3D2-13.

[10] D. Mai and K. Hoang, "Motorbike theft detection based on object detection and human activity recognition," *Control, Automation and Information Sciences*, 2013, pp. 358–362.

[11] T. Chitmaitredejsakul et al., "Multiple objects monitoring based on 3D information from multiple cameras," *Electrical Engineering Congr.,* 2014, pp. 1–5.

[12] *Vision. Cascade Object Detector System object* [Online]. Available: http://www.mathworks.com/help/vision/ref/vision.cascadeobjectdetector-class.html

[13] *Sendmail* [Online]. Available: http://www.mathworks.com/help/matlab/ref/sendmail.html

Image Optimization under APC Constraint

CAI Tiefeng[1,2,3], **ZHU Feng**[1,3], **HAO Yingming**[1,3], **FAN Xiaopeng**[1,2,3]
1. Shenyang Insititute of Automation, Chinese Academy of Sciences, Shenyang , China;
2. University of the Chinese Academy of Sciences, Beijing,China;
3. Key Laboratory Opto-Electronic Information Processing, Chinese Academy of Sciences, Shenyang , China

Abstract - *Image information for objects detection and recognition(IDR) is mainly reflected in the gray value relationship of adjacent pixel pairs. Under adjacent pixel pairs' gray value preserving constraint(APC), an optimization method is designed to Max Min perception degree of adjacent pixel pairs. Experimental results show that perception degrees of adjacent pixel pairs have been significantly improved.*

Keywords: Image optimization, objects detection and recognition, image enhancement

1 Introduction

When human watch images, if gray value difference between the objects and background is too small, human eyes will not be able to distinguish objects from the background. Image optimization increase human eyes' perception degree of gray difference between the objects and background through transforming image gray value. Therefore human eyes can correctly perceive the existence of the objects. Thus image optimization plays an important role in assisting human detecting and recognizing objects.

IDR is reflected in the relationship between pixels gray value including greater-than, less-than, and equal-to three relations, rather than reflected in the pixel gray values themselves[1]. Pixel pairs can be classified into adjacent pixel pairs and non-adjacent pixel pairs. Since scene outline and texture is directly related to the relationship between adjacent pixel pairs. Adjacent pixel pairs' relationship are more important than non-adjacent pixel pairs for human detecting and recognizing objects.

To maintain IDR , it is needed to ensure that gray value relationships of all pixel pairs are preserved. But in order to make as much as possible adjacent pixel pairs' gray value relationship good perceived by human eyes, this article will give up preserving the relationship between non-adjacent pixels, only optimize images under APC.

There are few image optimization method under APC[7]. Kartic Subr et al proposed an optimization method under APC, this method takes maximizing image local contrast as optimization objectives, and optimize images based on greedy algorithm[2]. This method can improve the image contrast, but not completely optimize the image for human eyes detection and recognition.

The minimum perception degree of adjacent pixel pairs is required to be maximized to Max Min adjacent pixel pairs' perception degree[1]. however, there are many optimized images are satisfied with this condition. It is further required to maximize the second minimum perception degree , maximize the third minimum, and so on. For this optimization objective, under APC, an optimization method is proposed in this paper.

2 Human vision characteristics

When the adjacent pixel pair's gray values are identical, human eyes can perceive the equal-to relationship, but when the difference of adjacent pixel pairs is non-zero, human eyes will not always perceive the presence of gray value difference.

Perception degree of adjacent pixel pairs' gray value difference can be calculated out by the two human visual characteristics.

1) Detection probability function. Foly and Legge in 1981 measured detection probability function[3]. The function returns the probability that human correctly have detected the gray value difference. The function is expressed as formula (1)[3-4]. s_0 is the gray value difference at which the probability human eyes correctly detecting the difference is 50% . Value of s_0 / δ ranges from 2.5 to 4.

$$p(s) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{(x-s_0)^2}{2\sigma^2}} dx \qquad (1)$$

2) JND Curve. JND is the just noticeable difference of human eyes[5]. At 1995， Chun-Hsien Chou et al experimentally draw a JND curve for a monitor[6]. Although for different monitors and lighting conditions JND measured values will be different, but the JND curve can be measured by the method proposed by Chou. This paper will study image optimization with the JND curve measured out by Chou[6]. This curve is approximated to formula (2). $T_0 = 17$, $\gamma = 3/128$. Essentially, the JND is the s_0 in detection function.

$$J(x)=\begin{cases} T_0\cdot(1-(x/127)^{1/2})+3 \\ for \quad x\le 127 \\ \gamma\cdot(x-127)+3 \\ for \quad x>127 \end{cases} \qquad (2)$$

JND curve and detection probability function make up an imperfect but simple expression of gray value difference's perception degree. The expression is showed as formula (3). a and b are gray values of adjacent pixel pair, and at default $a<b$. s_0/δ is set to 3.

$$p(a,b)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{b-a} e^{-\frac{(x-J(a))^2}{2\sigma^2}}dx \qquad (3)$$

3 Image optimization method

The gray values of adjacent pixels with identical gray value in the original image inevitable remains identical in the optimized image under APC. Therefore the adjacent pixel grays with identical gray value are grouped into a primitive called S-pixel.

The proposed optimization algorithm has three steps:
1)Extreme points assignment;
2)Max the minimum perception degree.
3)Max the minimum perception degree of path between the assigned super-pixels

2	51	74	60
50	73	89	122
72	88	121	136
87	120	135	210

→

?	?	?	?
?	?	?	?
?	?	?	?
?	?	?	?

(a) (b)

Fig. 2. (a) is the original image, (b) is the unknown optimized image.

The following describes the meaning of each step in the example of Fig. 2.

3. 1 Extreme points assignment

Extreme points include maxima and minima. Maxima is the S-pixel of which the gray value is greater than his adjacent S-pixels in the original image. Minima are the S-pixel of which the gray value is less than the adjacent S-pixels in the original image. Under APC, to max the perception degrees of the gray value difference between the extreme S-pixels and their adjacent S-pixels, the gray value of all maxims are assigned to the maximum gray value 255, the gray value of all the minima are assigned to the minimum value 0 in the optimized image as shown in Fig. 6, the circled S-pixels in the original image are extreme points.

2	**5**	**74**	**60**
4	**7**	**89**	**122**
8	**88**	**121**	**124**
87	**120**	**123**	**210**

→

0	**?**	**?**	**0**
?	**?**	**?**	**?**
?	**?**	**?**	**?**
?	**?**	**?**	**255**

(a) (b)

Fig. 3. (a) is the original image, (b) is the unknown optimized image with extreme points assigned

3. 2 Max the minimum perception degree

Suppose $P=<p_1,p_2,p_3,\cdots p_N>$ is a S-pixel sequence, where $I_1(p_i)$ is the gray value of p_i in the original image. If p_i is adjacent to p_{i+1} in the image and $I_1(p_i)>I_1(p_{i+1})$, $i=1,2,3,\cdots,N-1$,then sequence P is called an path. The total amount of S-pixels in a path is called the length of the path.

In the unknown optimized image, there are many paths from the maxima S-pixels to the minima S-pixels. Under APC, the length of the path are larger, it is more difficult to stretch the perception degree between the adjacent S-pixels. To Max the minimum perception degree of the unknown optimized image, it is needed to max the minimum perception degree of the longest path(see Fig. 4). Under APC, it will max the minimum perception degree of the longest path that stretching the longest path with uniform perception degree interval between adjacent S-pixels. The fourth part of this article will give the gray value assignment method for uniform perception degree interval stretching.

2	**5**	**74**	**60**
4	**7**	**89**	**122**
8	**88**	**121**	**124**
87	**120**	**123**	**210**

→

0	**?**	**?**	**0**
137	**?**	**?**	**?**
155	**?**	**?**	**?**
174	**199**	**224**	**255**

(a) (b)

Fig. 4. Image optimization process (1),(a) is the original image, (b) is the unknown optimized image with longest path assigned.

3. 3 Max the minimum perception degree of path between the assigned super-pixels

From already assigned S-pixels to other assigned S-pixels, there are many paths. Calculate the uniform perception

degree interval of all these paths. Find out the path with the smallest uniform perception degree interval. Assign the path in uniform perception degree interval(see Fig. 5).

After the assignments of the path with the minimum uniform perception degree interval, it is needed to return to the beginning of the procedure (see Fig. 6) until all of the S-pixels in the image have been assigned gray values(see Fig. 7). The original image in Figure 8 is the gray image of the original image in Figure 4, while the optimized image in Figure 8 is obtained by the method.

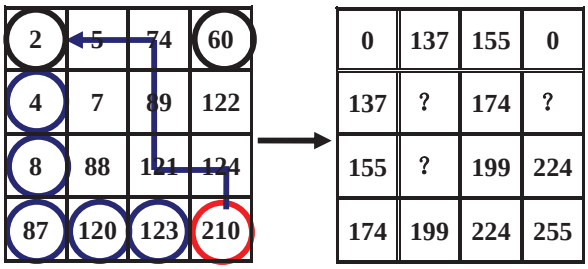


Fig. 5 Image optimization process (2)

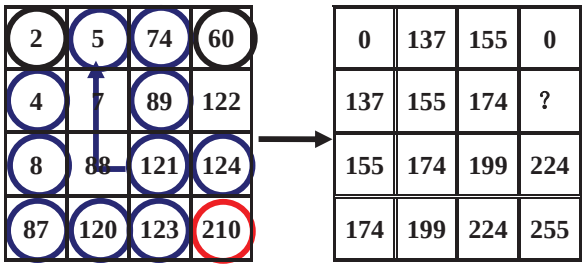
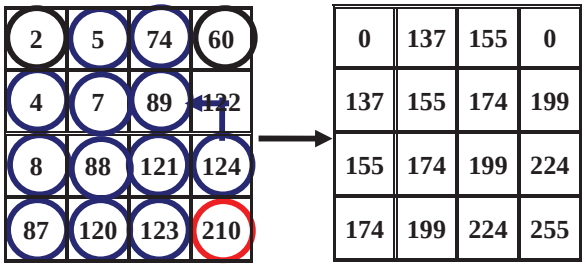
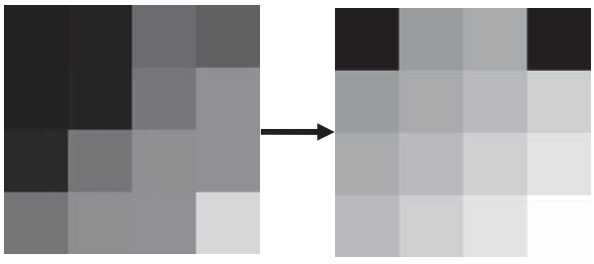


Fig. 6 Image optimization process (3)



(a) (b)

Fig. 7. (a) is the original image, (b) is the optimized image.



(a) (b)

Fig. 8.(a) is the gray image of Fig.7.(a), (b) is the gray image of Fig.7.(b)

4 Gray value assignment with uniform perception degree interval.

The gray values of S-pixels u_0 and u_m are g_0 and g_m respectively. the uniform perception degree interval and the assignments of all element except u_0 and u_m of path $U =< u_0, u_1, \cdots, u_m >$ are needed to be solved out. The unassigned gray value of u_i is denoted by $I_2(u_i)$,where $i = 1, 2, \cdots, m-1$.

Because $$p(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{(x-s_0)^2}{2\sigma^2}} d(\frac{x}{\sigma})$$

then $$p(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{(\frac{x}{\sigma}-\frac{s_0}{\sigma})^2}{2}} d(\frac{x}{\sigma}).$$

Assign $s_0 / \delta = 3$, then $p(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{3s}{s_0}} e^{-\frac{(x-3)^2}{2}} d(x).$

Therefore, $p(s) = h(\frac{s}{s_0})$, where $h(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3t} e^{-\frac{(x-3)^2}{2}} d(x)$.

When the path U is stretched with uniform perception degree interval,

$$\frac{s}{s_0} = \frac{I_2(u_{i-1}) - I_2(u_i)}{J(I_2(u_i))} = B \qquad for\ i = 1, 2, \cdots, m \tag{4}$$

where B is an constant. Then

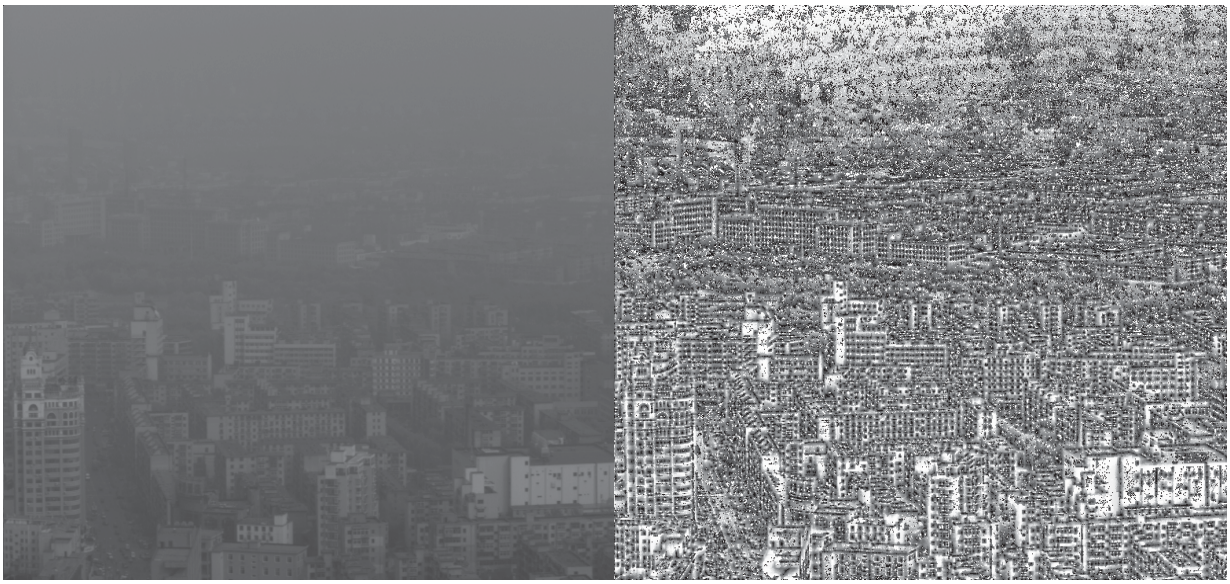
$$p(R_1(A_{j-1}), R_1(A_j)) = C \qquad j = 1, 2, \cdots, m \tag{5}$$

Set M_0 to be the largest JND value of collection $U =< u_0, u_1, \cdots, u_m >$, while set M_1 to be the lest JND value of the collection. Set $l = \frac{A_0 - A_m}{m \times M_0}$ and $u = \frac{A_0 - A_m}{m \times M_1}$,then $B \in [l, u]$.

The value of B can be searched through these steps:

1) Create a sequence $< g_{i,0}, g_{i,1}, \cdots, g_{i,m} >$, where $g_{i,0} = v_0$ and $g_{i,j} = (l+u) \times J(g_{i,j-1}) / 2 + g_{i,j-1}$ for $j = 1, 2, 3, \cdots m$.
2) If $g_{i,m} > v_m$, refresh the upper boundary value of domain $[l, u]$ with setting $u = (l+u)/2$, else refresh the lower boundary value of the domain with setting $l = (l+u)/2$.
3) If $|g_m - g_{i,m}| \le 0.0001$, set $B = (u+l)/2$, else go back to step 2) .

Then $C = h(B)$,the gray value of the super-pixels can calculated from formula(4)..

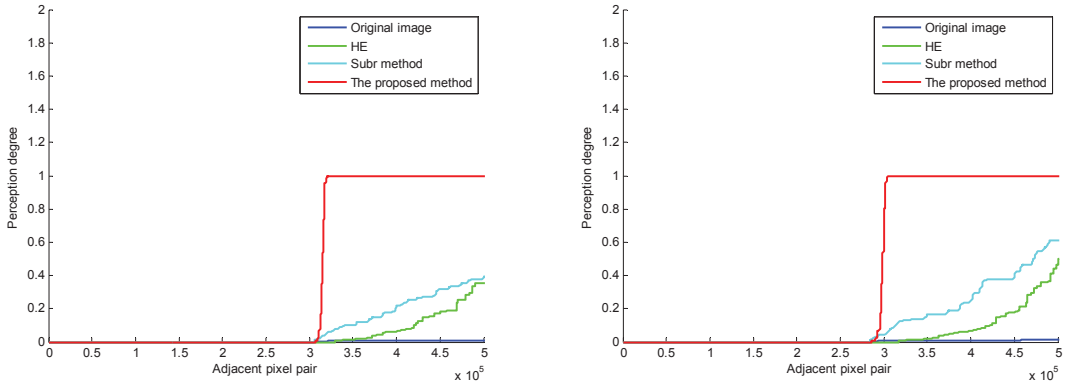


(a)Original image (b) Optimized image by the proposed method

Fig.9 image optimized by the proposed method

(a)Original image (b)Histogram equalization (c)The greedy method (d)The proposed method

Fig. 10. Methods comparison



(a) (b)

Fig 11. (a) is perception degree comparison of upper images in Fig. 10, (b) is the comparison of lower images.

5. Experiments

In this paper, adjacent relationship is defined in 4-nighbourhood. In Fig. 9, the left is the original image, and the right is the optimized image by the proposed method. The proposed method are compared with the Subr method and histogram equalization (see Fig. 10). The adjacent pixel pairs' perception degrees are sorted ascending respectively. In Fig. 11, the perception degree sequence is compared. It is clear that the perception degrees of the optimized images are significantly greater than the images processed by other two methods and the original image.

6.Discussion and future works

This paper proposed a image optimization method under APC. The optimization method maximize the minimum perception degree of adjacent pixel pairs. The perception degrees of adjacent pixel pairs have increased significantly in images optimized by the proposed method.

Without preserving non-adjacent pixel pairs' gray value relationship, optimized images turn out unnatural. The unnatural is unfavorable to the completion of the human eye detection and recognition tasks.

In the perception degree function, perception degree of gray value difference $s = J(x) + 3\delta = 2J(x)$ is 0.975, where $J(x)$ is the JND of gray level x and $3\delta = J(x)$. Therefore, when the gray value difference of adjacent pixel pair is $s > J(x) + 3\delta = 2J(x)$, perception degree of the difference approximates to constant 1. So it is rational that the gray value of extreme points are not need assigned 0 or 255 to maximize the perception degree of adjacent pixel pair. In the future work, it is possible to preserve non-adjacent pixel pairs' gray value relationship as much as possible after maximizing the minimum perception degree of adjacent pixel pairs.

Reference

[1] T. Cai, *et al.*, "A method to enhance images based on human vision property," in *2012 11th International Conference on Signal Processing, ICSP 2012, October 21, 2012 - October 25, 2012*, Beijing, China, 2012, pp. 952-955.

[2] A. M. a. S. I. Kartic Subr, "Greedy Algorithm for Local Contrast Enhancement of Images," *Image Analysis and Processing-ICIAP 2005,* 2005.

[3] J. M. F. a. G. E. Legge, "Contrast detection and near-threshold discrimination in human vision," *Vision Research,* vol. 21, p. 13, 1981.

[4] P. G. J. Barten, "Contrast sensitivity of the human eye and its effects on image quality," 1999.

[5] N. Jayant, "Signal Compression:Technology Targets and Research Directions," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS,* vol. 10, p. 23, JUNE 1992.

[6] C.-H. C. a. Y.-C. Li, "A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY,* vol. 5, p. 10, DECEMBER 1995.

[7] Liu Jian xin. A Spacial Approach to Image Enhancement[J]. Information and control, 1986,15(2): 44-47.

Deep Learning of Principal Component for Car Model Recognition

Yongbin Gao[1], Hyo Jong Lee [1, 2]
[1]Division of Computer Science and Engineering,
[2]Center for Advanced Image and Information Technology
Chonbuk National University, Jeonju 561-756, Korea

Abstract - *Vehicle detection and analysis is widely used in various applications, such as automatic toll collection, driver assistance systems. Among these applications, car make and model recognition is a challenging task due to the close appearance between car models. In this paper, we proposed a novel algorithm based on deep learning of principal component (DLPC) to recognize the car make and model. Considering the 3D complexity of a car, we extract the frontal view of a car to recognize the make and model. After that, we transform the frontal view of a car to its feature mapping using principal component analysis (PCA). Finally, we use deep learning with three layers of restricted Boltzmann machines (RBMs) to recognize the car make and model. Experiment results show that our proposed framework achieves favorable recognition accuracy.*

Keywords: Car model recognition; Deep Learning; Principal component analysis.

1 Introduction

Vehicle analysis is an essential component in many intelligent applications, such as automatic toll collection, driver assistance systems, self-guided vehicles, intelligent parking systems, and traffic statistics (vehicle count, speed, and flow). Specially, an electronic toll collection system can automatically collect tolls according to the identification of vehicle models. Also, the identification of vehicle models can provide valuable information to the police for searching suspect vehicles. The appearance of a vehicle will change under varying environmental conditions and market requirements. The shapes of vehicles between companies and models is very similar, which results in confusion in vehicle model recognition. This makes the vehicle model recognition a challenging task.

Vehicle detection is prerequisite of vehicle analysis. Background subtraction [2]–[5] is widely used to extract motion features to detect moving vehicles from videos. However, this motion feature is not available for still images. To address this problem, Wu et al. [6] proposed the use of wavelet transformation to extract texture features to locate possible vehicle candidates. Tzomakas and Seelen [7] found that the shadow of a vehicle is a good cue for detecting vehicles. Ratan et al. [8] localize the possible vehicles based on the detection of vehicle wheels and verify the candidate vehicle by a diverse density method.

There are various application of vehicle analysis. Chen et al. [9] proposed to use SVM and random forests to classify vehicles on the road into four classes, namely, car, van, bus, and bicycle/motorcycle. Ma and Grimson [10] used edge points and modified SIFT descriptors to represent vehicles. AbdelMaseeh et al. [11] used the combination of global and local cues to recognize car model. Hsieh et al. [12] proposed a symmetrical SURF for both vehicle detection and model recognition. Considering the favorable performance of deep learning [13], we apply it to car model recognition. This is the first attempt to use deep learning for car model recognition to our knowledge.

In this paper, we proposed a novel algorithm based on deep learning of principal component to recognize the car make and model. Considering the 3D complexity of a car, we extract the frontal view of a car to recognize the make and model. After that, we transform the frontal view of a car to its feature mapping using principal component analysis. Finally, we use deep learning with three layers of restricted Boltzmann machines to recognize the car make and model.

The remainder of this paper is organized as follows. Section II describes the framework of our designed system. We then introduce the car model recognition based on deep learning of principal component in Section III. Section IV applies the above algorithm to our car database, and presents the experiment results. Finally, we conclude this paper in Section V.

2 Framework of our system

The framework of our system is shown in Fig. 1. We first detect the moving car by frame difference, it is effective in our system due to the fact that our camera is fixed on the street. Frame difference is simple and sufficient in this scenario, which enables real time application. It is not wise to recognize a car by its whole appearance due to its 3D characteristic. The 3D object results in large variance within one class incurred by pose changes. In this paper, we proposed to use frontal view of a car to recognize the car model. Frontal view of a car offer sufficient features to represent a car model. Also, a portion of the car image reduce the computation time than an entire car image. Furthermore, the frontal view of a car is basically

symmetrical, thus, we use a symmetrical filter to detect the frontal view. After the binary image is calculated from the frame difference, a symmetrical filter is used to extract the symmetrical region of the binary image. As a result, the symmetrical region is regarded as a frontal view of a car. The extracted frontal database in gallery is used to train the PCA model, which is applied to test images to extract principal component. The principal component of a car is fed into deep learning to recognize the car make and model.

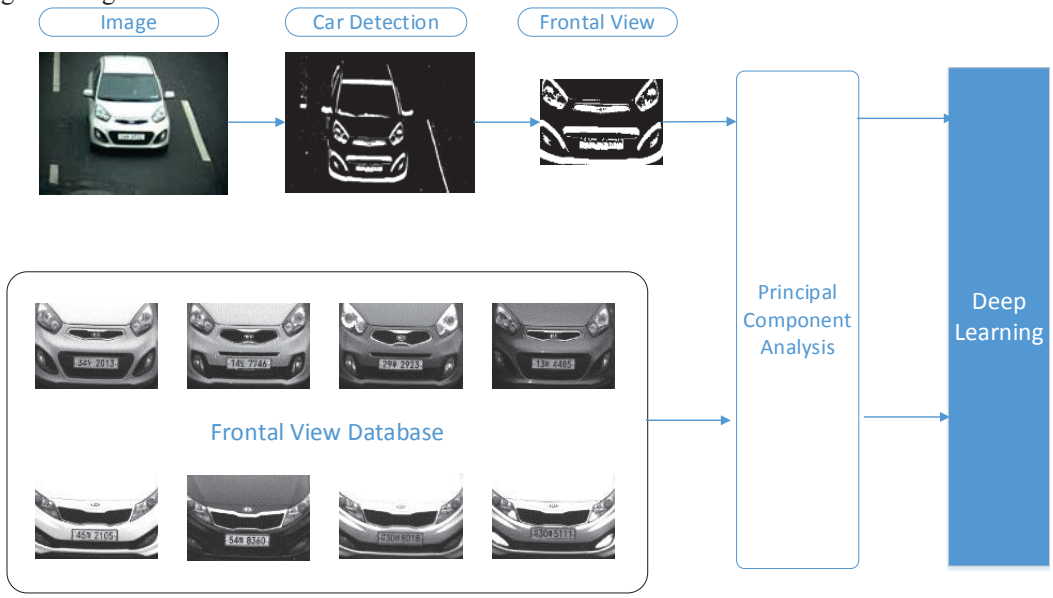


Fig. 1. Framework of proposed car detection and model recognition system based on deep learning of principal component analysis.

3 Deep Learning of Principal Component

3.1 Principal Component Analysis

PCA is a popular algorithm for dimensional deduction as well as feature extraction, which acquainted us with its successful application to face recognition using eigenfaces [16]. We try to learn the principal component of the frontal view of a car prior to feeding it into a deep network.

Let the training set be $T_1, T_2, \ldots T_M$, where each image with size $N * N$ is unrolled to a vector T_i of dimension N^2. The average of training images is calculated as $\overline{T} = \frac{1}{M}\sum_{n=1}^{M} T_n$, each training image is deducted from the average resulting in a vector $D_i = T_i - \overline{T}$. The covariance matrix of the difference image is calculated as follows:

$$C = \frac{1}{M}\sum_{i=1}^{M} D_i D_i^T$$

Let e_k and λ_k be the eigenvector sand eigenvalue of the covariance matrix C, respectively. By ranking the eigenvalues, we are able to see the efficacy of their associated eigenvector in handling the image variation. A projection matrix P_K consisting of K eigenvectors is generated by seeking the largest K associated eigenvalues. This projection matrix enables us to transform the original N^2 image space to K-dimension subspace, which is a better representation of original image. As for a new image T, it can be projected into the subspace y as follows:

$$y = (T - \overline{T})\, P_K$$

3.2 Deep Learning

Deep network aims at transforming the input data into other representations layer by layer. Conventional back-propagation suffers from the poor local optima problem as well as time consuming of weight learning. Also, labeled training data is a necessary for back-propagation, which is not always satisfied in case of small dataset. Deep learning method [13] provides a solution to address all these problems, which enable us to use unlabeled data to initialize the deep network. The idea behind the deep learning method is to learn $p(image)$ instead of $p(label|image)$, which is to maximize the probability that a generative model would have produced the input data.

To make use of the unlabeled data that is easy to acquire, we use the auto-encoder to pre-training the deep network. As a result, an initial point of weights close to an optimal solution is obtained. Restricted Boltzman machine (RBM) offers us an effective pre-training method, which is a two layers network with stochastic, binary pixels as units. These two layers comprise of pixels of "visible" units and "hidden" units that are connected using symmetrically weighted connections. The RBM model can be represented as an energy model:

$$E(v,h) = -\sum_{i\in visible} b_i v_i - \sum_{j\in hidden} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (1)$$

where v_i and h_j are the binary values of unit i and j, b_i and b_j are their corresponding biases, and w_{ij} is the weight of their connection. The probability of a possible image is assigned according to this energy. Given a training image, we alternately calculate the probability of binary state v_i and hidden units h_j to be set to 1 as follows:

$$p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{ij}) \quad (2)$$

$$p(v_i = 1|h) = \sigma(b_i + \sum_j h_j w_{ij}) \quad (3)$$

By updating the states of hidden units and reconstruction of visible units, we can obtain a change in weight by:

$$\Delta w_{ij} = \varepsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}) \quad (4)$$

where ε is the learning rate, $\langle v_i h_j \rangle_{data}$ is the fraction of multiplication of visible and hidden units driven by data, while $\langle v_i h_j \rangle_{recon}$ is the fraction driven by reconstruction images.

In our experiments, three layers RBMs are used to car model recognition. The binary image of frontal part of a car with size 40*50 is used as input. This binary image is unroll into a vector with dimension of 2000. The following three RBM layers are used as pre-training to obtain an initial weight. After the pre-training by RBM, we use traditional back-propagation method to fine-tuning the deep network using the labels of 107 car models.

4 Results

Since there is no standard car model database available, we built a car database to evaluate the performance of our proposed framework, which consists of 3210 car images with varying companies and models. These images composed of 107 car models with 30 images for each model. We are apt to use image instead of video, because it is easy to measure the accuracy. However, to use the frame difference for images, we shifted each image with 10 pixels to generate a neighboring image. The difference image was generated by an image and its shifted image.

As for our frontal view detection, the results of four cars is shown in Fig. 2, which includes five car models with four companies. Left column shows the original car image with detected frontal region marked by red box. Right column shows the binary representation of frontal view. Our system is able to detect the frontal view of each car image accurately. We test the detection algorithm on all 3210 images in the system, and get 100% accuracy with regard to the detection accuracy. Thus, the detection algorithm is effective and fast in our system.

The frontal views of a car are feed into the trained deep model as input, the output label is used to recognize the car model. Fig. 3 shows the six eigen cars learned from PCA, these eigen cars have the largest eigen values, with which the test cars are represented as the linear combination of these eigen cars. We compared our DLPC method to the prestigious features as follows: local binary pattern (LBP) [14], local Gabor binary patterns (LGBP) [15], and scale-invariant feature transform (SIFT) [1] and the primitive deep learning method. The experimental results is shown in Table 1. In our experiments, we use 29 images of each model for training, and the left one image for testing. The results show that DLPC can achieve impressive performance compared with other methods.

Fig.2 Results of frontal view extraction on five car images with four companies and five models.

Table 1. Performance comparison of car model recognition with prestigious methods.

Algorithm	Accuracy (%)
LBP	46.0
LGBP	68.8
SIFT	78.3
Deep Learning	88.2
DLPC	90.6

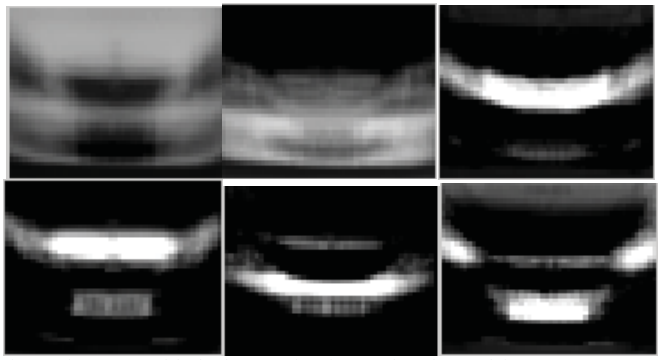


Fig.3 Eigen cars with the largest six Eigen values learned from PCA.

5 Conclusions

In this paper, we proposed a novel algorithm based on deep learning of principal component (DLPC) to recognize the car make and model. Considering the 3D complexity of a car, we extract the frontal view of a car to recognize the make and model. After that, we transform the frontal view of a car to its feature mapping using principal component analysis (PCA). Finally, we use deep learning with three layers of restricted Boltzmann machines (RBMs) to recognize the car make and model. Experiment results show that our proposed framework achieves better results than some prestigious methods.

Acknowledgment: This work was supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea. This work was also supported by Business for Academic-industrial Cooperative establishments funded Korea Small and Medium Business Administration in 2014 (Grants No. C0221114).

6 References

[1] D. G. Lowe. "Distinctive image features from scale-invariant keypoints"; Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[2] A. Faro, D. Giordano, and C. Spampinato. "Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection"; IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1398–1412, Dec. 2011.

[3] H. Unno, K. Ojima, K. Hayashibe, and H. Saji. "Vehicle motion tracking using symmetry of vehicle and background subtraction"; in Proc. IEEE Intell. Veh. Symp., 2007, pp. 1127–1131.

[4] A. Jazayeri, H.-Y. Cai, J.-Y. Zheng, and M. Tuceryan. "Vehicle detection and tracking in car video based on motion model"; IEEE Trans. Intell. Transp. Syst., vol. 12, no. 2, pp. 583–595, Jun. 2011.

[5] G. L. Foresti, V. Murino, and C. Regazzoni. "Vehicle recognition and tracking from road image sequences"; IEEE Trans. Veh. Technol., vol. 48, no. 1, pp. 301–318, Jan. 1999.

[6] J. Wu, X. Zhang, and J. Zhou. "Vehicle detection in static road images with PCA-and-wavelet-based classifier"; in Proc. IEEE Conf. Intell. Transp. Syst., Aug. 25–29, 2001, pp. 740–744.

[7] C. Tzomakas and W. Seelen. "Vehicle detection in traffic scenes using shadow"; Inst. Neuroinf., Ruhtuniv., Bochum, Germany, Tech. Rep. 98-06, 1998.

[8] A. Lakshmi Ratan,W. E. L. Grimson, and W. M.Wells. "Object detection and localization by dynamic template warping"; Int. J. Comput. Vis., vol. 36, no. 2, pp. 131–148, Feb. 2000.

[9] Z. Chen, T. Ellis, and S. A. Velastin. "Vehicle type categorization: A comparison of classification schemes"; in Proc. 14th Int. IEEE Conf. Intell. Transp. Syst., Oct. 2011, pp. 74–79.

[10] X. Ma, W. Eric, and L. Grimson. "Edge-based rich representation for vehicle classification"; in Proc. IEEE Int. Conf. Comput. Vis., 2005, pp. 1185–1192.

[11] M. AbdelMaseeh, I. BadreIdin, M. F. Abdelkader and M. EI Saban. "Car Make and Model recognition combining global and local cues"; in Proc. IEEE Int. Conf. Pattern Recognition, 2012, pp. 910-913.

[12] J. W. Hsieh, L. C. Chen and D. Y. Chen. "Symmetrical SURF and Its Applications to Vehicle Detection and Vehicle Make and Model Recognition"; IEEE Trans. Intell. Transp. Syst., vol. 15, no. 1, pp. 6-20, Feb. 2014.

[13] G. E. Hinton, R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks"; Science, vol. 313, pp. 504-507, 2006.

[14] T. Ahonen, A. Hadid, M. Pietikainen. "Face Description with Local Binary Patterns: Application to Face Recognition"; IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 2037-2041, 2006.

[15] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang. "Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition"; in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 786-791, 2005.

[16] M. Turk and A. Pentland. "Face recognition using eigenfaces"; in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591, 1991.

Off-line Handwritten Arabic Character Recognition: A Survey

Maad Shatnawi
Higher Colleges of Technology, Abu Dhabi, UAE

Abstract - *The automatic recognition of text on scanned images has several applications such as automatic postal mail sorting and searching in large volume of documents. Although Arabic handwritten text recognition has been addressed by many researchers, it remains a challenging task due to several factors. This paper presents an overview of off-line handwritten Arabic character recognition and summarizes the main technical challenges and characteristics of Arabic. It also investigates the relevant existing research carried out towards this perspective.*

Keywords: Handwritten character recognition, OCR, Arabic text recognition,

1. Introduction

Handwritten character recognition is the ability of computers to convert human writing into text either on-line or off-line. On-line recognition is performed by writing directly to a peripheral input device such as a personal digital assistant (PDA) and a cellular phone. Off-line recognition, also called optical character recognition (OCR), is the ability of computers to convert human writing into text by scanning. The early attempts in the area of Latin character recognition were made in the middle of the 1940s with the development of digital computers. One of the early attempts in Chinese character recognition was made in 1966. However, the first publication of Arabic text recognition was in 1975 [1].

Some research effort has been done on surveying handwritten Arabic character recognition. Amara and Bouslama [2] presented a review of the classification techniques used in the optical character recognition of Arabic script. Lorigo and Venu Govindaraju [3] reviewed off-line handwritten Arabic character recognition methods. Beg et al. [4] presented the current OCR products and reviewed the work done in the hardware domain of Arabic OCR. Harous and Elnagar [5] presented handwritten character-based parallel thinning Algorithms. AL-Shatnawi and Omar [6] reviewed the various Arabic baseline detection methods. Alginahi [7] surveyed Arabic character segmentation methods. Al-Shatnawi et al. [8] evaluated and compared skeleton Arabic character extraction methods. This paper reviews the main stages of handwritten Arabic OCR systems, presents the major technical challenges of these systems, and provides a comprehensive review of the research done in this field.

The remainder of this paper is organized as follows. The next section addresses the main technical challenges in the field of Arabic handwritten character recognition. Section 3 presents the main characteristics of Arabic writing. Section 4 briefly describes the five stages of Arabic character recognition. Arabic baseline detecting methods and approaches are presented in section 5. Section 6 focuses on the major classification techniques and approaches that are used in Arabic OCR. The use of synthetic data in Arabic handwritten OCR systems is illustrated in section 7, and concluding remarks are presented in Section 8.

2. Technical Challenges of Handwritten Arabic Character Recognition

Although the latest improvements in Arabic character recognition methods and systems are very promising, the automatic recognition of Arabic handwritten characters remains a challenging task due to many factors that are summarized as follows. The first factor is the lack of sufficient support in terms of funding, books, journals, and conferences at all levels including governments and research institutions, and the lack of interaction between researchers of this field [1]. Secondly, the lack of enough Arabic digital dictionaries and programming tools, and the absence of large public databases of Arabic handwritten characters and words when compared to English where large databases such as CEDAR [9] have been publicly available for a long time. The lack of large Arabic databases is, in part, due to the difficult, time consuming, and error prone process of generating ground truth1 for Arabic on the character level [10, 11]. Thirdly, the start of Arabic character recognition is very late compared to other languages such as Latin and Chinese. In addition to that, the unique characteristics of Arabic writing can be considered as a great challenge. The next section presents some of these characteristics.

3. Main Characteristics of Arabic Writing

Arabic is a native language for more than 250 million people. It is the third largest international language used by over one billion Muslims in their different religious activities. In addition to the Arabic language, there are several languages that use the Arabic alphabet, such as Urdu, Farsi (Persian), Pashto, Jawi, and Kurdish. The

[1] Ground truth refers to assigning correct symbolic text to images which will be used for learning.

Arabic text is written from right to left and is always cursive in both machine printed and handwritten text [12].

The Arabic alphabet set is composed of 28 basic letters which consist of strokes and dots. Dots, above and below the characters, play a major role in distinguishing some characters that differ only by the number or location of dots e.g. Ba (ب), Ta (ت), and Noon (ن).

Table 1. Different shapes of Arabic letters.

Character	Isolated	Beginning	Middle	End
Alef	ا			ـا
Ba	ب	بـ	ـبـ	ـب
Ta	ت	تـ	ـتـ	ـت
Tha	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Ha'	ح	حـ	ـحـ	ـح
Kha	خ	خـ	ـخـ	ـخ
Dal	د			ـد
Thal	ذ			ـذ
Ra	ر			ـر
Zy	ز			ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Sad	ص	صـ	ـصـ	ـص
Dhad	ض	ضـ	ـضـ	ـض
Tah	ط	طـ	ـطـ	ـط
Dha	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Ghain	غ	غـ	ـغـ	ـغ
Fa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Ha	ه	هـ	ـهـ	ـه
Waw	و			ـو
Ya	ي	يـ	ـيـ	ـي

Table 2. Numerals that are commonly used in Arabic writing.

Arabic	Indian
0	٠
1	١
2	٢
3	٣
4	٤
5	٥
6	٦
7	٧
8	٨
9	٩

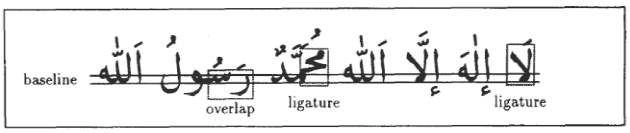


Figure 1. A sample of handwritten Arabic showing some of its characteristics [1].

The shape of an Arabic letter changes according to its location in the word, as shown in Table 1. For each character, there can be two to four different shapes: isolated, connected from the left (beginning of a word), connected from the left and right (middle of a word), and connected from the right (end of a word). Out of the 28 basic Arabic letters, six can be connected from the right side only while the other 22 can be connected from both sides. These six characters are: Alef (ا), Dal (د), Thal (ذ), Ra (ر), Zy (ز), and Waw (و) .These six characters have only two shapes, the isolated shape and the end shape, whereas the rest of the alphabets can appears in any of the four shapes mentioned above [13]. Consequently, each word may form one or more sub-words, where a sub-word is one or several connected characters, for example مستشفى, نور, and مناسبات. Moreover, in certain fonts, several characters can be vertically combined to form a ligature, especially in typeset and handwritten text. Ligatures can be formed out of two, three, or four characters [2]. Characters in a word may also vertically overlap without touching. Figure 1 presents some of these characteristics.

The use of special stress marks called diacritics is another distinguishing characteristic of Arabic. Diacritics such as Fat-ha (ً), Dhammah (ٌ), Shaddah (ّ), Maddah (~), Sukun (ْ), and Kasrah (ٍ) may change the pronunciation and the meaning of the word. The diacritics significantly affect the OCR performance [6, 14].

There are nine Arabic letters; Sad (ص), Dhad (ض), Tah (ط), Dha (ظ), Fa (ف), Qaf (ق), Meem (م), Ha (ه), and Waw (و) that have closed loops. This makes the closed loop an important feature in recognizing Arabic characters. One of the important characteristics of Arabic text is the presence of a baseline which is an imaginary horizontal line running through the connected portions of the text. If the script is handwritten, the baseline is not straight, and may only be estimated. Another feature of Arabic characters is that they do not have a fixed width or size, even in printed from. The character size varies according to its shape which is, in turn, a function of its position in the word.

In addition to the 28 characters, Arabic has additional non basic characters such as Hamzah (ء) and Ta marboota (ة). Hamzah can be isolated, on Alef (أ), on Waw (ؤ), or on Ya (ئ). Ta marboota is a special form of the letter Ta(ت) that only appears at the end of words. There are two

types of numerals that are commonly used in Arabic; the Indian and Arabic numerals as shown in Figure 2.

4. Stages of Arabic Character Recognition

The process of Arabic character recognition consists of five stages; preprocessing, segmentation, feature extraction, classification, and post-processing [1]. The preprocessing enhances the raw images by reducing noise and distortion. This stage includes thinning, binarization, smoothing, alignment, normalization, and base-line detection.

Since Arabic text is cursive, segmentation is an important step in Arabic text recognition. Segmenting a page of text includes page decomposition and word segmentation. Page decomposition separates different logical parts, like text from graphics and lines of a paragraph, while word segmentation is the breakdown of words into isolated characters. The feature extraction stage analyzes a text segment and selects a set of structural or statistical features that can be used to uniquely identify the text segment. These features are extracted and passed in a form suitable for the recognition phase. Selecting the most suitable features plays a crucial role in the performance of the classification stage.

The recognition or classification stage is the main decision-making stage of an OCR system. This stage uses the features extracted in the previous stage to identify the text segments based on structural or statistical models. The classification stage uses machine learning techniques such as Artificial Neural Networks (ANN), support vector machines (SVM), k-nearest neighbors (*k*-NN), and Hidden Markov Models (HMM). The post-processing stage improves the recognition by refining the decisions taken by the previous stage and recognizes words by using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information [1].

5. Arabic Baseline Detection Methods

Arabic baseline detecting methods can be categorized into four groups; the horizontal projection, the word skeleton, the word contour representation, and the Principal Components Analysis [6]. This section presents the current approaches of Arabic baseline detection. The horizontal projection method reduces the two-dimensional data into a one-dimension based on the pixels of the text image by summing up the pixel values of each row, and the row that obtains the highest score will be considered as the baseline. Although this method is easy to be implemented to printed text, it cannot easily detect handwritten text and it can be easily fooled by the diacritics.

Pechwitz and Maergner [15] used the word skeleton method to detect the baseline. This method creates the skeleton of the word based on polygon approximations. This method is applicable to both printed and handwritten Arabic text and is not affected by diacritics. However, it is computationally more expensive than other methods.

Farook et al. [16] detected the baseline according to the word Contour representation. This method finds the local minima of the word contour and then applies the linear regression technique to estimate the baseline of the text. This method can be applied to both printed and handwritten Arabic text and to both diacritized and non-diacritized text.

Burrow [17] applied the Principal Component Analysis for either foreground or background of the pixel distribution of the Arabic text in order to estimate the baseline direction and then applied the horizontal projection to detect the baseline. Detailed literature review of baseline detection methods can be found in [6].

6. Classification Approaches

There are several machine-learning (ML) classification techniques that are successfully applied to character recognition. These techniques can be classified into model-based and instance-based. ANN, HMM, and SVM are model-based classifiers that learn a model based on training examples to determine the decision boundaries between classes. Conversely, *k*-NN is an instance-based classification method, which assigns the class of the closet training examples in order to classify a new example. This section presents some of the existing approaches that use ML classification methods.

Graves and Schmidhuber [18] introduced a globally trained offline handwriting recognizer which is based on the raw pixel values of the input images. The system does not require any alphabet specific preprocessing and is applicable to any language. The two dimensional images were transformed into one dimensional label sequences of pixel values. The system employed the multidimensional long short-term memory (LSTM) neural networks as the classification technique. The system was evaluated on the IFN/ENIT database [10] where it outperformed the winner of ICDAR 2007 Arabic handwriting recognition contest [19] although neither author understands a word of Arabic.

Mahmoud and Awaida [20] described a technique for automatic recognition of off-line writer-independent handwritten Arabic (Indian) numerals using SVM and HMM. This work evaluated the use of the Gradient, Structural, and Concavity (GSC) features with a SVM classifier, and then the results are compared with HMM results using the same dataset and features. The SVM and HMM classifiers were trained with 75% of the data and tested with the remaining data. A two-stage exhaustive

parameter estimation technique is used to estimate the best values for SVM parameters. The achieved average recognition rates were 99.83% and 99.00% using the SVM and HMM classifiers, respectively. The recognition rates of SVM proved to be superior to those of HMM for all digits and tested writers.

Natarajan et al. [21] evaluated HMM in handwritten character recognition. The authors concluded that the HMM-based system have several advantages because no pre-segmentation of words is required. On the other hand, this system suffers from two limitations; the assumption of conditional independence of the observations given the state sequence, and the restrictions on feature extraction imposed by frame-based observations. The use of pixel-level features from narrow slices of the text, specifically, the narrow windows provide very little contextual information making the conditional independence assumption in these systems unrealistic.

Hamdani et al. [22] presented an off-line handwriting recognition system based on the combination of multiple HMM classifiers. The classifiers are based on three on-line features and one off-line feature. The three on-line features are pixel values, densities and Moment Invariants, and pixel distribution and concavities. The authors used the technique described in [23] which allows having the on-line trace of the writing in a given image. The system was implemented using the HMM Toolkit (HTK) [24] and the IFN/ENIT database [10]. The authors concluded that the combination of on-line and off-line systems significantly improves the recognition accuracy.

Al-Hajj Mohamad et al. [25] proposed Arabic handwritten city names recognition based on combining three HMM classifiers that include a set of baseline-dependent and baseline-independent features. The three classifiers are combined at the decision level using three combination schemes: the sum rule, the majority vote rule, and an original neural network-based combination whose decision function is learned through candidate words' scores.

Mahmoud and Abu-Amara [26] describes a technique for the recognition of off-line handwritten Arabic (Indian) numerals using Radon and Fourier Transforms. A database of 44 writers with 48 examples of each digit totaling 21120 examples is used for training and testing the classifier. Radon-based features are extracted from Arabic numerals. Nearest Mean, *k*-NN, and HMM are used as digit classifiers. The recognition rates of these classifiers are 98.66%, 98.33%, 97.1%, respectively.

Parvez and Mahmoud [27] proposed a structural-based character recognition method. An Arabic text line is segmented into words/sub-words and dots are extracted. An adaptive slant correction algorithm that is able to correct the different slant angles of the different components of a text line is presented. A polygonal approximation algorithm is employed for text segmentation. Dynamic programming is used to select best hypotheses of a sequence of recognized characters for each word/sub-word. Prototype selection using set-medians, lexicon reduction using dot-descriptors are also utilized.

Tomeh et al. [28] and Habash and Roth [29] incorporated linguistically and semantically related features to Arabic character recognition systems. They presented an error detection system that uses deep lexical and morphological feature models to locate words or phrases that are likely incorrectly recognized. They used BBN's Byblos HMM-based off-line handwriting recognition system [30] to generate an *N*-best list of hypotheses for each segment of Arabic handwriting.

Sahlol et al. [31, 32] proposed a feed-forward back-propagation Neural Network approach. The method consists of four stages; binarization, normalization, noise removal, feature extraction, and classification. Three types of features were extracted; structural, statistical, and topological features. Structural features are the upper and lower profiles that capture the outlining shape of a connected part of the character as well as horizontal and vertical projection profiles. Statistical features include the four neighboring pixels for each pixel. Topological features include end points, pixel ration, and height to width ratio.

7. Use of Synthetic Data in Character Recognition

Large databases of Arabic handwritten characters and words are not publicly available when compared to Latin languages. Many papers in Arabic character recognition used their own small datasets such as Al -Badr and Mahmoud [1], Amin [33], and Khedher et al. [34], or they talked about large databases that are not available to public such as Kharma et al. [35], and Al-Ohali et al. [36]. Therefore, the use of synthetic data in building character recognition systems of different languages has been discussed and examined by many researchers.

Margner and Pechwitz [37] introduced a system for automatic generation of synthetic printed data for Arabic OCR systems. This system can be described as follows. First, the Arabic text has to be typeset. Then, a noise-free bitmap of the document and the corresponding GT is automatically generated. And finally, an image distortion can be superimposed on the character or word image to simulate the expected real world noise of the intended application.

Elarian et al. [38] presented an approach to synthesize Arabic handwriting text. First, real word images are segmented into labeled characters which are then concatenated in an arbitrary way to synthesize artificial word images. The nearest Euclidean-distance neighbor is used for matching characters that can be concatenated to produce natural-looking words. This synthesized text is

used to train and test OCR system. Although the proposed approach is still infancy and was tested on only two writers from the IFN/ENIT dataset [10], the authors reported promising results.

Dinges et al. [39] presented a method for Arabic handwriting synthesis using Active Shape Models (ASM) computed based on 28046 online samples of multiple writers. ASMs were used to generate unique letter representations for each synthesis. Subsequently these representations were modified by affine transformations, smoothed by B-Spline interpolation and composed to text.

Shatnawi and Abdallah [40] extended the spatial congealing technique [41] and evaluated it on Arabic characters. Congealing models human distortions in the examples of one or more handwritten characters, and then uses this model to generate synthetic examples of other characters that are similarly distorted. The extended congealing approach along with different types of systematic affine transformations including rotation, scaling, and shearing were used to synthesize a large number of virtual training examples of isolated Arabic characters. Experimental results proved significant improvements across three machine-learning classification algorithms; k-NN, Naïve Bayes, and SVM.

8. Conclusion

This survey is focused on off-line handwritten Arabic character recognition. The most important challenges that face handwritten Arabic OCR systems and the main characteristics of Arabic writing were addressed. We investigated the major relevant existing approaches carried out towards this perspective. There are several approaches in this field. However, the research in handwritten Arabic character recognition is still in an early stage when compared to Latin and other languages.

Acknowledgment

The author would like to thank Prof. Boumediene Belkhouche for reviewing an earlier draft of this work.

References

[1] Al-Badr B, Mahmoud SA (1995) Survey and Bibliography of Arabic Optical Text Recognition. Elsevier Science B.V., Signal Processing 41: pp 49-77.

[2] Amara N, Bouslama F (2003) Classification of Arabic script using multiple sources of information: State of the art and perspectives. International Journal on Document Analysis and Recognition, 5(4): 195-212.

[3] Lorigo LM, Govindaraju, V (2006) Off-line Arabic Handwritten Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. EEE Trans. Pattern Analysis and Machine Intelligence 28(5): 712–724.

[4] Beg A, Ahmed F, Campbell P (2010) Hybrid OCR Techniques for Cursive Script Languages –A Review and Applications. 2nd International Conference on Computational Intelligence, Communication Systems and Networks, Liverpool, United Kingdom.

[5] Harous S, Elnagar A (2009) Handwritten Character-Based Parallel Thinning Algorithms: A Comparative Study. University of Sharjah Journal of Pure & Applied Sciences 6(1): 81-101.

[6] AL-Shatnawi A, Omar K (2008) Methods of Arabic Language Baseline Detection – The State of Art, IJCSNS International Journal of Computer Science and Network Security, 8(10).

[7] Alginahi, Y. M. (2013). A survey on Arabic character segmentation. International Journal on Document Analysis and Recognition (IJDAR), 16(2), 105-126.

[8] AL-Shatnawi, A. M., AlFawwaz, B. M., Omar, K., Zeki, A. M. (2014). Skeleton extraction: Comparison of five methods on the Arabic IFN/ENIT database. In Computer Science and Information Technology (CSIT), 2014 6th International Conference on (pp. 50-59). IEEE.

[9] CEDAR (Center of Excellence for Document Analysis and Recognition) (2006) USPS Office of Advanced Technology Database of Handwritten Cities, States, ZIP Codes, Digits, and Alphabetic Characters, [online]. Available at http://www.cedar.buffalo.edu.

[10] Pechwitz M, Maddouri S, Margner V, Ellouze N, Amiri H (2002) IFN/ENIT - Database of Handwritten Arabic Words. CIFED: 1-8.

[11] Kanungo T, Resnik P, Mao S, Kim D, Zheng Q (2005) The Bible and Optical Character Recognition. Communications of the ACM, 48(6): 124-130.

[12] Khorsheed M (2002) Off-Line Arabic Character Recognition – A Review. Pattern Analysis and Applications, 5:31-45.

[13] Al-Shoshan AI (2006) Arabic OCR Based on Image Invariants. Proceedings of the Geometric Modeling and Imaging — New Trends (GMAI'06), IEEE.

[14] Zeki AM (2005) The segmentation problem on Arabic character recognition – the state of the art. 1st International Conference on Information and Communication Technology (ICICT) Karachi, Pakistan: 11-26.

[15] Pechwitz M, Maergner V (2002) Baseline estimation for Arabic handwritten words. Frontiers in Handwritin Recognition: 479–484.

[16] Farooq F, Govindaraju V, Perrone M (2005) Preprocessing Methods for Handwritten Arabic Documents. (ICDAR'05) Proceedings of the 2005

Eight International Conference on Document Analysis and Recognition, IEEE 1: 267-271.

[17] Burrow P (2004). Arabic handwriting recognition. [M.Sc. thesis]. Edinburgh (England): University of Edinburgh.

[18] Graves A, Schmidhuber J (2008) Offline handwriting recognition with multidimensional recurrent neural networks. 22sd Conference on Neural Information Processing Systems (NIPS).

[19] Margner V, Abed HE (2007) Arabic handwriting recognition competition. 9th International Conference on Document Analysis and Recognition (ICDAR 2007) 2: 1274–1278, Washington, DC, USA, IEEE Computer Society.

[20] Mahmoud SA, Awaida SM (2009) Recognition of Off-line Handwritten Arabic (Indian) Numerals Using Multi-scale Features and Support Vector Machines Vs. Hidden Markov Models. The Arabian Journal for Science and Engineering (AJSE), 34: 429-444.

[21] Natarajan P, Subramanian K, Bhardwaj A, Prasad R (2009) Stochastic Segment Modeling for Offline Handwriting Recognition. 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain.

[22] Hamdani M, Abed HE, Kherallah M, Alimi AM (2009) Combining Multiple HMMs Using On-line and Off-line Features for Off-line Arabic Handwritten Recognition. 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain.

[23] Elbaati A, Kherallah M, Ennaji A, Alimi AM (2009) Temporal Order Recovery of the Scanned Handwriting. 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain.

[24] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu XA, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2006) The HTK Book. Cambridge University, England.

[25] Al-Hajj Mohamad, R., Likforman-Sulem, L., and Mokbel, C. (2009). Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(7), 1165-1177.

[26] Mahmoud SA, Abu-Amara MH (2010) Recognition of Handwritten Arabic (Indian) Numerals using Radon-Fourier-based Features. 9th WSEAS international conference on Signal processing, robotics and automation (ISPRA '10), Cambridge, UK, portal.acm.org.

[27] Parvez, M. T., and Mahmoud, S. A. (2013). Arabic handwriting recognition using structural and syntactic pattern attributes. Pattern Recognition, 46(1), 141-154.

[28] Tomeh, N., Habash, N., Roth, R., Farra, N., Dasigi, P., and Diab, M. T. (2013). Reranking with Linguistic and Semantic Features for Arabic Optical Character Recognition. In ACL (2): 549-555.

[29] Habash, N., and Roth, R. M. (2011, June). Using deep morphology to improve automatic error detection in Arabic handwriting recognition. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1: 875-884.

[30] Saleem, S., Cao, H., Subramanian, K., Kamali, M., Prasad, R., and Natarajan, P. (2009) Improvements in BBN's HMM-based offline Arabic handwriting recognition system. In 10th International Conference on Document Analysis and Recognition (ICDAR'09): pp. 773-777. IEEE.

[31] Sahlol, Ahmed T., Cheng Y. Suen, Mohammed R. Elbasyouni, and Abdelhay A. Sallam. (2014) A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters. Journal of Pattern Recognition and Intelligent Systems.

[32] Sahlol, A., and Suen, C. (2014). A Novel Method for the Recognition of Isolated Handwritten Arabic Characters. arXiv preprint arXiv:1402.6650.

[33] Amin A (1998) Off-line Arabic Character Recognition: The State of the Art. Pattern Recognition Society, 31(5): 517-530.

[34] Khedher M, Abandah G, Al-Khawaldeh A (2005) Optimizing Feature Selection for Recognizing Handwritten Arabic Characters. Proceedings of the Second World Enformatika Conference, WEC'05, Istanbul, Turkey: 81-84.

[35] Kharma N, Ahmed M, Ward R (1999) A New Comprehensive Database of Hand-written Arabic Words, Numbers and Signatures used for OCR Testing. IEEE Canadian Conference on Electrical and Computer Engineering, Alberta, Canada: 766-768.

[36] Al-Ohali Y, Cheriet M, Suen C (2000) Database For Recognition of Handwritten Arabic Cheques. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam: 601-606.

[37] Margner V, Pechwitz M (2001) Synthetic Data for Arabic OCR System Development. In: Sixth International Conference on Document Analysis and Recognition (ICDAR'01), IEEE: 1159-1163.

[38] Elarian YS, Al-Muhtaseb HA, Ghouti LM (2011) Arabic Handwriting Synthesis. 1st International Workshop on Frontiers in Arabic Handwriting Recognition.

[39] Dinges L, Al-Hamadi A, and Elzobi M (2013) An Approach for Arabic Handwriting Synthesis Based on Active Shape Models. In Document Analysis and

Recognition (ICDAR), 2013 12th International Conference on. IEEE, 1260–1264.

[40] [40] Shatnawi M. and Abdallah S. (2015). Improving Handwritten Arabic Character Recognition by Modeling Human Handwriting Distortions. ACM Transactions on Asian and Low-Resources Language Information Processing.

[41] [41] Learned-Miller E (2006) Data Driven Image Models through Continuous Joint Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2) pp 236-250.

A Pre-sorting Method for Iron Pan Surface Inspection Based on Image Recognition

[1]Yujiao Liu, [2]Yongwei Zhang, [3]Xitao Zheng
[1,2]Shanghai Moses Marine Engineering Corp., Shanghai, 201306, China
[3]Zhejiang Ocean University, Zhejiang, 316022, China

Abstract-*In this paper, we propose a method to evaluate the polishing quality of iron pan by using image recognition technology. The method is designed to enable the automatic sorting for iron pot. The inner surface image of iron pan is acquired by a CCD camera, then the image information entropy is selected as the texture feature of these images. 40 groups of information entropy from pre-qualified the iron pan samples are used as the quality detection standard. Then the test samples are chosen, the information entropy of which are calculated, and being compared with the standard. The comparison result shows that it is practicable to pick out the failed sample by the experimental threshold. The method is used in pan inspection line and considerable reduced the load of human inspectors.*

Keywords: *image recognition, iron pan, surface polishing, quality detection.*

1 Introduction

While the human cost is increasing quickly in China, the production and quality inspection process need heavy human intervention, which prevent the production scale. The method proposed in this paper is used in the pot inspection line to enable he system to pick out specked pot and let the inspection workers to focus on the minor defect on the processed surface, which can reduce working load and increase the efficiency for human inspection.

There are lots of surface roughness detection methods of work items. Among them there are three traditional ones: sample block comparison method, moulage method and stylus measurement method. Sample block comparison method is a qualitative visual comparison method, which generally relies on the experienced worker. In the moulage method, some plastic block materials are coated on the surface which is to be measured as the moulage. The shape and profile information of the surface is pressed on the moulage, then the surface roughness of workpiece will be measured from the surface of the moulage. But the measuring accuracy of this method is low and the process is tedious. While in the stylus measuring method, a special pin is moved along the surface texture in the direction that is perpendicular to the measured surface at a certain speed, then the roughness values of the surface can be displayed on the recorder. It is easy to scratch the surface in this way, and the measurement speed is low. Due to the diameter of the probe, it may filter out the high-frequency information of the surface. With these features, its application is limited [1].

With the development of computer and laser technology, many new detection methods were produced, such as interferometry [2], scattering method [3], and the speckle measurement method [4], etc. In addition to the above mature method that based on optical technology, people also explored other means to detect the surface roughness, including atomic force microscope method (ATM) [5], optical sensor method [6] and so on. Usually these method result higher device cost.

In this paper, the author proposed a method to automatically detect some surface defects at a low price. The process is embedded in automatic pan production line which is now under testing. This project designed to be human free and using robot arm to handle most of the moving and positioning of the pans. Thus the visual detection and sensation of the progress is an essential part to provide a robust and fault-tollerenceing system. With the fast detecting speed, high efficiency and low cost, this method is expected to be used in more pot process quality control system.

2 The research of image technology on surface roughness evaluation

Surface roughness is an important indicator to reflect the micro geometry of machine part surface, also a widely used characteristic of the parts surface inspection. Now many scientific researches provide key technologies for the surface roughness measurement, which result the development of the surface roughness measuring technique. The application of image processing technology in the surface roughness measurement is also the mutual fusion among theses technical application field.

In 1998, Du-Mine Tsai and others did some research on surface roughness by using neural network technology based on visual system [7].

In 2000, B.Y.Lee built the relationship between surface roughness and surface image texture by using the abductive network, the training parameters are the main peak frequency, the squares of frequency amplitude and gray standard deviation. With these parameters the surface roughness value could be estimated by the precise modeling [8]. ShinnYing Ho, from China Taiwan, predicted the surface roughness through the establishment of adaptive fuzzy neural network (ANFIS) model in 2002, He used the cutting speed, feed rate, back engagement and image grey value as the network training parameters [9].

In the next year, Kuane - Chyi Lee and others also used ANFIS network to carry on the related research, but they did not use the cutting parameters as the training parameters, they used image spatial frequency, arithmetic mean and the gray level standard deviation as the network training parameters [10]. Rajneesh Kumar, from India, used cubic convolution interpolation method to zoom in the image, and took the linear edge sharpening operator to improve image quality, then used the method of regression analysis to extract the grayscale average as image characteristics, and established the relationship between the mean gray level and surface roughness [11]. In 2006, Yun Xian Ho and other people studied the influence of the light condition changes to the roughness detection systematically, concluded that the light source angles from 50 to 70 degrees were beneficial for surface roughness detection [12].

India's V.E lango and L. Karunamoorthy also did some research on the effects of light source to the roughness detection in 2007, and reached a similar conclusion [13]. Yujing Wang and his team did the research on the roughness measurement of turner workpiece by using image method in 2007. They extracted the characteristic parameters of the cutting surface and compared them with stylus measurement results, and got the relationship between them, then built the mathematical formula, finally completed the measurement of cutting surface roughness [14]. In 2008, Chunya Wu used two-dimensional discrete Fourier transform into the frequency analysis to the grinding surface image; he took the power spectrum radius, average power spectrum and the center power spectrum percentage as the characteristics of grinding surface images, and used them to establish the BP neural network, then completed the grinding surface roughness measurement [15]. In 2009, Ghassan A.AI-Kindi carried on an analysis on workpiece surface roughness from different processing methods based on machine vision measurement. He compared the results obtained from machine vision method with the traditional stylus measurement results, and it is concluded that machine vision method could be widely applied to the workpiece surface roughness detection [16].

3 Theories

In this paper, the iron pan image process takes the following steps: image acquisition, color image converted to gray image, image background deduction, gray histogram display, image enhancement and image texture feature extraction process. The following is the introduction of each process.

3.1. Image acquisition

Image acquisition: under the condition of uniform illumination by CCD camera with 2 million pixels.

The original image size is 500 pixels by 509 pixels.

3.2. Gray scale image conversion

In this paper, the conversion form color image to gray image took the weighted average method. It is shown the color image and the transformed gray image of the polished iron pan.

$$f(x,y) = 0.299R(x,y) + 0.578G(x,y) + 0.114B(x,y) \tag{1}$$

In this formula, f(x,y) represents gray image，R, G and B represent for the three component of color image.

Figure 1 the color image (left) and the transformed gray image (right) of the polished iron pan

3.3. Image background deduction

We defined the center of the gray image as the origin, of which a circular area with a radius of 250 pixels was established, then grey values of the area were filled with 0, thus the iron pan image background was formed. After that we took the deduction between the gray image and the background to get rid of the image background interference, as shown in figure 2.

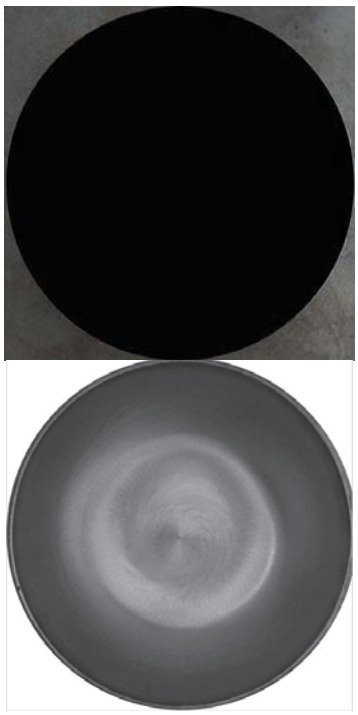


Figure 2 The image background (up) and the image after deduction (down)

3.4. Gray histogram

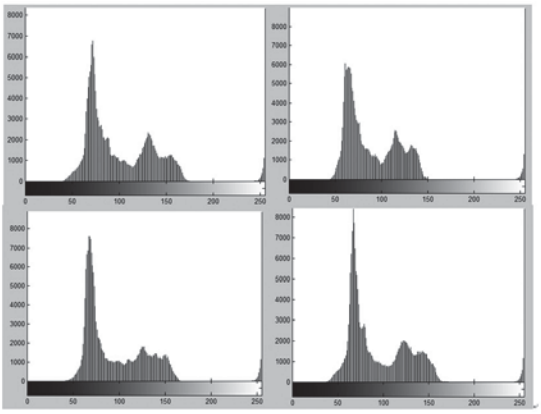


Figure3. 4 gray histogram images of iron pans with good polishing quality

In figure 3, there are 4 gray histogram images of iron pans which are chosen by the experienced worker for their good polishing quality. It can be seen from the figure that these images have very similar grayscale curve characteristics.

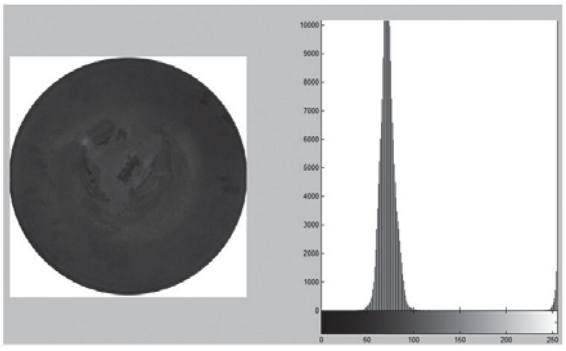


Figure 4 The example of an unpolished iron pan image and its gray histogram

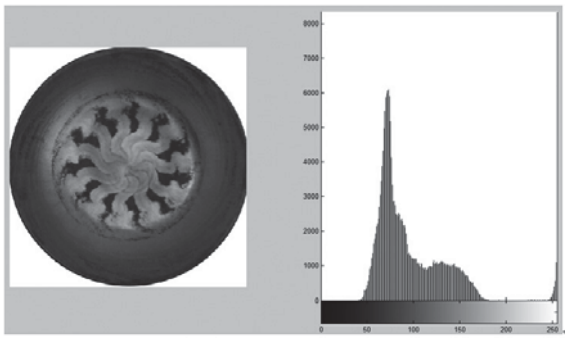


Figure 5 The example of an iron pan with spots at the bottom and its gray histogram

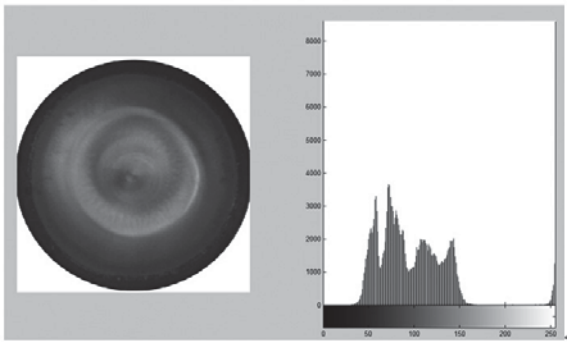


Figure 6 The example of an iron pan with its edge unpolished and its gray histogram

It is shown a iron pan that is chosen without polishing and its gray histogram in figure 4, the gray level histogram of which is significantly different from that in figure 3. There is an iron pan with spots at the bottom and its gray histogram figure 5; the gray values of the histogram range in [50,100] are different from that in figure 3. Figure 6 shows an iron pan with its edge unpolished and its gray histogram. The gray values of the histogram range in [50,100] are different from that in figure 3.

3.5. Image enhancement

Further we will make some enhancement to the image in this step. The low-level grayscale are compressed and the high-level grayscale are extended, so as to reduce the calculation amount of the image texture characteristic.

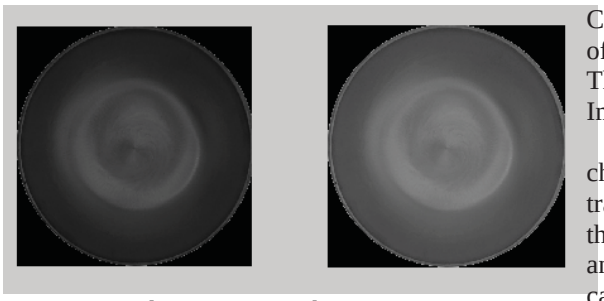


Figure 7 The comparison between iron pan images before and after its enhancement

3.6. Image texture feature extraction

This paper selects information entropy of the gray image after being enhanced as the iron pan image texture feature. The gray histogram is as a probability density function, the image information entropy calculation formula is as follows:

$$H(x,y)= - C \bullet \sum\sum p(xi,yj) \bullet \ln p(xi,yj) \quad (2)$$

In formula 2, $C=\log_2 e$, p(x, y) means the probability density function. We calculated the information entropy of iron pan in Matlab 2014b. The calculation code is just as bellows:

```
function [ h ] = entropy( x, n )   %Image information entropy calculation function.
error (nargchk (1,2,nargin));
if nargin < 2
n = 256;
end
x = double(x);
xh =hist(x(:),n);    % Image gray histogram.
xh = xh / sum(xh(:)); % The probability density function.
i = find (xh);
h=-sum(xh(i).*log2(xh(i))).    %Information entropy calculation.
```

4 The Experiment

This experiment is carried out as a part of the inspection line project of Shanghai Moses Marine Engineering Corp. The iron pan automatic polishing robot system enable to computer to pick pan from the feed line, position it to correct holding place, then hold tight and start the polishing process. The robot will ensure the smooth polishing and handle the abnormal conditions. When the polishing is done, the robot arm will remove the pan from its seat and then put it to the inspection line, where or image processing system is waiting. On this inspection line, the pan will be lined to pass the CCD checking. Disqualified pan will be removed to a separate place for special process. So the image acquisition platform is on this line. The image acquisition includes a uniform light source and a CCD camera with 2 million pixels, the position of which is directly above the circle of iron pan. The image size is 500 pixels by 509 pixels. Image processing software is Matlab 2014 b.

First there were 40 well polished iron pans chosen by the experienced worker. They were transported to the image acquisition platform through the assembly line for image acquisition and processing, after that, 40 sets of data were calculated, as shown in table 1.

Table 1 Forty image information entropy from the chosen good samples

Name	Entropy	Name	Entropy
Sample1	5.239	Sample21	5.040
Sample2	5.180	Sample22	5.020
Sample3	5.325	Sample23	5.121
Sample4	4.978	Sample24	5.234
Sample5	4.986	Sample25	5.221
Sample6	5.023	Sample26	4.933
Sample7	5.224	Sample27	4.987
Sample8	5.420	Sample28	5.039
Sample9	5.316	Sample29	5.289
Sample10	5.211	Sample30	5.158
Sample11	4.988	Sample31	5.195
Sample12	4.956	Sample32	5.261
Sample13	4.977	Sample33	5.163
Sample14	5.194	Sample34	5.145
Sample15	5.187	Sample35	5.056
Sample16	5.086	Sample36	4.964
Sample17	5.062	Sample37	4.921
Sample18	5.256	Sample38	4.995
Sample19	4.949	Sample39	5.023
Sample20	4.962	Sample40	5.010

Then, according to the forty sets of Information entropy data， the average of information entropy is calculated as 5.107。

The following are 20 group test sample entropy data calculated by adopting the method of this article, the data of which come from 20 iron pans without detecting by the experienced worker. And their differential ratios with the average entropy are listed in table 2.

Table 2 Twenty image information entropy from the test samples and their differential ratios with the average entropy

Name	Information entropy	Differential ratio
Test Sample1	5.232	2.4%
Test Sample2	5.177	1.4%
Test Sample3	5.325	4.3%
Test Sample4	4.983	2.4%
Test Sample5	4.946	3.2%
Test Sample6	5.013	1.8%
Test Sample7	5.202	1.9%
Test Sample8	5.320	4.2%
Test Sample9	5.310	4.0%
Test Sample10	5.119	0.2%
Test Sample11	3.685	27.8%
Test Sample12	3.916	23.3%
Test Sample13	2.967	41.9%
Test Sample14	3.394	33.5%
Test Sample15	4.225	17.3%
Test Sample16	4.386	14.1%
Test Sample17	3.162	38.1%
Test Sample18	3.251	36.3%
Test Sample19	3.969	22.3%
Test Sample20	4.132	19.1%

As it Can be seen from table 2, the differential ratios of test sample1 to sample 10 with the average entropy are less than 5%, and that of sample11 to sample 20 are between 14% ~ 42%, thus we concluded that the test sample 1 to 10 were well polished iron pans, which passed through the quality detection, while sample 11 to 20 didn't pass the quality detection. The result was evaluated by the experienced workers of our company, and they confirmed that the recognition rate is 95%.

5 Conclusion

This paper presents a method to measure information entropy, which is related to the inner surface polishing quality of iron pan using image recognition technology. It takes the average information entropy of 40 groups of high quality iron pan sample images as the detection standard. Then the test samples are chosen, the information entropy of which are calculated, and being compared with the standard. Test results are satisfied with controllable fail rate. The comparison result is evaluated by the experienced workers, and the calculation can be embedded to chip computer with CCD and robot arm, and more algorithms or standard can be added to enable its function in the working environment.

6 References

[1] Bin Liu, Qibo Feng, Cuifang Kuang. The review of surface roughness measurement methods. Optical instruments. 2004, 26 (5) : 54 ~ 58.

[2] Xifang Zhang. Research on surface roughness of optical interferometry

measurement [D]. Harbin: Harbin Engineering University, 2006

[3] Qiliang Ni, Bo Chen. Surface roughness measuring based on optical scattering method. Optics and Precision Engineering. 200l, 9 (2): 151-154.

[4] Jiyang Huang. The measurement of surface roughness based on Electronic speckle correlation system [D]. Chongqing: Chongqing University, 2006.

[5] Ulf Persson. Surface roughness measurement on machined surfaces using angular speckle correlation. Journal of Materials Processing Technology. 2006, 180: 233-238.

[6] Cong Ma, Ming Li, Changkun Wang. The visual inspection on mechanical processing surface roughness. Journal of Nanchang institute of Aeronautical Technology (natural science edition), 2003 (3) : 88-94.

[7] Du-Ming Tsai. Jeng-Jong Chen. Jeng-Fung Chert. A Vision System for Surface Roughness Assessment Using Neural Networks. The International Journal of Advanced Manufacturing Technology, 1998, 14, 412-422.

[8] B.Y.Lee, YS.Tarng. Surface roughness inspection by computer vision in turning operations. International Journal of Machine Tools & Manufacture, 2001, 41: 1251-1263.

[9] Shinn-Ying Ho, Kuang-Chyi Lee, Shih-Shin Chen, Shinn-Jang Ho. Accurate Modeling and Prediction of Surface Roughness by Computer Vision in Turning Operations Using an Adaptive Neuro-fuzzy Inference System. International Journal of Machine Tools and Manufacture. 2002. 42(13): 1441-1446.

[10] Kuang-Chyi Lee. Shinn-Jang Ho. Shinn-Ying Ho. Accurate Estimation of Surface Roughness from Texture Features of the Surface Image using an Adaptive Neuro-fuzzy Inference System. Precision Engineering, 2005, 29(1): 95-100.

[11] Rajneesh Kumar. P. Kulashekar. B.Dhanasekar, B.Ramamoorthy. Application of Digital Image Magnification for Surface Roughness Evaluation Using Machine Vision. International Journal of Machine Tools and Manufacture. 2005, 45(2): 228-234.

[12] Yun-Xian Ho. Michael S.Landy, Laurence T. Maloney. How direction of illumination affects visually perceived surface roughness. Journal of Vision. 2006, 6: 634-648.

[13] V.Elango, L.Karunamoorthy. Effect of lighting conditions in the study of surface roughness by machine vision - an experimental design approach. The international Journal of Advanced Manufacturing Technology. 2008, 37: 92-103.

[14] Yujing Wang. A cutting surface roughness measurement based on images processing [D] Harbin: Harbin university of science and technology, 2007

[15] Ghassan A.AI-Kindi, Bijan Shirinzadeh. Feasibility assessment of vision-based surface roughness parameters acquisition for different types of machined specimens. Image and Vision Computing, 2009. 27: 444-458.

[16] Chunya Wu. The grinding surface roughness based on machine vision detection [D] Harbin: Harbin University of science and technology, 2008.

A Real-time Traffic Sign Recognition System Based on Local Structure Features

Kwangyong Lim, Hyeran Byun
Department of Computer Science
Yonsei University
Seoul, Republic of Korea
{ kylim, hrbyun }@yonsei.ac.kr

Yeongwoo Choi
Department of Computer Science
Sookmyung Women's University
Seoul, Republic of Korea
ywchoi@sookmyung.ac.kr

***Abstract*—We present an accurate and efficient system for traffic sign recognition in a real-world driving scene video. The proposed system uses local structure features to achieve high, illumination-invariant accuracy in detection and recognition. We exploit a property of traffic signs, namely, shared boundary shapes, to enhance the speed and accuracy of the detection step. A multi-level SVM structure is employed for stable recognition. The proposed method can process real-world road driving scene video in real time with high accuracy, over 98%, in both detection and recognition**

Keywords—Traffic Sign Detection, Traffic Sign Recognition, 8-bit Modified Census Transform, Illumination invariant, Multi-level SVM

I. INTRODUCTION

Intelligent vehicles are an aggregation of cutting-edge techniques from various fields. Specifically, the Advanced Driver Assistance System (ADAS) is a system that supports drivers while driving. ADAS can be used in various environments that drivers might encounter for better safety of driving. Recently, computer vision methods have been expanding the scope of ADAS to address more problems. Traffic Sign Recognition (TSR) requires high accuracy, low computing cost, and verification based on real-world driving scene video. Germany, [which owns advanced vehicle technology], devotes considerable attention to research for further TSR enhancements, even providing the traffic sign detection and recognition benchmark [1]. The overall TSR procedure consists of detection and recognition. Most efforts to develop traffic sign detection are based on color and shape [2]. However, such methods show unstable accuracy in the presence of change and distortion, which occur quite often because vehicles cover a wide variety of environments.

TSR has been addressed by several approaches. Zaklouta et al. [1] extract features based on encoded gradient histograms. However, they cannot achieve high accuracy in a nighttime driving scenario. Ciresan et al. [3] use the deep learning method to achieve over 99% recognition performance. [However, they have the limitation that their target is the traffic signs that had been detected in advance, and also requires enormous computational costs in both training time and test time], which makes it impractical for application in a real-time system. Lim et al. [4] proposed a speed-limit sign detection and recognition method with illumination-invariant features. However, their method cannot be applied to various traffic signs.

In this paper, we propose a real-time traffic sign detection and recognition system. The proposed method uses the 8-bit Modified Census Transform (8-bit MCT) and its descriptors [4]. 8-bit MCT is a modification of the Census Transform for reducing dimensions while maintaining its properties. Traffic signs are detected by a combination of 8-bit MCT and landmark based detector and efficiently recognized by a multi-level Support Vector Machine (SVM).

II. TRAFFIC SIGN DETECTION BASED ON LANDMARKS

Fröba et al. [5] introduced MCT with the ability to describe 512 types in a 9-bit local area. MCT is suitable for the detection and recognition of objects in forward cameras of vehicles that drive through various lighting conditions because it is invariant to illumination changes. However, traffic signs are special in that they share the same shapes, and we denote them as objects with uniformity. MCT performs unnecessary computation in objects with uniformity, such as traffic signs. Hence, we introduce 8-bit MCT, a reduction of 9-bit MCT that eliminates the center bit. Since the center bit depends on the surrounding bits, 8-bit MCT is more robust to noise than 9-bit MCT. Additionally, because 8-bit MCT can fit in a single byte, it reduces computational cost in training as well as in detection with Adaboost.

Let $N'(x)$ be a 8-neighbor local region of a pixel location x such that $x \in N$; and let $\mathrm{I}'(x)$ be the mean of the pixel intensities in $N'(x)$. 8-bit MCT at x can be written as follows:

$$\Gamma(x) = \otimes_{y \in N'} \xi(\mathrm{I}'(x), \mathrm{I}(y)) \quad (1)$$

where $\xi(\mathrm{I}'(x), \mathrm{I}(y))$ is a comparison function that yields 1 if $\mathrm{I}'(x) < \mathrm{I}(y)$ and 0 otherwise. The symbol $\otimes$ denotes the concatenation operation to form a local structure index bit

Fig. 1. Example of properties of illumination invarianve

vector. If we consider an 8-neighbor 3x3 region, we can determine 256 local structure kernels in total. Fig. 1 shows an example of the process of 8-bit modified census transform. In this figure, $I'(x)$ is rounded down to 191 for rapid integer

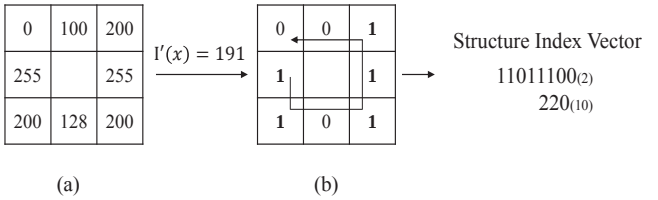


Fig. 2. Example of the process of 8-bit Modified Census Transform (a) Pixel intelsity values, (b) 8-bit Modified Census Transform

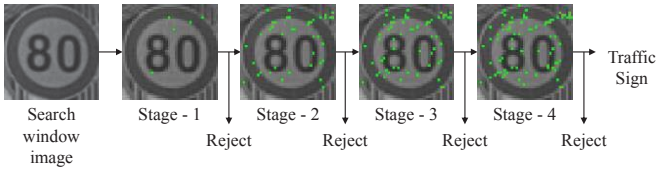


Fig. 3. Example of 4-stage cascade classifier

computation, whereas the original mean value was 191.14.

Each pixel on $N'(x)$ is transformed into binary value by applying $\xi(I'(x), I(y))$. A Structure Index Vector is generated by concatenating the bits incounterclockwise order, as shown in Fig. 2.

We used a 4-stage cascaded classifier trained by Adaboost to detect candidates of traffic signs from the transformed space via MCT. In the final stage, Froba [5] detected objects using the same number of weak classifiers as the number of pixels. However, this approach exhibits low accuracy and requires high computational costs when detecting objects with uniformity, such as traffic signs in a driving video depicting various environments. Weak classifiers in Adaboost provide a class of corresponding pixels [while reporting the importance of the classification.] A weak classifier is considered an important weak classifier if it has low errors in classification.

Hence, we collect only the important classifiers and stop the iteration when the sum of the errors in important classifiers becomes smaller than a pre-defined threshold. Important weak classifiers develop in uniform parts that can be referred as landmarks. Fig. 3 shows an example of proposed 4-stage cascaded classifier and the landmark points are marked in green.

III. Traffic Sign Recognition

A. Feature extraction

The candidate traffic signs detected in real-world road driving scenes may contain false positives due to various environmental changes, especially in urban areas where complex structures reside. Hence, we designed a verification and recognition features, which are vital for a satisfactory performance. Here, the two features based on 8-bit MCT are introduced below.

Verification of sign candidate regions is an important stage of the process. Since an erroneously detected region can also be classified as one of the trained classes, all candidate regions must be validated before the sign classification. Lim et al. [4] propose a simple feature for sign validation based on a histogram of 8-bit MCT as a *verification feature*. Its capability rests on the statistical property of 8-bit MCT kernel indices distributions on candidate regions. The verification feature is defined as below:

$$f = \sum_{i=0}^{255} \frac{1}{sv_{max}} H_i \quad (2)$$

where H is the histogram of the structure index for 8-bit MCT in the image of candidate region, and v_{max} is set to be the most highly value of structure index excluding the zero-structure index. In general, the accumulated value of zero-structure is the highest value and has no strong gradient. Thus, before feature normalization, histograms excluding the zero-structure index must be created.

Finally, verified traffic signs undergo a recognition step. We designed another feature for the further recognition of detailed contents in the signs because the verification feature uses a histogram based on the entire image while ignoring the spatial distribution of edges, thus does not distinguish efficiently. We built separate histograms for each quarter of the detected signs to consider the spatial distribution of edges. The recognition feature is defined as below:

$$d = \sum_{i=1}^{4} \sum_{j=0}^{255} \frac{1}{v_{\max\_i}} H_{ij} \quad (3)$$

TABLE I. SPECIFICATIONS OF THE REAL-WORLD VIDEO TEST SETS

Test set	Frames	Scenes	Driving Environments
Germany-I	4,829	161	Daytime, Rainy, Foggy
Germany-II	1.632	43	Daytime, Rainy, Foggy
Korea-I	4,528	99	Daytime, Rainy, Foggy
Korea-II	3,085	110	Night

TABLE II. SUMMARY OF OUR DETECTION AND RECOGNITION PERFORMANCE

Test set	True Positive	False Negative	False Positive	Precision	Recall	FPS
Germany-I	158	1	2	0.981	0.99	28.0
Germany-II	43	0	0	1.000	1.00	13.4
Korea-I	99	0	1	0.980	1.00	28.4
Korea-II	109	0	1	0.991	1.00	28.8

TABLE III. SUMMARY OF DETECTION AND RECOGNITION PERFORMANCE USING LBP

Dataset	True Positive	False Negative	False Positive	Precision	Recall	FPS
Korea-I	92	3	10	0.9019	0.96	27.7
Korea-II	87	11	8	0.9157	0.88	25.7

where H is the histogram of the structure index for 8-bit MCT in the image of candidate region. In Eq. (3), i is a sub-section index in the region; and $v_{\max\_i}$ is set to be highest value of structure index excluding the zero-structure bin.

B. Classification

Various traffic-sign recognition methods are available. For the traffic sign recognition, a multi-class SVM is useful for which the performance is quite high. However, it cannot distinguish a trained-object from an untrained-object. Thus, we designed a multi-level SVM structure for traffic sign verification and recognition. Fig. 4 shows an example of a multi-level SVM structure. The first-level SVM distinguishes traffic signs erroneously detected regions. The second-level SVM distinguishes speed-limit signs from other limitation signs. The third-level SVM recognizes the speed limits. Each classifier in the third-level SVM recognizes numerals and special symbols. The speed limits of 30, 60 and 80, which look similar, are packed into a single class to be recognized by the fourth-level classifier.

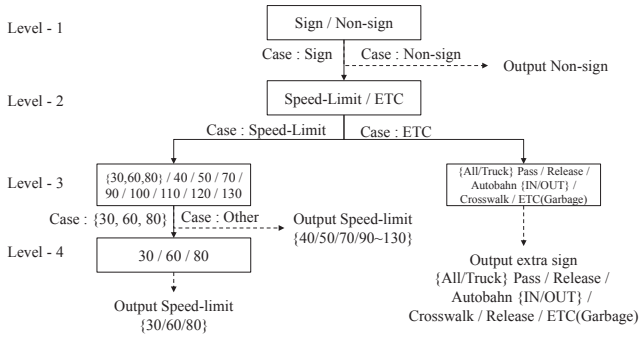


Fig. 4. Example of multi-level SVM structure

IV. EXPERIMENTAL RESULTS

To evaluate the proposed method, we used real-world driving scene video test sets captured by highly dynamic range cameras installed at the backside of Electronic Chromic Mirror (ECM) in Germany and South Korea. The test sets consist of various road environments, including highways, downtowns and local streets. We refer a subsequence of frames containing traffic signs as [scenes]. Each scene is annotated of the traffic sign it contains. The test sets were taken under various driving environments, such as daytime, nighttime, rainy and foggy. This test sets consist of 14,074 continuous image sequences in total with a 1280×672 resolution and 413 driving scenes in total. The test sets consist of four individual sets. Table I shows the specifications of the real-world video test sets and Table II shows the performance of our proposed method. The performances of detection and recognition were measured in four different test sets.

The proposed method shows 99.6% recall with an average of 22.1 fps. Our detection step achieved 99.0% precision. In the dataset Germany-II, the computational complexity costs were twice as high because there were additional signs, i.e., Autobahn In/Out and Crosswalk signs. Our method shows high accuracy even with severe noise in night driving scenes because 8-bit MCT is invariant to illumination changes and robust to noise.

V. CONCLUSION

In this paper, we propose a real-time TSR System for intelligent vehicles. We used 8-bit MCT and its descriptors to solve the problems resulting from illumination changes. The proposed features also show stable performance even in the presence of severe noises. Robustness against occlusion and

distortion is achieved because only the important weak classifiers are collected in the boosting step. Furthermore, we utilize a multi-level SVM structure to improve performance when there are similar traffic signs. Especially in an embedded environment, 8-bit MCT will compute faster than 9-bit MCT because 8-bit MCT is compatible with byte operation.

The proposed method is proven to be accurate, robust, and fast when applied to 4 datasets incorporating various illumination changes and weather. The experiments show credible results in all cases.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation (NRF) of Korea grant funded by the Korean government (No. NRF-2012R1A1A2041343).

REFERENCES

[1] F. Zaklouta, B. Stanciulescu, "Real-Time Traffic-Sign Recognition Using Tree Classifiers," IEEE Trans. Intelligent Transportation Syst. vol. 13, Issue 4, pp. 1507-1514, November 2012.

[2] M. Liang, M. Yuan, X. Hu and J. Li, "Traffic Sign Detection by ROI Extraction and Histogram Features-based Recognition," International Conf. Neural Network, Dallas, pp. 1-8, August 2013.

[3] D. Ciresan, D. Meier, J. Masci and J. Schmidhuber "A Committee of Neural Networks for Traffic Sign Classification," International Joint Conf. on Neural Networks, pp.1918-1921, August 2011.

[4] K. Lim, T. Lee, C. Shin, S. Chung, Y. Choi and H. Byun, "Real-time illumination-invariant speed-limit sign recognition based on a modified census transform and support vector machines," ACM International Conf. on Ubiquitous Information Management and Communication, Jan 2014.

[5] B.Froba, A. Ernst, "Face Detection with the Modified Census Transform," IEEE Conf. Automatic Face and Gesture Recognition, pp.91-96, May 2004.

SESSION

IMAGE SEGMENTATION METHODS + LOW-LEVEL IMAGE PROCESSING AND PREPROCESSING

Chair(s)

TBA

Automatic Tracking of Coronal Mass Ejection using STEREO Red-colored RGB Coronagraph Images

V. Kirnosov[1], L.-C. Chang[1], and A. Pulkkinen[2]
[1]Electrical Engineering & Computer Science Dept., The Catholic University of America, Washington, D.C., USA
[2]NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

Abstract - *In this paper we propose a new technique to segment Coronal Mass Ejection (CME), track its leading edge and estimate the propagation parameters using STEREO A/B SECCHI COR2 red-colored RGB images. The algorithm consists of two modules: pre-processing and classification. The pre-processing module uses a red component of the images to segment CME and produces a set of running-difference binary images. This set is then fed into the classification module that transforms images into polar coordinates, detects CME front edge, and computes the CME propagation parameters. The method was validated using total 30 CMEs, 15 STEREO A and 15 STEREO B events captured in the period from 1 May 2008 to 31 August 2009. The results demonstrate that the proposed method is effective for CME tracking and estimation of propagation parameters. The proposed method can be integrated with the triangulation technique to estimate propagation of CME in three-dimensional space*

Keywords: STEREO; Coronal Mass Ejection; Leading Edge Detection; Space Weather

1 Introduction

Coronal Mass Ejections (CMEs) are large-scale expulsion of plasma and magnetic field from the solar atmosphere [1]. It is well established that they play a crucial role in disturbing the space weather environment as they propagate into the interplanetary medium and interact with the Earth's magnetic field producing strong geomagnetic storms [2]. To study the nature of CMEs NASA and ESA have launched several missions. One of the missions is the twin Solar Terrestrial Relations Observatory (STEREO) [3]. STEREO uses two spacecraft that track CMEs from two different viewpoints and allow for three-dimensional (3D) observation of the events. The images of the CMEs are captured by coronagraphs on board the spacecraft. Individual coronagraphs provide a two-dimensional (2D) representation of the CME 3D structure projected onto the plane of the sky. As a consequence, the propagation parameters like primary angle, angular width, height, and speed are also projected onto this plane and can be measured [4]. The parameters derived for CME using data from both spacecraft allow to reconstruct the propagation path of the CME in 3D space. The reconstruction can be accomplished using a technique of geometric triangulation proposed by Liu et al. [1]. This technique requires the propagation parameters derived by tracking a CME leading edge.

The CME propagation parameters can be estimated manually, for example, using a Stereoscopic CME Analysis Tool (StereoCAT) [5]. This tool allows for manual tracking of the CME leading edge and estimation of the parameters. Although such manual tools provide great help to researchers, the manual processing is a very time consuming task. The result of manual visual analysis is subjective to the experience and perception of the observer. In addition to that, the size of the STEREO COR2 data set is big and has been constantly increasing. All these facts stress a need for the automatic method that will make estimation of the CME kinematic properties faster and more accurate.

There are several automatic methods to detect the CME leading edge have been proposed. Young et al. [6] have proposed a multiscale edge detection method that was further studied by Byrne et al. [7-9], Gallagher et al. [10], Pérez-Suárez et al. [11], and Morgan et al. [12]. This method is employed to process SOHO LASCO data in a CORIMP catalog. Another technique to track the front edge was introduced by Olmedo et al. [11] in SEEDS catalog. But, as noted on the SEEDS website [13], the module to process STEREO SECCHI COR2 images was added to the SEEDS recently. This module has not yet been rigorously tested and the data should be treated carefully. Despite these earlier works, there are currently no automatic, reliable methods to track the CME leading edge and estimate the propagation parameters from the STEREO A/B COR2 images.

In this paper we would like to propose a novel technique for efficient CME segmentation, CME leading edge detection and extraction of the parameters for each 2D CME representation. Our ultimate goal is to perform the automated triangulation of the CME parameters, extracted from the STEREO A/B images, to estimate the CME propagation path in 3D space. We validate our approach using the data set with series of STEREO A/B COR2 RGB images. The validation data set consists of 30 CME events, 15 for STEREO A and 15 for STEREO B captured in the period of the primary phase of the STEREO mission, from 1 May 2008 to 31 August 2009.

2 Method

In this section we provide an overview of the method that we propose for efficient CME segmentation, tracking and estimation of its propagation parameters. The method consists of two modules: pre-processing and classification. The pre-processing module removes the background from the images, segments the CME and extracts its moving features. The set of segmented binary images is then fed into the classification module for CME tracking and estimation of propagation parameters.

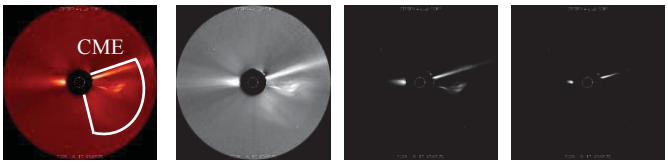
2.1 Data

The data used in this study are the STEREO A/B COR2 red-colored RGB 8-bit images captured in the period from 1 May 2008 to 1 August 2009. The dimension of the images is 256 × 256, temporal cadence is 30 minutes. These image files were generated at STEREO Science Center from science quality 16-bit raw FITS files collected by several NOAA and international ground-tracking stations. Both data sets can be accessed using public and scientific STEREO Web pages at STEREO Science Center [3].

2.2 Pre-processing

The input data set consists of series of 40-48 images. Each image is decomposed into its three channels (red, green, and blue) and only the red channel grayscale image is saved into a new stack that is used for further processing. The green and blue channels are ignored since they are not representative for CME signature in its full extent as can be seen in Figure 1.

In order to reduce noise, the stack of red channel images is processed with a 3×3 kernel of a median filter. The kernel size was experimentally determined to be effective for noise reduction in this application.



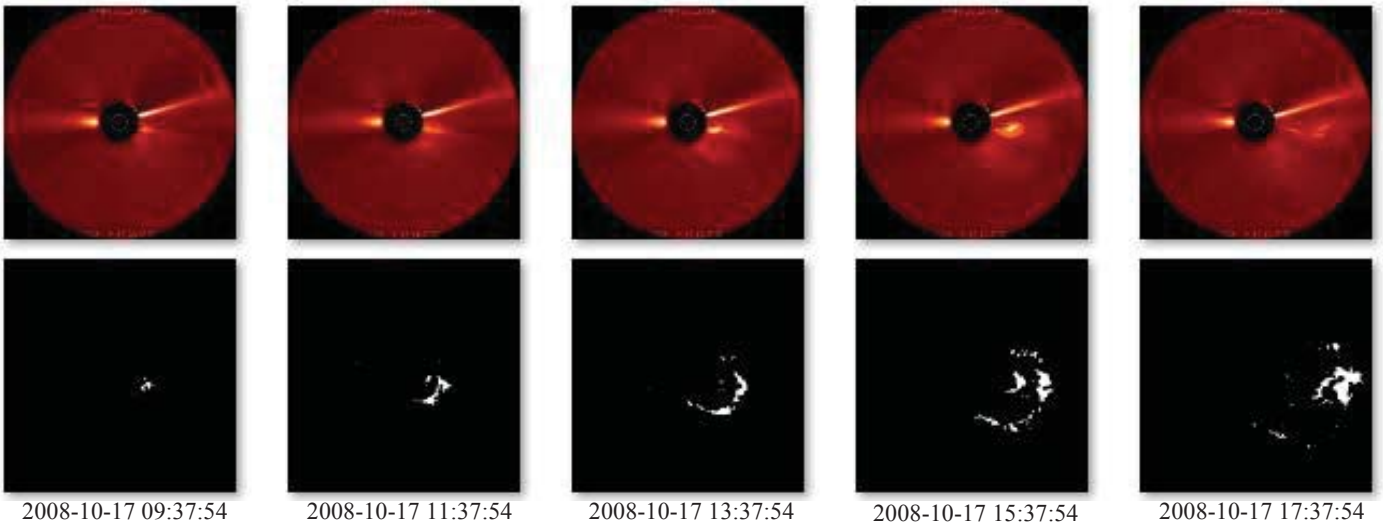
(a) RGB image (b) red channel (c) green channel (d) blue channel

Figure 1. Decomposing a red-colored RGB COR2 image into the R, G, and B channels.

The next step in this module is to remove constant bright areas (e.g. streamers) and segment CME mass. To perform these tasks, the mean image of the stack is first computed. This mean image is then subtracted from each image in the stack thus producing a set of images which contain pixels that differ from the mean image. The streamers and other constant non-CME regions are eliminated by this procedure.

The next step is to extract the moving features that represent the CME mass and remove any unwanted regions that still might be present in the images. The moving features are detected by subtracting the preceding image from the current one, producing a set of running-difference images. A temporary binary version of this set of images is generated and morphological erosion with a structuring element (SE) 3×3 is applied to the images. The pixels that were set to zero during the erosion in the binary images are set to zero in the relevant running-difference images as well. The set of running-difference images is then processed by a histogram equalization procedure to enhance and equalize intensities of the images.

The last step in this module is to convert the set of running-difference images into the binary format and apply a morphological closing to restore the segment's sizes with the same structuring element SE. Figure 2 shows an example of the final pre-processing result.



2008-10-17 09:37:54 2008-10-17 11:37:54 2008-10-17 13:37:54 2008-10-17 15:37:54 2008-10-17 17:37:54

Figure 2. An example of the output produced by the pre-processing module.

2.3 Classification

The input to the classification module is the set of binary running-difference segmented images produced during the pre-processing step (see the example in the lower panel of Figure 2). To classify the groups of pixels into CME and non-CME we selected 2 features: the movement of the segment's leading edge in the outward direction and the distance that it travels in the field of view (FOV) of the image.

First, the running-difference images, which are in [x, y] Cartesian coordinates are transformed into the [r, θ] polar coordinates. This type of transformation has been employed in other CME detection algorithms, for example, in the method proposed by Olmedo et al. [14]. The transformation is done by rotating the image in the clockwise direction starting from the solar north and transforming each angle onto the new [r, θ] FOV [14]. The dimension of the polar transformed image is 128 × 360 where 128 is the half height of the input running-difference image, and 360 is the number of degrees. Figure 3 shows the polar transformed images using the segmented images from Figure 2.

Second, the leading edges of non-zero regions are identified and their coordinates are stored in the 2D array of size 360 × *n* that we refer to as SEG_EDGE matrix where *n* is the number of images in the set. This procedure is done by scanning each image from top to bottom (i.e, toward the Sun), row by row, until a non-zero pixel is identified or last row is reached. The index of the row where the non-zero pixel is identified is assigned to the element [*i, j*] in the SEG-EDGE matrix, where *i* is the current column (angle) index and *j* is the current image index.

Third, the distance that edges travel is estimated as the difference between the lowest and highest positions reached by the edges in the FOV of the image. The estimation is stored in the 1D array of size 360 referred to as ANGLE_PASS array. This is done by scanning through each row of the SEG_EDGE matrix. For each row, the difference between the min and max values is identified and assigned to the *i*-th element in the ANGLE_PASS array, where *i* is the current row (angle) index. These values represent the distances that edges travel along particular angle.

Fourth, the neighboring non-zero entries within the ANGLE_PASS array are identified and combined into the CME candidate groups. The neighboring entries are combined into the sets and then the gaps (number of zero valued entries) between these sets are determined. The sets which have gaps lower than eight are combined together and become one of the CME candidate groups. The eight-index threshold was empirically found to be reliable. All other stand alone sets are also considered as the CME candidate groups and classified as CME or non-CME during the next step.

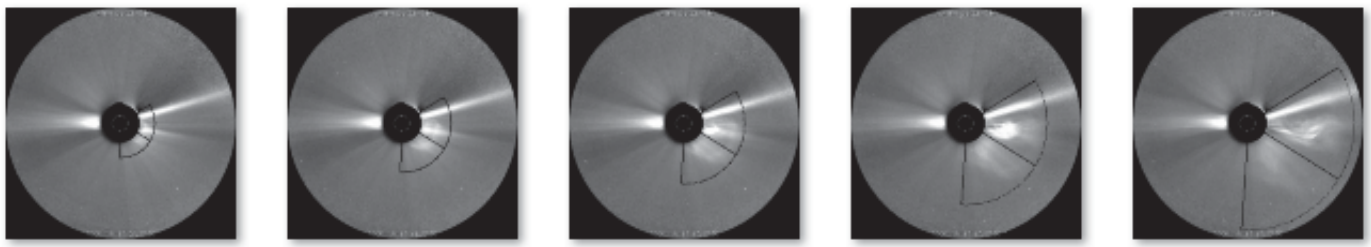
Fifth, the candidate groups are classified as CME or non-CME, so the true CME is determined. This step is done by using a hard threshold method. We have experimentally found that the threshold value of 30 (40% of the STEREO A/B COR2 FOV) is effective for classification in the proposed method. The local maximum is identified for each candidate group. If this value is less than 30 the segment is classified as non-CME and all relevant entries will be ignored in further processing.

The final step is to overlap the results onto the input images in the Cartesian coordinates for visualization purposes (Figure 4) and to estimate the CME propagation parameters (principal angle, angular width, projected height, and speed).

The principal angle parameter is estimated as the midpoint between the beginning and end angles of the CME candidate group. The angular width is computed as the width of the CME candidate group. The projected height is determined separately for each image in the stack as the distance from the position of the Sun to the local max of the CME candidate group. The speed is calculated from the ratio of the distance (projected height) the CME front edge moves to the amount of time it moves.

2008-10-17 09:37:54 2008-10-17 11:37:54 2008-10-17 13:37:54 2008-10-17 15:37:54 2008-10-17 17:37:54

Figure 3. Example of transformation to polar coordinates using the segmented images from Figure 2.



2008-10-17 09:37:54 2008-10-17 11:37:54 2008-10-17 13:37:54 2008-10-17 15:37:54 2008-10-17 17:37:54

Figure 4. Visual result of detections made by the classification module.

2.4 Validation

To validate our approach we used a set of randomly chosen 30 CME events, 15 for STEREO A and 15 for STEREO B, with series of STEREO A/B COR2 red-colored RGB images. These events were captured in the period from 1 May 2008 to 31 August 2009. This timeframe falls into the primary period of the STEREO mission when the separation between the spacecraft was between 50 and 110 degrees and the triangulation of CMEs was considered to be optimal [3]. The chosen events were manually examined by a domain expert using the StereoCAT tool [5] to obtain the CME parameters (principal angle, angular width, and speed). The same set of events was automatically processed by the proposed method, and the estimated CME parameters were compared against the CME parameters computed using the manual method.

3 Results

Figure 5a and 5b show that the principal angle estimates using the automatic and manual methods are in very good agreement. Figure 5c and 5d show that the speed parameter estimation given by the automatic method is also in accordance with the manual analysis. The angular width detections are also closely converged as can be seen in Figure 5e and 5f. It can be noted that the automatic method tends providing bigger angular width than the manual one. The reason for that will be described in the discussion section.

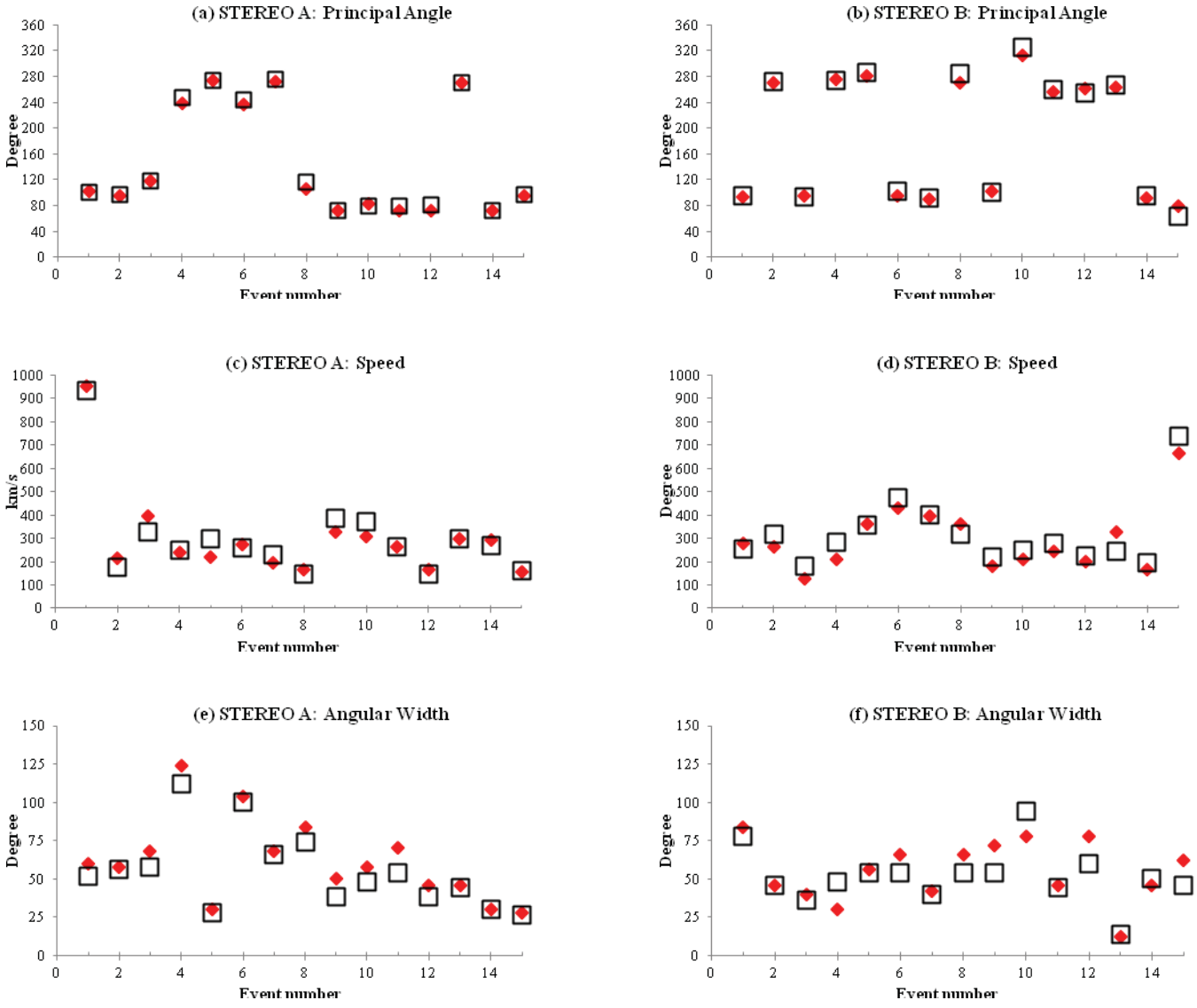


Figure 5. Results of comparison between the automatic and manual methods. Red diamond – automatic detection, black square – manual estimation. a) primary angle (STEREO A); b) primary angle (STEREO B); c) angular width (STEREO A); d) angular width (STEREO B); e) speed (STEREO A); f) speed (STEREO B).

4 Discussion and Conclusion

The propagation path, density, and distribution of each CME are unique. In our validation process, we randomly chose CME events that appear as side-propagating in the STEREO A/B FOV. The halo CMEs were not used because our automatic method and the geometric triangulation technique are not designed to handle this type of events. When the CME propagates in the direction toward or from the observer (i.e., halo CME) its leading edge cannot be seen and this type of events is challenging for detection.

The experiments have demonstrated the efficiency of the proposed algorithm to process the side-propagating CMEs. High agreement in estimation of the primary angle parameter (Figure 5a and 5b) means that the automatic method is very robust in detecting the CME core. In our algorithm, the estimation of the speed parameter fully depends on the CME leading edge detection and tracking. The results in Figure 5c and 5d show that in general, the estimates of the speed parameter using the proposed and manual methods are in close agreement with each other. The differences are due to the subjectivity of the manual tracking of the leading edge and unique perception of CME features by the expert. As can be noted in Figure 5e and 5f, the angular width detected by the proposed method for most of the events is larger when compared with the manual estimation. The investigation in this matter showed that our method is capable of detecting low brightness parts of CME that are hardly visible when observing visually. As a consequence these parts were underestimated during the manual analysis demonstrating the limitation of human eyes in difficult cases where automatic method can provide a great help.

This paper presented a novel automatic algorithm for CME tracking and estimation of propagation parameters. The method includes a unique segmentation approach that allows for effective background removal and extraction of CME signature. To estimate the propagation properties a novel algorithm to detect the front edge of the CME is proposed. These improvements allow us moving toward development of automatic method for estimation of 3D CME properties.

5 Acknowledgement

The STEREO/SECCHI data used here are produced by an international consortium of the Naval Research Laboratory (USA), Lockheed Martin Solar and Astrophysics Lab (USA), NASA Goddard Space Flight Center (USA) Rutherford Appleton Laboratory (UK), University of Birmingham (UK), Max-Planck-Institut für Sonnensystemforschung (Germany), Centre Spatiale de Liège (Belgium), Institut d'Optique Théorique et Appliqué (France), Institut d'Astrophysique Spatiale (France).

6 References

[1] Liu, Y.; Davies, J.A.; Luhmann, J.G.; et al. "Geometric triangulation of imaging observations to track coronal mass ejections continuously out to 1 AU"; Astrophys. J. Lett., 710, 1, L82-L87, 2010.

[2] Srivastava, N.; Inhester, B.; Mierla, M.; et al. "3D Reconstruction of the Leading Edge of the 20 May 2007 Partial Halo CME"; Solar Phys, 259, 213-225, 2009.

[3] Kaiser, M. L.; Kucera, T. A.; Davila, J. M.; et al. "The STEREO Mission: An Introduction"; Space Sci. Rev., 136, 1-4, 5-16, 2008.

[4] Mierla, M.; Inhester, B.; Antunes, A.; et al. "On the 3-D reconstruction of Coronal Mass Ejections using coronagraph data"; Ann. Geophys, 28, 203-215, 2010.

[5] Stereoscopic CME Analysis Tool (StereoCAT). On-line Web page. http://ccmc.gsfc.nasa.gov/analysis/stereo/

[6] Young, C. A.; Gallagher, P.T. "Multiscale Edge Detection in the Corona", Solar Phys., 248, 457-469, 2008.

[7] Byrne, J. P.; Gallagher, P. T.; McAteer, R. T. J.; et al. "The kinematics of coronal mass ejections using multiscale methods". Astron. Astrophys., 495, 1, 325-334, 2009.

[8] Byrne, J. P.; Maloney, S. A.; McAteer, R. T. J.; et al. "Propagation of an Earth-directed coronal mass ejection in three dimensions". Nat. Commun., 1, 74, 2010.

[9] Byrne, J. P.; Morgan, H.; Habbal, S. R.; et al. "Automatic Detection and Tracking of Coronal Mass Ejections. II. Multiscale Filtering of Coronagraph Images". Astrophys. J., 752, 145, 2012.

[10] Gallagher, P. T.; Young, C. A.; Byrne, J. P.; et al. "Coronal mass ejection detection using wavelets, curvelets and ridgelets: Applications for space weather monitoring". Adv. Space Res., 47, 12, 2118-2126, 2011.

[11] Pérez-Suárez, D.; Higgins, P. A.; Bloomfield, D. S.; et al. "Automated Solar Feature Detection for Space Weather Applications". Book chapter in "Applied Signal and Image Processing: Multidisciplinary Advancements", 207-225, edited by R. Qahwaji, R. Green and E. Hines, 2011.

[12] Morgan H.; Byrne, J. P.; Habbal, S. R. "Automatically Detecting and Tracking Coronal Mass Ejections. I. Separation of Dynamic and Quiescent Components in Coronagraph Images". Astrophys. J., 752, 2, 14, 2012.

[13] SEEDS catalog. On-line Web Page. http://spaceweather.gmu.edu/seeds/

[14] Olmedo, O.; Zhang, J.; Wechsler, H; et al. "Automatic Detection and Tracking of Coronal Mass Ejections in Coronagraph Time Series". Solar Phys, 248, 2, 485-499, 2008.

Document image segmentation by a cascade of Pseudo-Word contextual labelings

A. Belaïd[1], A. M. Awal[1], S. Kébairi[2] and V. Poulain d'Andecy[2]
[1] LORIA, Campus scientifique, 54500 Vandœuvre-Lès-Nancy, France
{abdel.belaid, ahmad-montaser.awal}@loria.fr
[2] ITESOFT, Parc d'Andron, Le Séquoia, 30470 Aimargues,
{saddok.kebairi, vincent.poulaindandecy}@itesoft.com

Abstract - *The aim of this work is to classify the document content in handwritten (H), printed (P) and noise (N). In a first step, based on smearing, writing pseudo-lines and pseudo-words are extracted. The latters are classified in (P, H, N) using SVM with a Gaussian kernel. In a second step, the context of pseudo-words is examined along their pseudo-lines, spread the type of script and correct errors. First, the word separation is modeled by a conditional random field. Then, the context is extended using a cascade of contextual propagation modules. Our system achieves a very good pseudo-word classification rate for both handwritten and printed text (97.3% and 99.5% respectively) for a total of 98.7%.*

Keywords: pseudo-line, pseudo-word, contextual labeling, CRF, class dominance propagation

1 Introduction

Administrative documents systems are often faced with a variety in the type of document, in terms of content, quality and structure mixing handwritten and machine printed information (see Figure 1). Documents can be skewed, contain noise and different objects like graphics, signatures, logos, annotations, etc. The document analysis system must rid the text of graphic objects and noise, extract the annotations whatever the document language and thus separate the content in typewritten and manuscript to allow specialized OCRs to handle them. This separation must be done regardless of the nature of documents, fully structured, semi-structured as forms, tables, etc. This variability prohibits any method requiring empirical and specialized parameters, and of course the processing time is a critical constraint for industrial application.

Scientific publications exhibit an abundant research on document segmentation. In some papers and surveys [2, 3, 6, 14], we can find a broad view of what currently exists in writing separation. The general principle is to extract first basic units of type: line, word or character and extend this classification in the neighborhood to prevent misclassification and find homogeneous writing groups, with for instance a post-processing contextual re-labeling

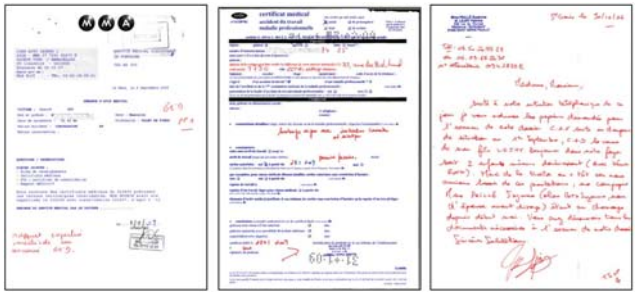


Figure 1: Example of recognized documents: printed text in blue, handwritten in red and noise in black color.

Regarding the best representation level to choose as basic unit for script separation, text lines are too global and may contain both types of scripts. Thus, it is not appropriate to most of real life applications. On the other hand, very small basic units (such as characters or connected components) might be ambiguous and do not hold enough discriminant information.

We introduce the notion of pseudo-word (PW) [1]. It represents a part of word which is a homogeneous "quantity of writing" in a text line, in terms of spaces and sizes inter PWs and intra PW. It seems to be more stable than all other representations, but this requires locating first pseudo-lines (PLs) and calculating space and density statistics to help PW segmentation.

This paper is organized as follows: in section 2, the proposed approach is described giving the ancient modules [1] with new contributions in the contextual re-labeling. In section 3, we expose the other new contributions related to PW regularity exploitation and ambiguity class introduction. We conclude and give perspectives of this work in section 4.

2 Proposed approach

Figure 2 shows the different modules of the proposed system. We will describe them below.

2.1 The dataset

The dataset represents business documents obtained from ITESOFT. For training, we use a set of 107 documents for a

total of 32.715 PWs. For the tests, 202 documents are used for a total of 77.964 PWs. All document images are labeled at the pixel level. For the evaluation, we use the same measure proposed in [13].

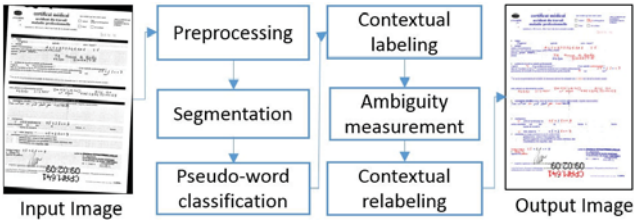


Figure 2: Different modules of the system.

2.2 Segmentation

As reported in [1], the preprocessing corrects the document skew by RAST algorithm [17] and deletes the current "salt and pepper" noise by a k-fill based algorithm [18]. The segmentation module provides regular textual units approaching the final targets. This is operated by a two level segmentation. First, PLs are delimited using a smearing technique. Then, PWs are extracted in each PL, using adapted space threshold to each PL performed by space histogram examination.

2.3 Classification

We started by a basic method using mutli-class support vector machines (SVM) [16] on PW images, with 137 features inspired from the literature [4, 6, 12]. The results are reported in Table 1, line 1.

2.4 Contextual re-labeling

The contextual labeling is achieved using different grouping techniques: local neighboring and neighboring within pseudo-lines. The grouping by local neighboring was tested using K_{NN}, confidence propagation and a new contribution: conditional random fields (CRF). We also introduce a new grouping method based on pseudo-lines neighbors achieved by probabilistic or deterministic models.

In K_{NN}, k nearest neighbors are taken into account if they are closer than a pre-defined threshold and the accumulated number of their pixels is significant compared to the number of pixels of the main PW. A constraint is added to avoid the labeling of small components: we check whether the accumulated number of pixels of the neighbors is significant compared to the number of pixels of the main component (See Table 1, line 2).

In grouping by confidence propagation, the idea is to check the confidence of the nearest horizontal neighbor of a given PW. If the latter is stronger than that of the PW, the neighborhood class is assigned. A Gaussian function weighs the neighbor confidence by its distance to the PW. The nearest the neighbor is, the more impact it has (see Table 1, line 3).

Finally, in grouping by CRF [15], the separation problem is modeled as the search of best configuration of a label field given observations (similar as in section 2.4.1). The model is defined by the product of exponential linear combination of k functions called 'feature functions'. These functions can be modeled by discriminant classifiers like Multi Layer Perceptron (MLP) or SVM (see Table 1, line 4).

Contextual neighborhood of a given PW, as defined above, only delivers local information. We have now extended the context to the entire PL. We define the dominant class in a PL as the class with the highest cardinality. It is obvious that we cannot assign the same label systematically to all the components of a PL. We explore below the use of the dominant class by probabilistic and deterministic models.

2.4.1 Probabilistic model

For all PWs composing a PL, a classification confidence is estimated, using a CRF model. We combine a local classifier (an SVM on local features) and a contextual classifier (an MLP on the neighborhood with respect to the dominant class). The extracted features are: a) normalized class cardinalities in the PL, b) structural features measuring the homogeneity of each PW with the dominant class, such as: height ratio, density ratio, connected components (CC) count ratio, and inter-CC distance variance ratio (see Table 1, line 5).

Table 1: Recognition score for different used methods

System	H%	P%	N%
Without contextual re-labeling [1]	97.7	96.5	**94.3**
k-NN with constraints [1]	95.5	97.5	92.3
Confidence propagation [1]	97.8	96.6	94.0
CRF [20]	98.5	97.1	94.2
Grouping by PL (CRF): Probabilistic Model [20]	98.9	97.5	93.5
Grouping by PL: Deterministic Model [20]	98.3	99.2	87.9
Improved segmentation + diacritic linkage [20]	**99.1**	**99.2**	90.1

2.4.2 Deterministic model

The dominant class label is associated to a PW if it verifies 1) the classification confidence is low or 2) the PW has a similar height as the dominant class in which case the classifier decision is ignored. This latter case is inspired from printed text lines where most of the words have a height similar to the height of the PL reflecting the regularity of the text line (see Table 1, line 6).

Table 2: Comparison with literature systems

System	Description					Pseudo-word_rate			Pixel_rate			
	Features	Classifier	Neighbors	Re-labeling	Docs	H%	P%	All%	H%	P%	N%	All%
Kandan et al. [12]	7	SVM	Delaunay triangulation	Majority voting	150	-	-	93.2	-	-	-	-
Peng et al. [7]	12	G-means	4-NN	MRF	82	93.8	95.7	95.5	-	-	-	-
Shetty et al. [9]	23	CRF	6-NN	CRF	27	-	-	-	94.8	98.4	89.8	95.8
Zheng et al. [4]	31	Fisher classifier	Horizontal Left – right	MRF	94	93.0	98.0	97.8	-	-	-	-
Grouping by pseudo-lines: deterministic	137	SVM	Pseudo-line	Dominant class	202	**97.3**	**99.5**	**98.7**	**99.1**	**99.2**	**90.1**	**96.8**

2.5 Segmentation improvement

We have frequently observed that our rule based on the proximity of the CCs in the same line, is not sufficient. Indeed, often in the presence of signatures or logos, several lines are attached. Thus, we proposed to solve this issue by introducing additional connectivity conditions such as two CCs will never be merged if they do not have a minimum horizontal overlap (less than 30% of the CC maximum size). In addition, to address the case of diacriticals, we link small connected components to the nearest PW (Table 1, line 7).

2.6 Contextual relabeling evaluation

The three proposed methods outperforms those proposed in [1] for the handwritten class. Indeed, the use of horizontal neighborhood overcomes an important drawback of the KNN method, where a nearby printed text can affect a handwritten PW and assign it to the wrong class. The use of the new contextual neighborhood based on PLs allows improving the performance of the CRF model based on the local neighborhood definition (98.5% to 98.9% for the handwritten class and 97.1% to 97.5% for the printed class).

On the other hand, the deterministic model allows improving significantly the printed text separation rate to 99.2% with a good rate for handwritten class. In addition, the deterministic model is better in correcting the small PWs (diacritics) which results in a very good PWs rate compared to the other methods. Furthermore, segmentation method improvements enhance the deterministic model rates to 99.1%, 99.2% and 90.1% for the handwritten, printed and noise classes respectively. At this level of the system, system's global performance is competitive compared to the state of the art, see Table 2.

3 Other new contributions

3.1 Feature selection

A concern is to improve the performance of the classifier in terms of run time and complexity. One improvement was to reduce the number of features. To this end, Fisher score and ReliefF methods have been adopted [19] and features are passed from 137 to 48 (see Table 3).

Table 3: Features selected by ReliefF

	N°	N° selected	Selected features
Structural descriptors	8	2	1, 2
Hu Moments	7	0	-
Vertical profile	1	1	All
Horizontal profile	4	4	All
Pixel distribution	1	1	All
Mean line	8	5	1, 2, 3, 4, 5
Run Length	20	17	All except {4, 14, 19}
Crossing count	10	9	All except {10}
Vertical segments	2	1	{2}
Texture 1D	16	5	{6, 9, 10, 13, 14}
Texture 2D	60	3	{15, 30, 45}
total	137	48	

Both methods have been evaluated using Tanagra machine learning tool. As shown in Table 4, his reduced the complexity by a third with a very small impact on the system performances.

Table 4: Class confidence scores

N° Features	Printed%	Handwritten %	Noise%
137	99.2	99.1	90.1
48	99.1	98.3	90.7

3.2 Ambiguity layer

Despite our classification performance, an issue is the misclassification between Printed and Handwritten PW. We propose a fourth output layer in addition to layers H, P, N to collect a maximum of misclassifications and errors produced by the modules. We consider each processing module as a classifier involved in a classifier cascade structure which takes a PW as input to associate it with one of the three classes (H, P, N) as output. The ambiguity calculation is mainly based on classification probabilities given by the first PWs classifier (SVM in our case), summarized in Figure 3.

3.2.1 PW ambiguity

With the output of the first PW classifier (SVM), probabilities of associating a PW to one of the three classes H, P, N are provided: $\{p_1, p_2, p_3\}; \sum_i p_i = 1; p_1 > p_2 > p_3$

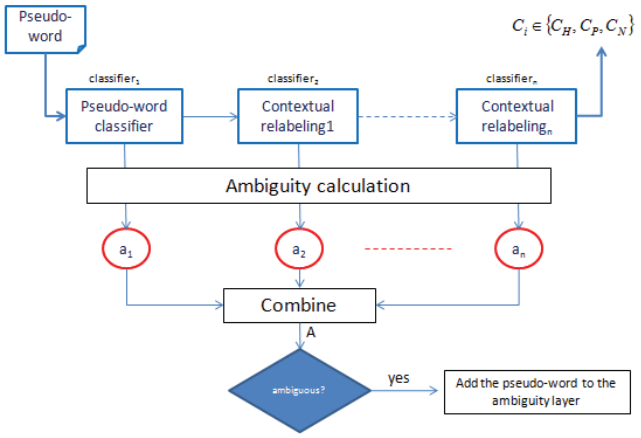


Figure 3: Ambiguities cascade classifier combination

We select the winner class as the class with the highest probability: $p_w=p_1$. The global classification ambiguity 'A' is the result of combining ambiguities of each of the classifiers $A=f(a_i); a_i \in [0,1]$. The ambiguity of the SVM classifier is given by the importance of the second class, normalized in the interval [0,1]: $a_{SVM} = 2.p_2$.

During all post-processing modules, the original label of the PW could be turned into a new one C_l depending on its contextual neighborhood. Thus, the ambiguity is calculated as the difference between the probability of the winner class and the probability of the PW being originally associated to the new class C_l: $a_{classifier_i} = p_w - p_{C_l}$.

3.2.2 Ambiguity fusion

Different ambiguities are then combined using a classical classifier combining methods as in [21]. We have implemented three methods: average, maximum and production. Finally, the PW with a global ambiguity higher than the average ambiguity of all the PWs in the current document is considered ambiguous and is added to the ambiguity layer.

3.2.3 Ambiguity evaluation

To evaluate the efficiency of the ambiguity layer, the classification rate is calculated ignoring the ambiguous PWs. Table 5 summarizes the global performance of the system when applying the three combination methods.

Table 5: Class confidence score with the combination methods

	P%	H%	N%
Original	99.1	98.3	90.7
Average	**99.7**	99.5	98.5
Maximum	99.7	**99.6**	**98.9**
Product	99.6	99.2	93.7

We can notice a better accuracy, the ambiguity layer enables to remove classification errors and increase the classification rate for the three classes. However, this evaluation is not sufficient since we have to remove a maximum of misclassified PWs but a minimum of PWs correctly classified. The Table 6 shows the rate of pixels associated to the ambiguity layer, and the rate of misclassified pixels correctly associated to the ambiguity layer.

Table 6: Effect of ambiguity layer pixel association

	Pixels associated to the ambiguity layer			Misclassified pixels correctly associated to the ambiguity layer		
	P%	**H%**	**N%**	**P%**	**H%**	**N%**
Average	11.2	24.4	27.3	**1.6**	3.1	48.3
Maximum	**10.3**	**22.7**	26.5	1.4	**3.9**	**52.2**
Product	11.1	24.6	**24.3**	1.0	1.3	35.6

We can explain above tables as follows: e.g. in the row of the combination method "maximum", we can notice that the classification accuracy of class P is improved to 99.7% and 1,4% of misclassified P are saved to the ambiguity layer. But the cost is10.3% of ambiguous printed pixels.

Hence the benefit for the end-user is to manage the accuracy versus the recall by merging when needed the ambiguity layer with another layer for instance with the layer H to maximize the handwriting annotation detection.

3.3 Detection of PW regularity

The objective is to detect misclassified printed PWs (classified as handwritten or noise). We started from the idea that in a printed word, the majority of characters are aligned within the central band, sharing the same height. In our system, a PW (PW) is a set of connected components (CCs): $pw = \{CC_1, CC_2, ..., CC_m\}$. We define the connected components with the same height as the average height (h) of CCs: $pw' = \{CC_i, h(CC_i) \approx H\}$. A PW is ambiguous if it verifies these conditions:

- C1: $\|pw\| > 2 \wedge \|pw'\| > \frac{2}{3}\|pw\|$
- C2: CCs with height similar to the central band height must be aligned to the upper and lower lines of the band.

Applying this method on the test database, 20443 PWs are declared in accordance with the two conditions. Among these detected PWs, 17,239 are correctly classified as printed (84.32%) of which 17,099 PWs were correctly classified as

printed (99.18%). As a result, only 140 misclassifications (0.82%) will be saved. In addition, 3204 PWs (15.68%) are the subject of false positives (regular manuscript or noise PWs reported as printed by the detection). This naïve approach keeps sense in case of maximization of printed detection but need further work to develop a robust improvement.

4 Conclusions

A Handwritten/printed/noise separation system is presented in this paper. A distance based segmentation method allows regrouping CCs to obtain PLs and PWs. A multiclass SVM is used as PWs classifier. PL based methods are proposed to correct classification errors (probabilistic and deterministic model). An ambiguity layer is introduced to manage misclassifications depending end-user needs. The prospects of this work are to increase the testing on other datasets and continue to ensure system stability.

5 References

[1] A. Belaïd, K. Santoch and V. Poulain d'Andecy, "Handwritten and Printed Text Separation in Real Document," *Machine Vision Applications,* vol. 2, 2013.

[2] U. Pal and B. B. Chaudhuri, "Machine-printed and hand-written text lines identification," *Pattern Recognition Letters,* vol. 22, pp. 431-441, 2001.

[3] E. Kavallieratou and S. Stamatatos, "Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics," in *International Conference on Pattern Recognition*, Cambridge, UK., pp. 437-440, 2004.

[4] Y. Zheng, H. Li and D. Doermann, "Machine Printed Text and Handwriting Identification in Noisy Document Images," IEEE Transactions on *Pattern Analysis Machine Intelligence,* vol. 26, pp. 337-353, 2004.

[5] J. K. Guo and M. Y. Ma, "Separating Hadwritten Material from Machine Printed Text Using Hidden Markov Models," in *International Conference on Document Analysis and Recognition*, Seattle, WA, USA, pp. 439-443, 2001.

[6] L. F. da Silva, A. Conci and A. Sanchez, "Automatic Discrimination between Printed and Handwritten Text in Documents," in *Brazilian Symposium on Computer Graphics and Image Processing*, Rio de Janeiro, Brazil, pp. 261-267, 2009.

[7] X. Peng, S. Setlur, V. Govindaraju and R. Sitaram, "Handwritten text separation from annotated machine printed documents using markov random," *International Journal on Document Analysis and Recognition,* vol. 16, pp. 1-16, 2011.

[8] K. Zagoris, L. Pratikakis, A. Antonacopoulos, B. Gatos and N. Papamarkos, "Distinction between handwritten and machine-printed text based on thebag of visual words model," *PatternRecognition,* vol. 47, pp. 1051-1062, 2014.

[9] S. Shetty, H. Srinivasan and S. Srihari, "Segmentation and Labeling of Documents using Conditional Random Fields," in *Proc. SPIE 6500, Document Recognition and Retrieval XIV*, 65000U, 2007.

[10] M. Shirdhonkar and M. B. Kokare, "Discrimination between Printed and Handwritten Text in Documents," *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition,* vol. 3, pp. 131-134, 2010.

[11] K.-C. Fan, L.-S. Wang and Y.-T. Tu, "Classification of machine-printed and handwritten texts using character block layout variance," *Pattern Recognition,* vol. 31, pp. 1275-1284, 1998.

[12] R. Kandan, N. K. Reddy, K. R. Arvind and A. G. Ramakrishnan, "A robust two level classification algorithm for text localization in documents," in *International conference on Advances in visual computing*, Lake Tahoe, NV, USA, pp.96-105, 2007.

[13] F. Shafait, D. Keysers and T. M. Breuel, "Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms," *Pattern Analysis and Machine Intelligence,* vol. 30, pp. 941-954, 2008.

[14] Ranjeet Srivastva, Aditya Raj, Tushar Patnaik, Bhupendra Kumar, "A Survey on Techniques of Separation of Machine Printed Text and Handwritten Text," *International Journal of Engineering and Advanced Technology*, Volume-2, Issue-3, February 2013.

[15] S. Nicolas, J. Dardenne, T. Paquet et L. Heutte, "Document Image Segmentation Using a 2D Conditional Random Field Model," in *Ninth International Conference on Document Analysis and Recognition*, Curitiba, Brazil, pp. 407-411, 2007.

[16] C.-W. Hsu and C.-J. Lin. "A comparison of methods for multi-class support vector machines", *IEEE Transactions on Neural Networks*, 13, pp. 415-425, 2002.

[17] van Beusekom J., Shafait F., Breuel T., « Combined orientation and skew detection using geometric text-line modeling », International Journal on Document Analysis and Recognition, vol. 13, p. 79-92, 2010.

[18] Chinnasarn K., Rangsanseri Y., Thitimajshima P., « Removing Salt-and-Pepper Noise in Text/Graphics Images », The Asia-Pacific Conference on Circuits and Systems, p. 459-462, 1998.

[19] I. Kononenko, E. Simec, M. Robnik-Sikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF (1997), Applied Intelligence, 7(1), p39-55.

[20] A. M. Awal, A. Belaïd, V. Poulain d'Andecy, Handwritten/printed text separation Using pseudolines for contextual re-labeling14th International Conference on Frontiers in Handwriting Recognition, Crete, 2014, pp. 29-34.

[21] L. A. Alexandre, A. C. Campilho, M. kamel, On combining classifiers,using sum and product rules, Pattern Recognition letters, n. 22, 2011, pp. 1283-1289.

Distribution Matching and Active Contour Model based Cardiac MRI Segmentation

Ruomei Wang, Jinping Feng, Zhong Wang*, Qiuyuan Luo
School of Information Science Technology, Sun Yat-sen University, Guangzhou 510006, China

Abstract – *In this paper an approach to extract the Left Ventricle (LV) endocardium contour by proposing an improved Distribution Matching (DM) algorithm is presented, the main idea of which is to match the distribution of grayscale and distance constrained by gradient vector flow (GVF) force field between the sample and the images input. The endocardium contour is then regarded as the initial iterative curve of Active Contour Model (ACM) to obtain the epicardium contour with a method of local circular constraints and adaptive changing parameter of external force field. The approach has been applied on data of 20 subjects, and results have demonstrated simpler user interaction, more robustness and higher segmentation accuracy compared to LV segmentation methods published previously.*

Keywords: distribution matching, image, contour model

1 Introduction

The technology of computer image processing and computer vision have been applied in different areas. In the real world, Heart Failure (HF) is one of the leading causes of death. According to statistics, there are 1200~1500 patients with HF all over the world, and up to 5% patients with HF among one billion of the population in European and American. Thus it's significant to develop an automatic detection method for HF's early diagnosis. Paulus et al. [1] and Doughty [2] detailed the importance of utilizing LV Ejection Fraction (LVEF) which can be obtained by Cardiac Magnetic Resonance Images (CMRI) segmentation, in HF's diagnosis. However, although the automatic segmentation of CMRI has been intensively studied in computer vision, because most existing LV segmentation algorithms requires either intensive user inputs, extensive training sets, or with low accuracy and efficiency produced, it might limit its further application in the medical process. To overcome these problems, in this paper, an improved Distribution Matching (DM) algorithm and Active Contour Model (ACM) are proposed to be used to extract the LV endocardium and epicardium. The experiment results show that our method is more efficiently.

2 Related works

The segmentation algorithms of medical images can be categorized in terms of the mathematical model they used, such as clustering analysis method, statistical method, graph cut method, deformable model methods, cardiac model-based method, distribution matching method and etc.

Clustering analysis method is often applied to MRI segmentation with Gaussian Mixture model (GMM) and K-means. For instances, Lynch et al. [3] used grayscale feature and K-means to cluster, and extracted the myocardium area. Klann et al. [4] used K-means constrained by boundaries (edge features) for breast MRI segmentation. But cluster analysis method is easily influenced by uneven grayscale distribution and need special threshold configuration.

Statistical method mainly involves two kinds of random field, Markov Random Field (MRF) and Conditional Random Field (CRF), such as the image segmentation method proposed by Chittajallu et al. [5], who utilized MRF with shape prior, edge prior and label prior for Computed Tomography images segmentation. When working with graph cut method, it performs well but needs extra computation. Similarly, Grosgeorge et al. [6] proposed a method to calculate the summary of grayscale energy and classification tags energy by graph cut method to obtain the optimal segmentation. And Mahapatra et al. [7] detailed a method for CMRI segmentation by combining graph cut method and shape priors.

Deformable model method is used to solve image segmentation problems with partial differential equations, by modeling the CMRI with energy function by the internal energy and external energy for numerical solutions. For instances, Wang et al. [8] combined ACM and GMM to obtain the epicardium and endocardium contour of LV, Klann et al. [9] proposed a Mumford-Shah Level-Set approach for tomography's reconstruction and segmentation, Lee et al. [10] detailed a method for CMRI segmentation by area growing and ACM, and etc. Deformable model method sometimes performs well, but needs intenin sive user inputs[11].

Cardiac model-based method is used to map a three dimensional cardiac model to a two dimensional CMRI by calculating the internal energy and external energy of the CMRI[12]. But it requires to generate a three dimensional CMRI model first, and the reflection also need extra computation although by using principal component analysis[11].

Distribution matching method is used to convert image segmentation problems to the functional optimization problems, by establishing energy functional equations and using Euler-Lagrange partial differential equation to solve the problems. Similarly, Njeh et al. [13] used two constraints to

solve the brain tumor segmentation problem, one is the intensity distribution prior, and another is the smoothness prior; and Nambakhsh [14] proposed a convex relaxed distribution matching (CRDM) method, which utilized three kinds of constraints to obtain the myocardium area: intensity, distance and edge. Distribution matching method cen be performed well because it converges to a distribution which has the most similarity of the distribution calculated by the sample. But it needs extra computation and sometimes too easy to be influenced by the sample distribution, which make the converged curve lack its meaningful description.

In this paper, a CMRI segmentation method combining with DM method is proposed to reduce amount of dataset, which only needs one CMRI input, and ACM is reported to be used to obtain a smooth and accurate contour with higher accuracy and more robustness. The improved DM algorithm is based on the grayscale and gradient features of LV chamber constrained by the gradient vector flow (GVF) force field. The original iterative curve of ACM is then obtained by using the contour of endocardium. A local circular constraint and a method of adaptive changing parameter of external force field are also proposed to improve the ACM in LV epicardium extraction. After being performed on data of over 20 subjects[1], the proposed algorithm has been demonstrated its robustness and accuracy when uneven grayscale distribution and mastoid muscle prompted. Besides, the character of simple user interaction has great application in HF's early diagnosis.

3 The improved left ventricular segmentation algorithm

There are two steps involved for the proposed LV segmentation method: 1) an image from base of heart with endocardium contour drawn from the CMRI sequence of a given subject is needed to be the sample to calculate the grayscale distribution, then match the distributions constrained by distances and GVF feature between sample picture and input picture every iteration to perform the endocardium contour extraction; 2) convert the results in step 1 as the initial iterative curve, and an improved ACM is used to obtain the epicardium contour.

3.1 Improved distribution matching

DM algorithm has been demonstrated its outstanding images segmentation results, such as in CRMI segmentation[14, 15] and in daily scene images segmentation[16]. Therefore, we propose a DM-based method to extract the chamber area.

Let $E_{\text{gray}}(k)$ be grayscale energy, $E_{\text{distance}}(k)$ be the distance energy, and $c_{\text{gv}}(k)$ be the edge energy of the input images. $E_{\text{gray}}(k)$ means the similarity of grayscale distribution between chamber area of the sample image, defined as $T_{\text{sam}}(x)$, and the image input later, defined as $T_{\text{in}}(x)$. If these two distributions are similar enough, $E_{\text{gray}}(k)$ will be small enough. $E_{\text{distance}}(k)$ is defined as the energy to make sure that in every iteration, the iterative area is similar to a circle, or at least, it's smooth enough. We define $c_{\text{gv}}(k)$ as the edge energy, which can be calculated by GVF [17]. Then we need to solve such energy functional optimization problem:

$$E_{\hat{R}_c} = min\varphi \int_{R_c} E_{gray}(k)ds + \theta \int_{R_c} E_{distance}(k)ds + \lambda \int_{\Delta R_c} c_{gv}(k)ds \quad (1)$$

here φ, θ and λ are the weighs of E_{gray}, E_{distance} and c_{gv}. R_{c} is the iterative area and ΔR_{c} is its boundary.

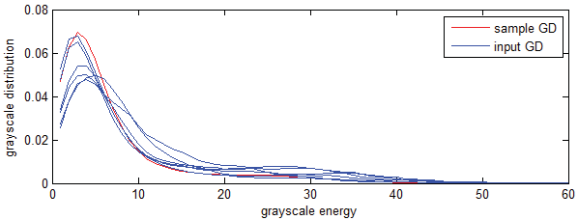


Fig.1. Grayscale distribution of chamber area in a series of input CMRI. The red line is the distribution of the sample images of the base and the blue lines are other flames of CMRI. It is clear that almost the chamber areas of all flames have the similar grayscale distribution.

Let $P(R,Z)$ be the grayscale distribution (Figure 1 shows the CMRI sequence input from base to apex, counts for 7 images), $I(x)$ be the grayscale feature of the input images, and $u(x) \in (0,1)$: when $u(x)=1$, the iterative area belongs to chamber area (R_{c}), and when $u(x)=0$, the area is out of chamber (R_{d}). $P(R,Z)$ is defined as follows:

$$P(R,Z) = \frac{\sum_{\text{R}} K_{\text{z}}(I(x))u(x)}{\sum_{R_{\text{c}}} u(x)} \quad (2)$$

$K_{\text{z}}(x)$ is given by Gaussian Core function, where $n \in \{1,2,3,4\}$, and when n=4, $K_{\text{z}}(x)$ is typical Gaussian Core function, defined as follows:

$$K_{\text{z}}(x) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{4}}} \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \quad (3)$$

$R(P,M)$ is defined as the similarity between distribution P of $T_{\text{in}}(x)$. And distribution M of $T_{\text{sam}}(x)$ calculated by Kullback–Leibler divergence is defined as follows:

$$R(P,M) = \sum_{z=0}^{N} \ln(\frac{P(z)}{M(R,z)})P(z) \quad (4)$$

But only $E_{\text{gray}}(k)$ is unable to obtain the optimal edge of endocardium, so we add another constraint: $E_{\text{distance}}(k)$, which is the energy of the distance from the points p of R_{c} to $O(x)$, the center point of $T_{\text{sam}}(x)$, where we can get $O(x)$ defined as follows:

$$O(x) = \frac{\sum_{R_{\text{c}}} x}{\sum_{R_{\text{c}}} u(x)}, R_{\text{c}} \in T_{\text{sam}}(x) \quad (5)$$

So $E_{\text{distance}}(k)$ can be obtained as follows:

[1] CMRI data source: Sunnybrook Cardiac MR Database Contours, gathered by GE Medical Systems

$$E_{\text{distance}}(k) = \sum_{R_c} \sqrt{(\|p - O\|^2 - \bar{R})^2} \qquad (6)$$

with $\bar{R}$ as the mean radius of R_c.

We calculate $c_{gv}(k)$ by GVF. Because the energy calculated by $\Delta I(x)$ can be only obtained on the edge, which makes $E_{\hat{R}_c}$ only be constrained by E_{gray} and E_{distance} when the iterative area excludes the edges or intensively constrained by the edge when the iterative area includes the edges. But GVF force field will lead the iterative area to the edges, which can converge to the optimal area. ε_{gv} and c_{gv} are defined as follows:

$$\varepsilon_{gv} = \sqrt{G(b,v)^2}, c_{gv} = \frac{1}{1 + \delta\varepsilon_{gv}} \qquad (7)$$

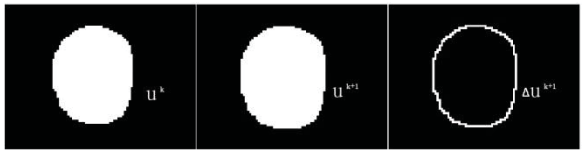
with $G(b,v)$ defined as GVF force field [17]:

$$\varepsilon_{gv} = \iint \mu(\nabla^2 G) + |\nabla I|^2 (G - \nabla I)\mathrm{d}x\mathrm{d}y \qquad (8)$$

To make $\varepsilon_{gv} \to 0$, we get b and v as follows, where k is the iteration times.

$$\begin{cases} b^{k+1} = b^k + \mu\nabla^2 b^k - \left(g_x{}^2 + g_y{}^2\right)\left(b^k - g_x\right) \\ v^{k+1} = v^k + \mu\nabla^2 v^k - \left(g_x{}^2 + g_y{}^2\right)\left(v^k - g_y\right) \end{cases} \qquad (9)$$

Now $E_{\text{gray}}(k)$, $E_{\text{distance}}(k)$ and $c_{gv}(k)$ are available to achieve, so when $E_{\hat{R}_c} \to \min$, $u(x)$ will converge to the optimal chamber area. Our algorithm of calculating area $u(x)$ is something like area growing algorithm. We define an initial area $u(x)^0$ with the center of $O(x)$ (e.g., a circle with radius of 5 px) and a dilation template B (e.g. 3x3 or 5x5). We perform B to $u(x)^k$ to achieve $u(x)^{k+1}$ and define $\Delta u(x)^{k+1} = u(x)^{k+1} - u(x)^k$ as shown in Figure 2.



(a) (b) (c)

Fig.2. $u(x)$ is calculated by using dilation template. (a) the area of $u(x)^k$, (b) the area of $u(x)^{k+1}$, (c) the area of $\Delta u(x)^{k+1}$.

Let Δu be constrained by energy function $E_{\hat{R}_c}$, so we convex $u \in \{0,1\}$ to $u(x) \in [0,1]$. Define $p(x) \in I(x) \cdot \Delta u,\ x \in R$.

For any $p(x)$, define $E_{\text{gray}}(p) = R\left(P_{\text{gray}}, M_{\text{gray}}\right)$, where P is the distribution of $p(x)$ and M is the distribution of sample $T_{\text{sam}}(x)$. The edge function of $p(x)$ is defined as following, where $C_{u^{k+1}}(x)$ is the edge of $u(x)^{k+1}$ calculated by Canny edge detector mentioned [18]:

$$E_{\text{edge}}(p) = \sum_{x \in R} \frac{C_{u^{k+1}}(x)}{1 + \delta\varepsilon_{gv}(x)} \qquad (10)$$

So $\Delta u(x)^{k+1}$ can be updated as following:

$$\Delta u(x)^{k+1^*} = \varphi R\left(P_{\text{gray}}, M_{\text{gray}}\right)^* + \theta E_{\text{distance}}(p)^* + \lambda E_{\text{edge}}(p)^* \qquad (11)$$

R^* being normalized by R, $E_{\text{edge}}{}^*$ being normalized by E_{edge}, $E_{\text{distance}}{}^*$ being normalized by E_{distance}, and $\varphi + \theta + \lambda = 1$. So $\Delta u(x)$ can be constraint to [0,1].

To make $u(x) \in [0,1] \to u(x) \in \{0,1\}$, we defined a constant ω to binary-value area $\Delta u(x)^{k+1^*}$, as follows:

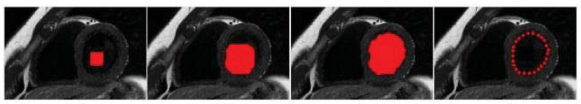
$$\Delta u(x)^{k+1^*} = \begin{cases} 1, & \text{when } \Delta u(x)^{k+1^*} \le \omega \\ 0, & \text{when } \Delta u(x)^{k+1^*} > \omega \end{cases} \qquad (12)$$

and redefine $u(x)^{k+1^*} = u(x)^{k^*} + \Delta u(x)^{k+1^*}$.

The global energy E^{k+1} is calculated as follows:

$$E^{k+1} = \varphi R\left(P^{k+1}{}_{\text{gray}}, M_{\text{gray}}\right)^* + \theta E_{\text{distance}}(p)^* + \lambda \sum_{x \in R} \frac{C_{u^{k+1}}(x)^*}{1 + \delta\varepsilon_{gv}(x)} \qquad (13)$$

Because R decreases monotonically with distribution P being similar with distribution M, E_{distance} also decreases monotonically with the area R_c being smooth, and edge energy decreases monotonically with edge of R_c approach the edge of input CMRI, E^{k+1} will obtain its minimum value when area R_c approaches the optimal endocardium area. So when $E^{k+1} - E^k < \vartheta$, with ϑ is defined as an infinitesimal constant, the algorithm is terminated, and the area $u(x)$ converges to the target chamber area. Figure 3 shows the progress of endocardium contour extraction by improved DM method. And figure 4 shows the distribution matching between the sample chamber area and automatic extracted chamber area.



(a) (b) (c) (d)

Fig.3. Endocardium contour segmentation by improved DM method. (a) initial area, (b) result area after 15 times iteration, (c) endocardium contour extraction result, (d) final endocardium contour.

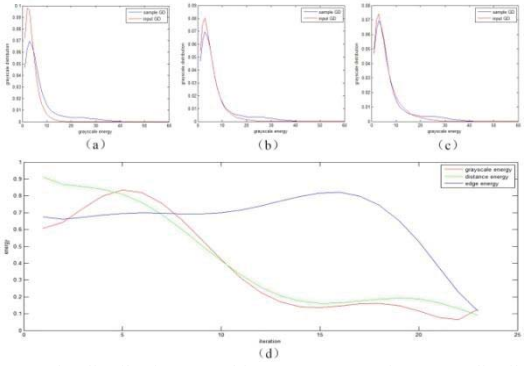


Fig.4. Grayscale distribution matching progress and energy distribution in each iteration.

In Figure 4 (a), (b) and (c), red line is the sample distribution, and blue line is the automatic extraction distribution. From Figure 4 (a) to (c), automatic extraction distribution becomes more and more similar to the sample distribution, which corresponds to progress of Figure 3 (a), (b) and (c) and (d) shows how the energy of grayscale, distance and edge change in each iteration, where red line is the grayscale energy, green line is the distance energy and blue line is the edge energy. Algorithm is terminated after 23 times iteration, when the global energy obtains its minimum. And when we obtain the final endocardium contour, the similarity of grayscale and distance distribution between the sample and automatic extraction result achieves its maximum

3.2 Improved ACM for epicardium contour extraction

The improved ACM epicardium contour extraction includes two parts: 1) obtain initial iterative curve and edge distribution, and 2) epicardium contour extraction by improved ACM with adaptive changing parameters and a local circular constraint of external force field.

Firstly, the input images are denoised by using Gaussian Core function. Then the improved Canny edge detecting algorithm[18] is used to obtain the optimal edge.

Let $f(x), x \in R^{N\times N}$ be the input image, g(x) be Gaussian denoising function, and $j(x)$ be Canny edge detector. We can obtain the edge of the images input like following:

$$f_{\mathrm{d}}(x) = j(x) * (\mathrm{g}(x) * f(x)) \qquad (14)$$

Let $f_{\mathrm{end}}(x),\ x \in R^{N\times N}$ be the function of endocardium profile received from section 3.1, we use it with linear sampling to obtain the initial iterative curve of ACM and convert it to be $w(x) \in \{0,1\}$. Define B as 3×3 dilation template and $\oplus$ is dilation operator.

$$w(x) = w(x) \oplus B \qquad (15)$$

Now we can get $D_{\mathrm{edge\ -final}}(x, y)$ filtered by $w(x)$ defined as follows, and the sample results have been presented in Figure 5.

$$D_{\mathrm{edge\ -final}}(x) = f(x) \cdot [1 - w(x)] \qquad (16)$$

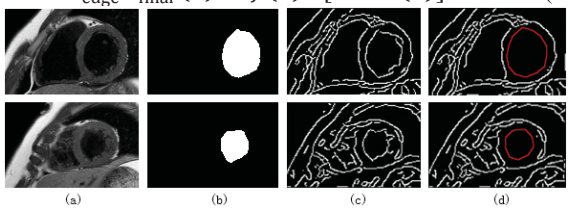


Fig.5. Results of middle and base of the heart in two rows: (a) the original images, (b) the chamber images received from section 3.1, (c) the edge images from Gaussian denosing and Canny edge detector, (d) the final edge images for ACM filtered by chamber area, where the red curve is the initial iterative curve.

ACM is a typical application of energy functional optimization problem, which finds minimum of its external and internal energy, as follows, where $v(s)$ is the active contour curve:

$$v(\mathrm{s}) \coloneqq \mathrm{argMin} \int \mathrm{t}E_{\mathrm{int}}(\mathrm{v(s)}) + \tau E_{\mathrm{ext}}(\mathrm{v(s)})\mathrm{ds} \qquad (17)$$

We can obtain the external energy E_{ext} by $-|\Delta f(x)|^2$ or by GVF, and get internal energy E_{int} as following equation, where E_{elastic} is elastic energy, E_{bend} is bend energy and E_{circle} is the proposed local circular constraint energy.

$$wE_{\mathrm{int}} = \alpha E_{\mathrm{elastic}} + \beta E_{\mathrm{bend}} + \gamma E_{\mathrm{circle}} \qquad (18)$$

with E_{elastic} is given by:

$$E_{\mathrm{elastic}} = \left|\frac{\mathrm{d}\bar{\mathrm{s}}}{\mathrm{d}\bar{\mathrm{x}}} - \frac{\mathrm{ds}}{\mathrm{dx}}\right| \qquad (19)$$

where

$$\frac{\mathrm{ds}}{\mathrm{dx}} = \sqrt{1 + \mathrm{y}'^2}, \qquad \frac{\mathrm{d}\bar{\mathrm{s}}}{\mathrm{d}\bar{\mathrm{x}}} = \frac{1}{\mathrm{s}} \int \sqrt{1 + \mathrm{y}'^2}\,\mathrm{ds} \qquad (20)$$

and E_{bend} can be easily calculated by distances between current point and its neighbor points.

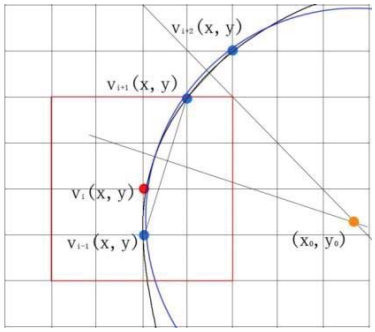


Fig.6. v_i is the current point, $v_{i-1}, v_{i+1}, v_{i+2}$ are its neighboring points, $C_0(x_0, y_0)$ is the center of the local circle calculated by crossover point of the perpendicular bisector of $[v_{i-1}, v_{i+1}]$ and $[v_{i+1}, v_{i+2}]$. Define point v_i' as the next iterative point of point v_i. We should make sure that v_i' is not away from the circle with center point $C_0(x_0, y_0)$ to lower the energy E_{circle}.

Because the shape of left ventricle is something like an oval in base and middle but like a circle in apex, and to make sure the epicardium contour smooth enough, we can use a local circular constraint additionally, as shown in Figure 6. Local circular constraint guarantees that the active contour curve can keep smooth locally, and easily converge to the optimal edge globally.

Therefore, we can get E_{circle} as equation below, where R_0 is the radius of the local circle:

$$E_{\mathrm{circle}} = \exp|R_{\mathrm{i}} - R_0| \qquad (21)$$

In our experiments, we find that ACM's iteration is influenced significantly by parameter τ of the external force filed. If τ is a constant, it will cause a multi-boundary convergence problem, which means that when different edges become close enough, the curve will be only controlled by E_{int}, and E_{ext} will cut no ice, as shown in Figure 7.

To overcome this problem, an adaptive changing parameter method is developed. Let ω_{τ} be the adaptive changing factor, x_f and E_f be the position and external energy

where the curve the first time meets a continuous boundary, so we constrain $\tau_i^{k+1} = \tau_i^k \cdot \omega_{\tau_i^k}$ by E_{ext} and the distance from x_f, defined as follows:

$$\omega_\tau = \frac{\theta \log \frac{E_{ext}}{E_f}}{\log(\|x - x_f\|^2 + 1)} \quad (22)$$

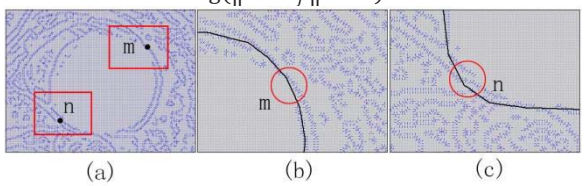


Fig.7. Multi-boundary convergence process. (a) the edge of CMRI. (b) the boundary is very simple at the point m, where the curve can converge to the optimal edge easily. (c) the boundary is complex at the point n. If the parameter is not influenced by different edges, the curve's changing will be only influenced by the internal force, as a result of which, it will expand to our undesirable edges.

Finally, by Newton iterative algorithm, which considers that the local optimal solution tends to the global optimal solution, we can obtain the epicardium contour $v(s)$ by calculating the energy of its 24 neighborhood, and consider that the point with the minimum E^i defined as following can make $v(s)$ converge to target optimal curve. After repeatedly iterating, when $v(s)$ almost doesn't change, our algorithm terminates.

$$E^i = \alpha E^*_{elastic} + \beta E^*_{circle} + \tau E^*_{ext} \quad (23)$$

Figure 8 presents the curve convergence procedure controlled by the improved ACM algorithm.

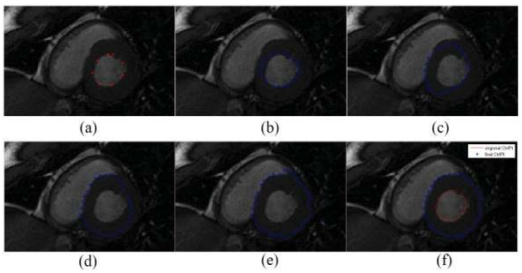


Fig.8. Epicardium contour convergence process. From (a) to (f): the snapshots of left ventricle epicardium contour extraction procedure by the improved ACM algorithm. (f) the result after 53 times iteration, where the red line is the initial iterative curve.

4 Segmentation Experiment

The proposed DM and ACM LV contour extraction methods are evaluated over a dataset which contains about 400 short-axis plane CMRI of over 20 subjects, and compared to other methods.

Figure 9 shows the chamber extraction results of the manual drawing, the proposed method, and methods proposed by Lee et al. [10] and Nambakhsh [14]. Methods mentioned by Lee et al. [10] and Nambakhsh [14] can be used to extracted endocardium contour, but significantly influenced by mastoid muscle and the fuzzy endocardium boundary of the apex. The contour extracted by the proposed method is more distinct, smoother and similar to the manual results.

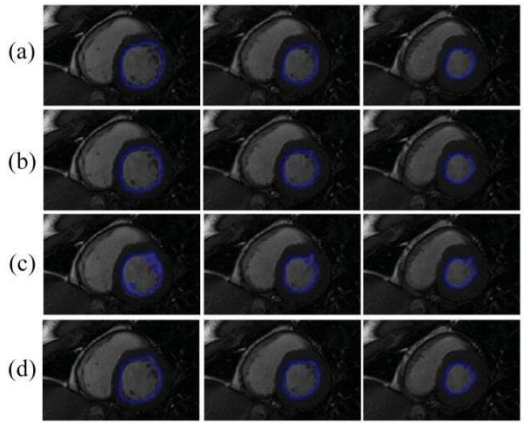


Fig.9. Endocardium contour extraction results comparison. From left to right are different planes of the heart. Blue line is the endocardium contour. (a) the manual contour, (b) the proposed method, (c) Lee et al. [10]' method, (d) Nambakhsh [14]' s method.

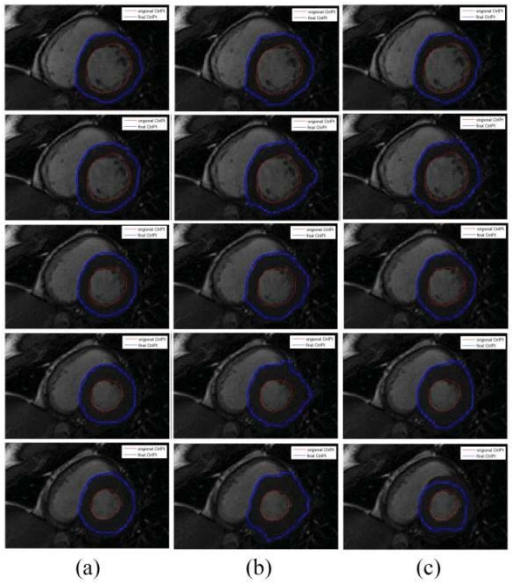


Fig.10. Epicardium contour extraction results comparison. From top to bottom is the end-diastole to end-systole of the heart. (a) the proposed method, (b) Nambakhsh [14]' s method, (c) Lee et al. [10]' s method. The blue line is the extracted contour of epicardium and the red line is the initial iterative curve calculated by endocardium contour received from section 3.1.

Figure 10 shows the epicardium extraction results of the proposed algorithm, Lee et al. [10] and Nambakhsh [14]'methods, with flames from end-diastole to end-systole, where we use the parameter as: $\alpha = 0.5, \beta = 0.01, \tau^0 = 0.2$, the threshold of Canny detector is 0.018, and σ of Guassian Denoising function is 31. Nambakhsh [14]'s method fails to extract a smooth enough contour of epicardium when handling the frames of end-systole. Lee et al. [10]'s method needs complex user input and sometimes can't obtain the accurate contour if the edge of epicardium is fuzzy or influenced by uneven grayscale distribution. Our approaches do not affected by those factors, and obtains the contour with higher accuracy and high robustness.

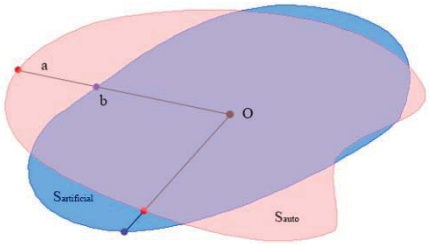


Fig.11. The way to calculate the differences of automatic segmentation method (s_a) with the manual contour (s_m). Point b is an arbitrary point in the contour of manual drawing. Point a is the nearest point nearby the line connecting b and the center point $O(x)$ of chamber.

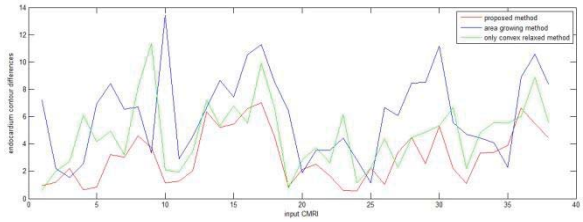


Fig.12. Y-axis: the differences between the extracted epicardium contours with the manual contour. X-axis : the input images. Red line: the proposed method; blue line: Region Growing from Lee et al. [10]' s study; green line: Only Convex Relaxed DM from Nambakhsh [14]' s study. The proposed method performs the lowest differences compared with other algorithms.

We use the following approach to calculate the differences of automatic segmentation method (s_{auto}) with the manual contour ($s_{artificial}$), where $s_{auto}(x)$ is the nearest point nearby the line connecting with $s_{artificial}$ (x) and $O(x)$, as shown in Figure 11.

$$h(s_a, s_m) = \frac{1}{N}\sqrt{\left(\sum_{x \in N}(s_m(x) - s_a(x))^2\right)} \qquad (24)$$

The comparison result in Figure 12 shows that the result of the proposed method is the most similar to the manual result where we perform the automatic LV segmentation methods in data of 14 subjects.

After epicardium and endocardium contours segmentation, we can use Simpson method [19] to calculate the cardiac volume to obtain LVEF, defined like following:

$$V = \sum_{i=1}^{N} h_i \cdot S_i \qquad (25)$$

with being the thickness of the frames and being the area of specific frame. And we deploy the results in HF' s early detection with semantically labeling related patient ' s information with HF ontologies. The formula below descripts how to calculate the LVEF, where is the volume of end-diastole, and is the volume of systole:

$$\mathrm{LVEF} = \frac{\mathrm{LVEDV} - \mathrm{LVESV}}{\mathrm{LVEDV}} \times 100\% \qquad (26)$$

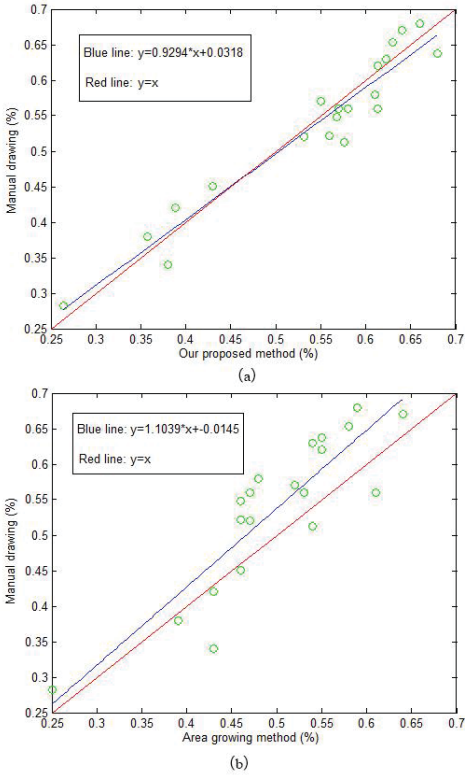


Fig.13. Comparison of LVEF calculated by different CMRI extraction method. (a) the proposed method, (b) area growing method in Lee et al. [10]' s study. The green points are distribution calculated by results received from automatic segmentation method and manual drawing results. The similarity with red line and blue line is the same with the similarity with the results of manual contour and auto-segmentation results. The comparison between (a) and (b) can be told that the LVEF calculated by proposed method performs better.

Figure 13 shows the LVEF results comparison estimation by linear fitting of the automatic segmentation calculation and the manual drawing results by performing in over 20 subjects (each subject uses about 4*10 frames to calculates the LVEF). The linear fitting equation of manual drawing results and the proposed method is y=0.9294x+0.0318, and the linear fitting equation of manual drawing results and Lee et al. [10]'s method is y=1.1039x-0.0145. After comparing with equation y=x, the proposed method is high performance.

5 Conclusions

Based on the computer technology, this study aims to seek an endocardium and epicardium contour extraction method for LVEF calculation from the CRMI, with simpler user interaction, more robust performance, and higher accuracy. More than 20 subjects have been used to demonstrate the method's performance with comparing to the research results[10] [14] to calculate the endocardium and epicardium contour. Meanwhile, when extracting the epicardium by the improved ACM algorithm, we calculated the differences between the results of manual drawing and the automatic segmentation methods, and a superior performance is revealed by the proposed method. The experiment results show that our method is more efficiently.

Acknowledgement

This research is supported by the National Key Basic Research and Development Programof China (973)(No. 2013CB329505), the National Natural Science Foundationof China (61320106008, 61370160). Zhong Wang is the corresponding author.

References

[1] Paulus W. J., T.C., Sanderson J. E., How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography. Associations of the European, 2007. 28: p. 2539–2550.

[2] N., D.R., The survival of patients with heart failure with preserved or reduced left ventricular ejection fraction: an individual patient data meta-analysis. European Heart Journal, 2012. 33(14): p. 1750-1757.

[3] Lynch M., G.O., Whelan P., Automatic segmentation of the left ventricle cavity and myocardium in MRI data. Comput. Biol. Med., 2006. 36(4): p. 389-407.

[4] Kang D., S.S.Y., Sung C. O., Pack J. K., Kim J. Y., Choi H. D., An improved method of breast MRI segmentation with Simplified K-means clustered images. Proceedings of ACM Symposium on Research in Applied Computation, 2011: p. 226-231.

[5] Chittajallu D. R., P.N., Kakadiaris I. A., An Explicit Shape-Constrained MRF-Based Contour Evolution Method for2-D Medical Image Segmentation. IEEE Journal of Biomedical and Health Informatics, 2014. 18(1): p. 120-129.

[6] Grosgeorge D., P.C., Dacher J. N., Ruan S., Graph cut segmentation with a statistical shape model in cardiac MRI. Computer Vision and Image Understanding, 2013: p. 1027-1035.

[7] Mahapatra D., S.Y., Joint Registration and Segmentation of Dynamic Cardiac Perfusion Images Using MRFs. Medical Image Computing and Computer-Assisted Intervention MICCAI 2010 Lecture Notes in Computer Science, 2010. 6361: p. 493-501.

[8] Wang X., H.Q., Heng P., Left Ventricle Segmentation with Mixture of Gaussian Active Contours. IEEE International Symposium on Biomedical Imaging, 2012: p. 230-233.

[9] Klann E., R.R., Ring W., a Mumford-Shah Level-Set Approach for the Inversion and Segmentation of SPECT/CT Data. American Institute of Mathematical Sciences, 2011. 5(1): p. 137-166.

[10] Lee H. Y., C.N.C.F., Cham M. D., Weinsaft J. W., and Wang Y., Automatic Left Ventricle Segmentation Using Iterative Thresholding and an Active Contour Model With Adaptation on Short-Axis Cardiac MRI. IEEE Transactions Biomedical Engineering, 2010. 57(4): p. 905-913.

[11] Petitjean C., D.J.N., A review of segmentation methods in short axis cardiac MR images. Medical Image Analysis, 2011. 15: p. 169-184.

[12] Ecabert O., P.J., Walker M. J., Ivanc T., Lorenz C., Berg J., Lessick J., Vembar M., Weese J., Segmentation of the heart and great vessels in CT images using a model-based adaptation framework. Medical Image Analysis, 2011. 15: p. 863-876.

[13] Njeh I., A.I.B., Hamida A. B., A distribution-matching approach to MRI brain tumor segmentation. 2012 9th IEEE International Symposium on Biomedical Imaging 2012: p. 1707-1710.

[14] S., N.C.M., Left ventricle segmentation in MRI via convex relaxed distribution matching. Medical Image Analysis, 2013. 17: p. 1010-1024.

[15] Ayed I. B., C.H., Punithakumar K., Ross I., Li S., Max flow segmentation of the left ventricle by recovering subject specific distributions via a bound of the Bhattacharyya measure. Medical Image Analysis, 2012. 16: p. 87-100.

[16]Pham V., T.K., and Naemura T., Foreground Background Segmentation using Iterated Distribution Matching. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011: p. 2113-2120.

[17] Xu C., P.J., Snakes, shapes, and gradient vector flow. IEEE Trans. Image Process, 1998. 7(3): p. 359-366.

[18] Wang B., F.S., An improved CANNY edge detection algorithm. 2009 Second International Workshop on Computer Science and Engineering, 2009: p. 497-500.

[19]Jenkins C., B.K., Chan J., Hanekom L., Marwick T. H., Comparison of Two- and Three-Dimensional Echocardiography With Sequential Magnetic Resonance Imaging for Evaluating Left Ventricular Volume and Ejection Fraction Over Time in Patients With Healed Myocardial Infarction. The American Journal of Cardiology, 2007. 99(3): p. 300-306.

VHR image segmentation using MeanShift clustering and spatial neighborhood

J. Lopez[1] **and J.W. Branch**[2]
[1]Department of Engineering, Universidad del Tolima, Ibague, Tolima, Colombia
[2]Department of Computing and Decisions Sciences, Universidad Nacional de Colombia, Medellin, Antioquia, Colombia

Abstract – *The segmentation of very high resolution imagery (VHR) represents a challenge to existing algorithms, due to high intraclass variability that occurs with pixels inside objects of interest in the image, and the low separability between classes. The use of unsupervised algorithms, such as MeanShift represent a significant advance in that direction, since it allows groups to have arbitrary shapes; but using distances to label assignment, that removes the neighborhood contribution, which provides valuable information in complex environments such as frontier zones between two objects. Therefore, the incorporation of spatial neighborhood after clustering is proposed as a later phase to segmentation, where the dubious pixels, are compared to its near vicinity, and then labeled according to its spatial vicinity.*

Keywords: image segmentation, VHR imagery, clustering, Mean-Shift algorithm, spatial neighborhood

1 Introduction

The image segmentation of very high resolution (VHR) imagery is a challenge for traditional segmentation techniques, because they do not make use of rich amount of information they offer [1], and in particular the spatial neighborhood of pixels. The image segmentation of images using clustering is a technique useful recognition patterns when there is no a priori knowledge of the categories of interest. Although this method is considered as a statistical technique, they do not require common assumptions of statistical methods, but often assume that the distribution of classes presents a particular form, but this is not always true.

The clustering process is able to organize data under a underlying abstract structure, either in groups of individuals or groups hierarchies based on their similarity [2]. These groups may differ in shape, size and density; usually measured using similarity or distance metric, such as the Euclidean distance.

Among the best known clustering algorithms is the k-means, which recently met his half century of existence, and the MeanShift algorithm, more recent, with some advantages over the first, described later.

An important aspect of clustering methods with relation to segmentation, is that these not give importance to the sequence patterns or vicinity of pixels. Therefore, it is necessary to consider these attributes to improve the accuracy of clustering [3].

1.1 Mean-Shift algorithm (MS)

This clustering algorithm its a nonparametric technique whose main advantages are that not require to know a priori the number of groups and allows these have arbitrary shapes. It is based on pattern recognition technique called Mean-Shift [4,5], which considers the attribute space as an empirical probability density function. Consists of finding those stationary points of the density function (modes) from the initial points present in the vector of attributes, retaining only the local maxima, and removing the remaining points. The set of all locations that converge to a mode, known as base of attraction, so those points that fall in a given base attraction, are associated with their respective grouping. For a more detailed description, read the original paper [6].

Among of the MS advantages is that it does not assumes spherical clusters, delivering a variable number of modes, only requires a parameter corresponding to the window size to use, and is robust to noise and outliers; but has some disadvantages, such that it is computationally expensive (O Tn^2), where T is the number of iterations, n is the number of points, and also its result depends on the parameter initially specified.

In short, the algorithm works as follows: a) Sets a window around each point in attributes space, b) Calculate the average of data inside window, c) Mode the window to the media, and repeat the process until it converges.

1.2 Spatial neighborhood (SN)

The spatial neighborhood of a given pixel in a digital image is defined by an n-connected system (or structural element) that identifies the type of connectivity, either 4-, 6-, 8-connected to each unlabeled pixel defining their respective label from this neighborhood [7]. This type of segmentation technique can be interpreted as a region growing process, which identifies the object label, even in the presence of noise [8].

The figure 1, shows the 4 connectivity type used for each pixel (p), where every neighbor has its respective label (l#), or alternatively the value -1, when the label is unknown.

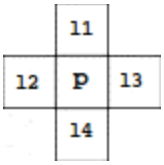


Figure 1. 4-c connectivity used to determine the pixel label of pixel 'p' under analysis.

The general operation for the spatial neighborhood (SN) can be summarized as follows: a) Establish a cross-shaped mask in each pixel of the image, b) Determine whether the pixel is unlabeled, c) Identify the label mostly displayed in that vicinity, d) Assign this label to the pixel under consideration, e) Continue to the next pixel.

2 Proposed algorithm

This algorithm seeks to complement with spatial neighborhood the benefits of clustering to improve their performance in very high resolution images (VHR), which have high intraclass variability that may be underestimated by the method of grouping.

2.1 Outline

For a gray-scale image (8-bit), denoting the pixel value as p(x,y), where 1 <= x <= N, and 1 <= y <= M of size MxN. Each pixel has a 4-connected neighborhood, vertical axis and horizontal axis, for a total of four neighboring pixels, where each neighbor has a particular label, or otherwise a value of -1, indicating no known label (Figure 1).

Initially, the MS method is applied onto the original image, with a chosen value of bandwidth equal to 0.2, although usually the default value is 0.3, but it can be set into any value in the range (0.0-1.0). This particular bandwidth value, allowed to obtain a reasonable number of labels (7-10). This method provides a set of centroids and their respective label. Subsequently, a segmentation of the original image is done using the centroids +/- 2 * (Std. Standard / number of labels). It is understood that pixel connectivity at the edge of the image is lower, that is, at the corners it will be equal to 2, while on the sides of the image is equal to 3.

The VHR image is processed from top left corner to the bottom right corner. Where p(x,y) is the current pixel under analysis, and its four-connected connectivity is represented by S = {L1, L2, L3, L4}, where each l# represents one of the possible labels provided by the MS method, otherwise the value -1, unlabeled.

For the label assignment of pixel p(x,y), there may occur one of the following cases: a) All 4-connected pixels have a value of -1 (eg S = {- 1, -1, -1, -1}), then retains the same value, b) If one of the pixels 4-connected has a value other than -1, that value will be assigned to the pixel under analysis is (eg S = {- 1, 2, -1, -1}), c) If a tie between two labels are present, or if all labels are different, its label value is chosen randomly (eg S = {-. 1, 2, 2, -1 } or S = {3, 4, 1, 2}). Table 1 listed all possible cases.

Table 1. Operations for pixel labeling under each possible case for p(x,y)

Case	p(x-1,y)	p(x+1,y)	p(x,y-1)	p(x,y+1)	Operation	Description
1	-1	-1	-1	-1	None	Continue to next pixel
2	-1	-1	-1	*l1*	Choose(*l1*)	Assign the unique label
3	-1	-1	*l1*	*l2*	Random(*l1* or *l2*)	Assign one of both
4	*l1*	*l1*	*l2*	*l2*	Random(*l1* or *l2*)	Assign one of both
5	-1	*l1*	*l1*	*l2*	Max()	Assign the most repeated label
7	*l1*	*l1*	*l2*	*l3*	Max()	Assign the most repeated label
8	*l1*	*l2*	*l3*	*l4*	Random(any)	Assign any of them

2.2 Implementation

The complete procedure of proposed method (MS+SN) is represented by Algorithm 1 (Figure 2), which also its flowchart is illustrated in Figure 3.

Algorithm 1.

```
Input: img1 (MxN digital image)
Output: img2 (MxN segmented image)
V = Choose(Higher_Variability(CV_htal(img1),CV_vert(img1)))
Vs = Sorted(V)
SN = label_mask(V)
Centroids, Labels = MS(Vs, bandwidth=0.1)
Intervals = centroids +/- (std.dev/#labels)
For x,y
      if p(x,y) = verify(p(x,y),intervals)
            break
      else
            p(x,y) = -1
For x,y
      if p(x,y) == -1
            label = SN(p(x,y))
            p(x,y) = label
```

Figure 2. Segmentation using the clustering method MeanShift (MS) and spatial neighborhood (SN), called MS + SN

The flowchart of proposed method is illustrated in Figure 3, where the spectral band used (R channel) is highlighted, with the required parameters such as bandwidth, and number of samples (points) used for grouping. The CV (Coefficient of Variation) allowed to determine which of the two (horizontal or vertical) lines has higher variability across the image, was the input required to build the vector of neighborhoods, required to incorporate the spatial neighborhood after the clustering.

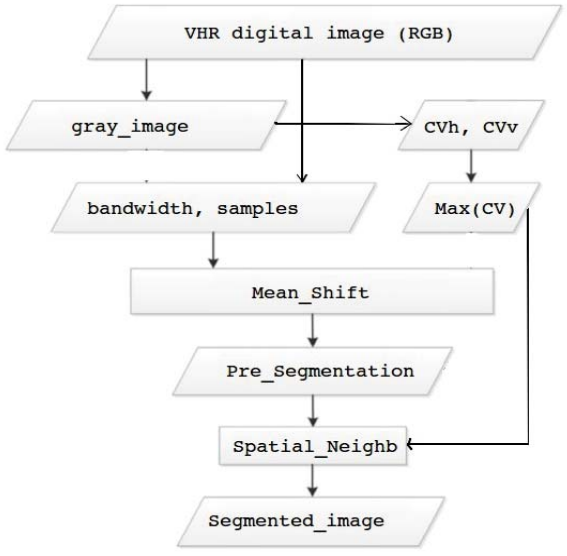


Figure 3. Proposed method flowchart

2.3 Results and discussion

The proposed method was applied on an VHR image, which results was under a qualitative and quantitative evaluation to measure their performance. Although there are many metrics for performance evaluation [9], for this case, the segmentation evaluation of particular interest objects (roofs, trees, cars, etc.) is performed, not assessed on the whole image; because usually when working with images from remote sensors, it is difficult to have an extensive segmented image of reference, which became not appropriate for images of high resolution, so it was of great relevance to use partial segmentation of reference objects, leaving the rest of pixels outside the evaluation, since its complete segmentation requires hard work and delayed time [10].

2.3.1. VHR digital image

The digital image used has a high quality natural color offering a spatial resolution of 30 cm with three spectral bands (RGB) with a size of 65 MB with dimension of 4723 x 7038 pixels, taken from the official website of the company Digital Globe. Although for testing the proposed method, only was used a fragment of it, with a size of 512 x 512 pixels (Figure 4), the R channel, corresponding to the spectral range 600-690 nm of EMS (Electromagnetic Spectrum).

The image covers a swath over city of San Diego (USA), corresponding to suburban upper and middle class with the presence of infrastructure (houses, roads, swimming pools, etc.), vegetation (trees, grasses, shrubs), and other (bare soil, roads, cars, etc.).

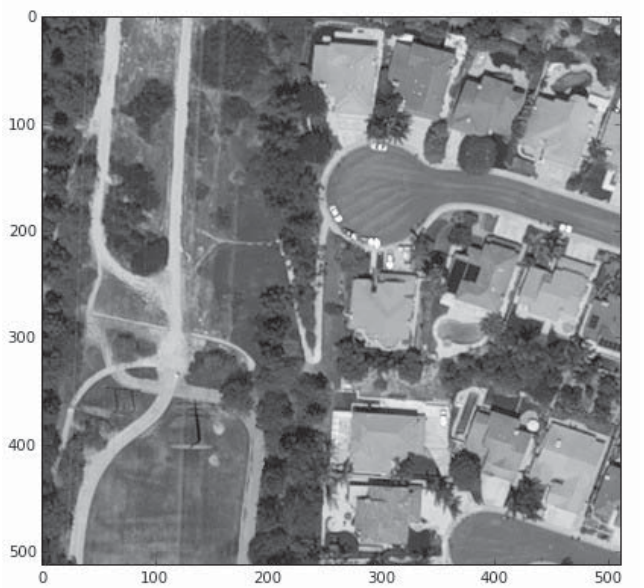


Figure 4. Original image of very high resolution (GSD = 0.3 m) corresponding to a suburb of San Diego, USA. (DigitalGlobe, 2012**)**

2.3.2. Qualitative evaluation

This type of evaluation is widely used to measure membership of an object class to a single region [11].

In Figure 5a, it can be seen at frames 6,8,9 and 10, as the proposed method (MS+SN) allows better discrimination or objects differentiation compared to MS method (Figure 5b), where some objects appear as one, such as access to a garage, a leafy tree, bare ground or pole holding power cables. This is significant visually, because it has been found that the proposed method is able to separate them from their environment, reaching more clear differentiation, while the MS method mixes with its surrounding environment. As for roofs (frames 4 and 5 from Figure 5a and 5b), was observed a similar performance between the two methods, both manage to separate the object from its neighbors, although its appreciated a slight superiority in terms of homogeneity under the method MS, against the proposed MS+SN method.

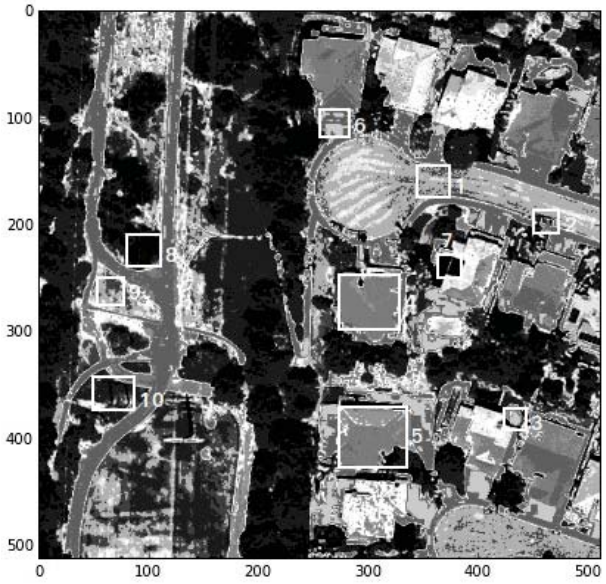


Figure 5a. Segmentation under proposed method (MS+SN)

Comparing the frames 1,2,3, from Figures 5a, we found a bad segmentation quality, because is observed greater blurring/dispersion, as in the case of object 'road' (frame 1), presenting a oversegmentation. For the object 'car' (frame 2), with white tone, the proposed method prevent its good visualization. Similar situation was presented with the water tanks, illustrated in frame 3. Meanwhile, in Figure 5b, the MS method performed well in the segmentation of these same objects, differentiating them from their immediate environment.

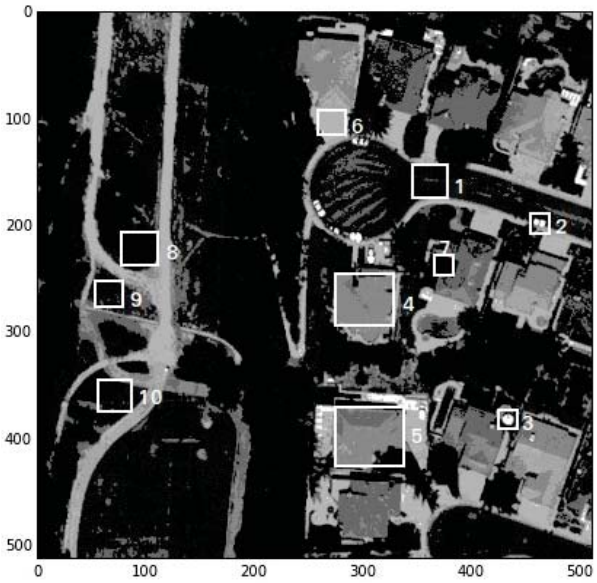


Figure 5b. Segmentation under MeanShift method (MS)

2.3.3. Quantitative evaluation

This type of evaluation measure the exact geometric position of the edges of each object or segment, using as quantitative measure the segmentation error, ie, calculating the number of pixels coincidences (AND operation) vs excess (XOR operation) between pairs of binary images (reference vs each method) for a given object (Eq. 1). This means that the error is between [0,1], approaching to zero, when the method delivers fewer excess and high coincidences between both binary images, and near to one when there are high excess and fewer coincidences; according to the following formula:

$$Error = 1 - (Pixel\ coincidences / Pixel\ excess) \quad (1)$$

Table 2 lists several objects analyzed showing their respective number of pixels after reference segmentation, and comparing the adjustment of each segmentation method against it.

Tabla 2. Error rates obtained by different objects under each segmentation method.

		MS			MS+SN		
Object	#Pixels	Coincidences	Excess	Error	Coincidences	Excess	Error
road	2340	1728	2196	0.21	720	12996	0.94
parking	79	13	68	0.81	62	62	0.00
tree	4428	720	6768	0.89	1404	9000	0.84
shallow	11	3	14	0.79	11	0	0.00
post	2628	2592	8352	0.69	1116	2772	0.60

These results shows contrasting agreement with previous qualitative evaluation, mainly depending on several object features, like size, geometry and intensity variability.

Figure 6 illustrates the three segmentation results for the object 'tree', where it can be seen that both methods has similar high error rate, , with values above 0.8 (Table 2). Figure 7 shows the results of AND and XOR operations between reference segmentation against the proposed and traditional method (PM and MS respectively), where it can be seen that the proposed method has more pixel coincidences, but at same time, it has too more pixel excess due to pixel intensity variability and geometry, causing its high error rate.

Figure 6. Segmentation of object 'tree' (left: Reference, center: PM method, right: MS method)

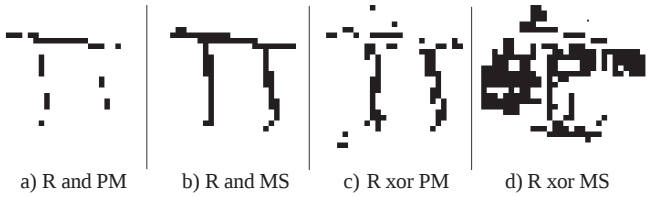
a) R and PM b) R and MS c) R xor PM d) R xor MS

Figure 7. AND and XOR operations for object 'tree' between reference segmentation (R), proposed method (PM), and traditional method (MS)

Figure 8 shows 'post' object segmentations results, the proposed method had labeled it properly (center), separating it from its environment (pastures), while the traditional method (MS), mixed two different objects, 'post' and 'grass', considering both as a single object (right).

Figure 8. Segmentation of object 'post' (left: Reference, center: PM method, right: MS method)

The coincidences and excess for object 'post' is illustrated in Figure 9, where the proposed method, behaves a little better than traditional method (0.6 vs 0.69), because although the last one has higher coincidences (9b), it suffer from having much excess pixels, as can be seen in Figure 9d.



a) R and PM b) R and MS c) R xor PM d) R xor MS

Figure 9. AND and XOR operations for object 'tree' between reference segmentation (R), proposed method (PM), and traditional method (MS)

For object 'road', the traditional method (MS) behaved better, because it fully identifies the object under analysis (Figure 10 right), while the proposed method (PM), had a really poor performance finding the correct object border (center). Figure 11, also illustrates that proposed method has few coincidences and many excess pixels, giving an incorrect segmentation result.

Figure 10. Segmentation of object 'road' (left: original, center: MS-SN, right: MS)

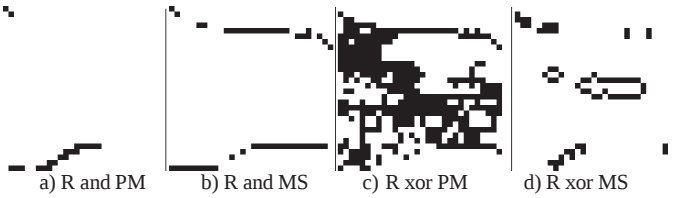


Figure 11. AND and XOR operations for object 'road' between reference segmentation (R), proposed method (PM), and traditional method (MS)

2.4 Analysis

According to obtained segmentation results, we can highlight the following facts: the addition of post-clustering spatial neighborhood analysis presents promising results, because it allowed differentiate better some objects, showing slight variations in its gray shades that clustering do not taken into account, but with the addition of neighborhood analysis, could improve in this area.

Another aspect that is important to mention, is related to the fact that was obtained a better performance with proposed method on objects that present some type of regular geometry and high intensity variability, while MS method, performs better on those objects with bigger areas, high contrast with neighbor objects and low intensity variability.

It is considered that spatial neighborhood used can be modified to improve their performance, maybe an increase in its size (6-c, 8-c, etc.), and also incorporate other elements such as texture, although these would increase its computational cost.

3 Conclusions

In this paper was proposed a method to combine a segmentation method based on MeanShift clustering, but with an added component, the spatial neighborhood. Although, this could turn image processing computationally intensive, because it is applied to each unlabeled pixel of image, its performance allowed detecting promising results, but this computational burden can be overcome to improve its operation and efficiency.

Spectral information (intensity gray levels) and spatial (neighborhood) for segmentation of objects of interest was utilized. Although the results show the benefits of this promising approximation, still is observed overlap between objects, or blurring of some of them, but a strength in the separation of small objects with high variability in their shades, characteristic identified in VHR images.

As future work, we consider increasing the size of the neighborhood, to improve the performance of this segmentation, and maybe incorporate additional information (shape, size, etc.), although this would increase the computational cost significantly.

4 References

[1] J. Inglada and J. Michel, "Qualitative Spatial Reasoning for High-Resolution Remote Sensing Image Analysis", Trans. on Geoscience and Remote Sensing (IEEE), Vol. 47, Iss. 2, February 2009.

[2] A.K. Jain, "Data Clustering: 50 Years Beyond K-Means", Pattern Recognition Letters, Vol. 31, Iss. 8, 651-666, 2010.

[3] Y. Xia, D. Feng, T. Wang, R. Zhao, Y. Zhang. "Image segmentation by clustering of spatial patterns", Pattern Recognition Letters. Vol. 28, Iss. 12, 1548–1555, September 2007.

[4] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", Transactions on Information Theory (IEEE), Vol. 21 , Iss. 1, 32-40, 1975.

[5] Y. Cheng, "Mean shift, mode seeking, and clustering", Transactions on Pattern Analysis and Machine Intelligence (IEEE), Vol.17, Iss. 8, 790–799, 1995.

[6] D. Comaniciu and P. Meer. “Mean shift: A robust approach toward feature space analysis”, Trans. Pattern Anal. Machine Intell. (IEEE), Vol. 24, Iss. 5, 603–619, May 2002.

[7] N. Makni, N. Betrouni, O. Colot. "Introducing spatial neighbourhood in Evidential C-Means for segmentation of multi-source images: Application to prostate multi-parametric MRI", Information Fusion, Vol. 19, 61–72, September 2014.

[8] Y. Zuo, C. Do, A. Neumann. "Automatic measurement of surface tension from noisy images using a component labeling method", Colloids and Surfaces A: Physicochemical and Engineering Aspects, Vol. 299, Iss. 1–3, 109–11615, May 2007.

[9] J. Michel, M. Grizonnet and O. Canevet, "Supervised re-segmentation for very high-resolution satellite images",

Geoscience and Remote Sensing Symposium (IGARSS), IEEE International, 68-71, 2012.

[10] J. Liu, P. Li, X. Wang. "A new segmentation method for very high resolution imagery using spectral and morphological information", Journal of Photogrammetry and Remote Sensing (ISPRS), Vol. 101, 145–162, 2015.

[11] J. Schiewe, "Segmentation of high-resolution remotely sensed data concepts, applications and problems", Geospatial Theory, Processing and Applications (ISPRS), 9–12, July 2002.

HISTO-BINARY COMBINED CORNER ENHANCEMENT (HBCCE) IMPROVED ALGORITHM FOR IMAGE PROCESSING CORNER DETECTION

Abdeslam El Harraj[1] **and Naoufal Raissouni**[2]
[1]and [2]: RSAID Laboratory: "Remote sensing/Signal-image Processing & Applied mathematics/Informatics/ Decision making". The National School for Applied Sciences of Tetuan. Univeristy of Abdelmalek Essaadi. BP. 2222. M'Hannech II. 93030. Tetuan. Morocco.

Abstract - *A novel algorithm, Histo-Binary Combined Corner Enhancement (HBCCE), for enhancing corner detection is discussed. The main goal is to enhance the repeatability for the corner by corner sharpening using a preprocessing filtering strategy. We are particularly interested in corner points because they are defined locally, usually in very small neighborhoods. The quality of the corners and the efficiency of the detection methods are both very important aspects that can greatly impact the accuracy, robustness and real-time performance of the corresponding corner-based vision system. There are many variant of corner detectors, but, no precise corner detector exists. The most used corner detectors our days are, Harris corner detector [1], Sh-Tomasi detector [2], SUSAN detector [3], CSS detector [4], FAST detector [5] and AGAST detector [6].*

In the present paper, we propose a new strategy to enhance the corner detectors by adding a preprocessing stack prior to any detection step, as mentioned previously the final goal is to sharpen the corners so they can be easily detected. Our approach is based on an efficient combination of filters. We verify the performance of the proposed method by measuring the repeatability rate under various JPEG compressions, rotation, scale, blur and illumination changes using a standard dataset [7].

Keywords: binary features, enhance corner detector, preprocessing, AGAST, ORB, BRISK

1 Introduction

Corners, also known as junctions, key points or interesting points, are used in many image applications such as image registration, shape analysis, object recognition/tracking, motion analysis, scene analysis, stereo matching, etc. Therefore, when working on two-dimensional features, a particular interest is given to corner detection [1] [2] [15] [16]. Because of the growing interest of using corner based features, many researches investigate on finding a nonlinear operator able to remove texture and noise, while preserving edges and corners [5] [6] [7] [8] [14]. The leading operators are based on median filtering [9], bilateral filtering [10], mean shift [11], total variation [12] and the most popular anisotropic diffusion [13]. The last one is not computationally efficient and has been subject of many improvements in last year's [14].
The purpose of these techniques is to process an image so that the final image gives more visual information than the original one, but none of these algorithms can give good results for all types of applications.

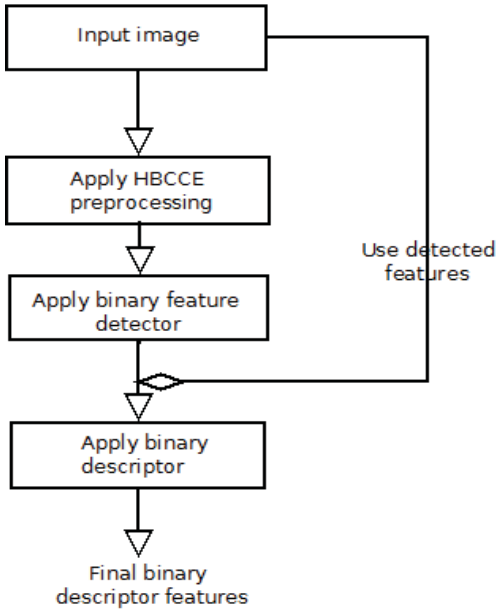


Fig. 1: How to integrate the HBCCE stack with any binary detector/ descriptor approach.

In this paper we propose a new preprocessing stack enhancing corner detection while removing noise, blur and minimizing the effect of illumination changes (Figure1). The purpose is to come out with a new model for building a more accurate local binary features detector based on a very efficient preprocessing stack. Our method proceeds as follows (Figure. 2):

First, the original image is converted to a 16-bit grayscale image.

Second, we apply a contrast-wise version of Contrast Limited Adaptive Histogram Equalization to reduce the

illumination effects.

Third, we use Unmask Sharpening, the purpose is to reduce the noise with a Gaussian blurring operator and then subtract the blurred version from the original one.

Fourth, we apply Laplacian Filter for shapes sharpening, in our case we work directly on the resulting image after Laplacian Filtering without subtracting the filtered image from the original one.

Finally, we convert the final enhanced image matrix to 8-bit grayscale, representing the final HBCCE enhanced image. The first aim of this paper is to set out a concise overview of the proposed HBCCE method and demonstrate its efficiency under numerous image changes. In the next section we describe the steps used in our approach.

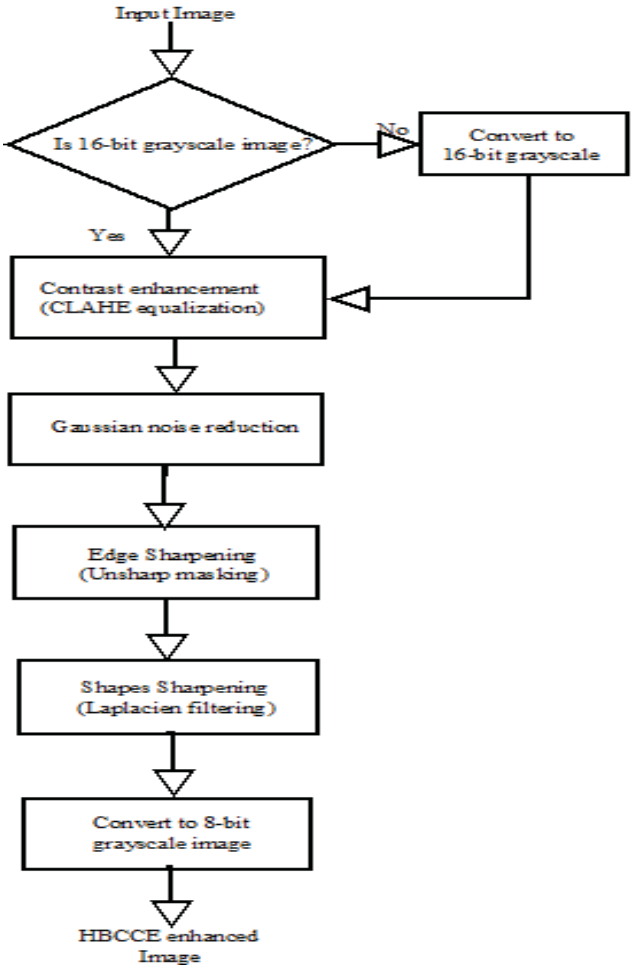


Figure 2: a – Diagram of different steps involved in HBCCE preprocessing algorithm.

2 CLAHE: Contrast-Limited Adaptive Histogram Equalization

The first step in our preprocessing stack is to improve the contrast of the overall objects presents in the processed image. For that raison, we will use CLAHE [18][32] as a powerful contrast enhancement.

Contrast enhancement methods are not intended to increase or supplement the intrinsic structural information in an image but rather to improve the image contrast and hypothetically to enhance particular characteristics. The input images are 8-bit grayscale images. We can process these images directly. But there is a slight problem with that. Black-to-White transition is taken as Positive slope (it has a positive value) while White-to-Black transition is taken as a Negative slope (It has negative value). So when we convert data, all negative slopes are made zero. And then we miss some edges. To bypass this problem, we convert data type to some higher forms like 16-bit, 64-bit etc (we use 16-bit grayscale images), process it and then convert back to original 8-bit

2.1 CLAHE concept:

Initially developed for medical images, CLAHE has demonstrated to be successful for enhancement of low-contrast images such as portal films [32]. CLAHE is an adaptive contrast enhancement method based on Adaptive Histogram Equalization (AHE) [17]. AHE proceed as follows: The histogram is calculated for the contextual region of a pixel. The resulting pixel's intensity is transformed to a value within the display range proportional to the pixel intensity's rank in the local intensity histogram but this process can over amplify the noise in the initial image.

Basically, developed to prevent the over amplification of noise that AHE can give rise to [18] [19], CLAHE refine AHE by imposing a user specified maximum, ie, Clip Limit, to, the height of the local histogram, and thus on the maximum contrast enhancement factor. The enhancement is thereby reduced in very uniform areas of the image "tiles". The resulting neighboring tiles are then stitched back seamlessly using bilinear interpolation, which prevent over enhancement of noise and reduce the edge-shadowing effect of unlimited AHE (Figure 3). Thus, CLAHE can limit the noise whereas enhancing the contrast [18] [32].

In our case we use a Uniform distribution with a clip limit equal to 0.1. (Figure 3) shows an example of the produced enhanced grayscale image by applying the CLAHE enhancement.

The clip limit can be obtained by: ß [19].

$$\beta = \frac{M}{N}\left(1 + \frac{\alpha}{100}(S_{max} - 1)\right) \quad (1)$$

Where α is clip limit factor, M region size, and N is grayscale value. The maximum clip limit is obtained for α=100.

The uniform CLAHE equalization is obtained by (2)

$$I = (I_{max} - I_{min}) * P(f) + I_{min} \quad (2)$$

Where:

I : computed pixel value

I_{max} : Maximum pixel value

I_{min} : Minimum pixel value

P(f) : Cumulative probability distribution

2.2 CLAHE Testing Results:

Testing CLAHE (Figure 3) on the image datasets proposed

by Mikolajczyk and Schmid [20] and available on[1], shows that the downsides of CLAHE equalization are:
Amplify image noise for flat regions.
Introduce ring artifacts for strong edges.
To reduce the noise introduced by the contrast enhancement, we apply Gaussian smoothing basically used for Gaussian noise reduction.

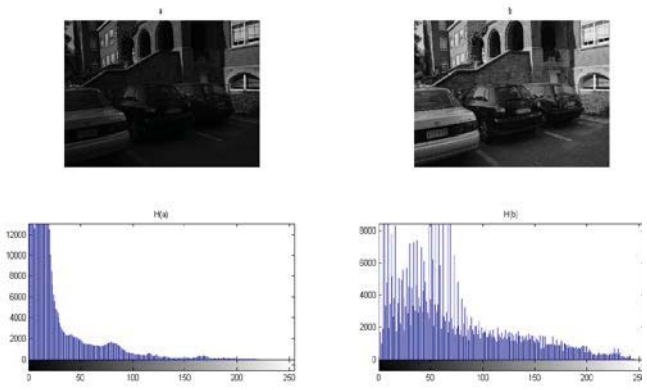


Fig. 3: a - original grayscale image (leveun-6), b - grayscale enhanced image with CLAHE (Distribution = unifrom , clipLimit=0.1,tileSize=(2,2)), H(a)- classical histogram image for original grayscale image and H(b)- CLAHE histogram for the transformed image.

3 Noise reduction: Gaussian blurring

Many attempts have been trying to construct digital filters which have the qualities of noise attenuation and detail preservation. One of the best known filters when dealing with impulsive noise is the median filter [23].
Median filter is less efficient in presence of Gaussian noise. Several researchers have attempted to generalize the standard median filter for removing Gaussian noise with more or less success. In our case we use Gaussian blurring for noise reduction.

3.1 Gaussian Blurring Concept:

Blurring filters generally follows the equation (3):

$$J(i,j) = \sum_{k,l} I(i+k, j+l) * G(k,l) \quad (3)$$

Where:
$J(i,j)$ is the blurred image
$I(i+k, j+l)$ is the input pixel values.
$G(k,l)$ is the kernel (the coefficient of the filter).

The Gaussian Blur effect is a filter that blends a specific number of pixels incrementally, following a bell-shaped curve. The blurring is dense in the center and feathers at the edge [28]. For Gaussian blurring, we replace in (3), $G(k,l)$ by the Gaussian kernel; which is given by (4):

$$G(k,l) = \frac{1}{2\pi\sigma^2} e^{-\frac{k^2+l^2}{2\sigma^2}} \quad (4)$$

To produce a discrete approximation of the Gaussian filter, we use the property that, the distribution approximate zero at about three standard deviations from the mean. 99% of the distribution falls within 3 standard deviations. Gaussian filter, is probably the most useful, simple and frequently used filter for noise reduction (but not the fastest). It's implemented by convolving each point in the input image with a Gaussian kernel and summing them to output the final blurred image. Gaussian filtering is more effective at smoothing images. It has been proven that neurons in the human visual perception system create a similar filter when processing visual images [29]. Gaussian smoothing is commonly the first step in edge and corner detection. It has the following characteristics [28]:

The Gaussian filter is a non-uniform low pass filter.

The kernel coefficients diminish with increasing distance from the kernel's centre.

Central pixels have a higher weighting than those on the periphery.

Larger values of σ produce a wider peak (greater blurring).

Kernel size must increase with increasing σ to maintain the Gaussian nature of the filter.

Gaussian kernel coefficients depend on the value of σ.

At the edge of the mask, coefficients must be close to 0.

The kernel is rotationally symmetric with no directional bias.

Gaussian kernel is separable which allows fast computation.

Gaussian filters might not preserve image brightness.

Figure 4: left: the CLAHE enhanced image, Right: the resulting image after CLAHE enhancement and Gaussian Blurring with σ=5

3.2 Gaussian Blurring Limitations:

Gaussian filtering is very useful for noise and detail removing (figure 4). But, contrary to median filter, Gaussian filter is not effective at salt and pepper noise removing [23].

The resulting image is a smoothed image. In next step we will sharpen this image to improve the clarity of details in the image

4 Unsharp Masking

Sharpness describes the clarity of detail in a photo, and can be a valuable creative tool for emphasizing texture.

We use Unsharp masking (UM) to emphasize texture and Detail [21].
Unsharp masking filter, also known as edge enhancement filter, is a simple operator to enhance the appearance of detail by increasing small-scale acutance without creating additional detail. The name was given because this operator improves details and other high frequency components in edge area via a process by subtracting a blurred version of the original image from the first one.

[1]http://www.robots.ox.ac.uk/~vgg/data/data-aff.html

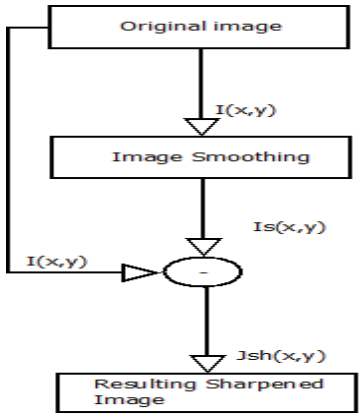


Fig. 5: Block diagram of the classical Unsharp masking

4.1 UM Concept:

Sharpening is a simple spacial filtering concept that produces an enhanced image J by increasing the contrast of the given image I along edges, without adding too much noise within homogeneous regions in the image.
The principle of UM is quite simple [21] [22]:
First a blurred version of the original image is created (we use a Gaussian blurring filter in our case).
Then, this one is subtracted from the original image to detect the presence of edges, creating the unsharp mask.
Finally this created mask is used to selectively increase the contrast of theses edges (fig. 5).
Mathematically this is represented by (5):

$$J_{sh}(x,y) = I(x,y) - I_s(x,y) \quad (5)$$

Where $J_{sh}(x,y)$ is the sharpened resulting image
$I(x,y)$ is the original image
$I_s(x,y)$ is the smoothed version of $f(x,y)$ obtained by

$$I_s(x,y) = I(x,y) - \{I(x,y) * HPF\} \quad (6)$$

Where HPF is a height pass filter.

4.2 UM Limitations:

Unsharp masking is a very powerful method to sharpen images (fig. 6). But, too much sharpening can also introduce undesirable effects such as "halo artifacts". These are visible as light/dark outlines or halos near edges. Halos artifacts become a problem when the light and dark over and undershoots become so large that they are clearly visible at the intended viewing distance.
Here we are addressing only the gray level images. The sharpened image may contain some introduced noise. To reduce this probably introduced noise we will apply a Laplacian filter to a smoothed version of the previously sharpened image.

Fig. 6: left: the CLAHE enhanced and Gaussian Blurred image, Right: the resulting image after Clahe enhancement and Gaussian Blurring and Unsharp Masking

5 Laplacian Filtering

Sharpening filters are used in order to highlight fine details within an image. They are based on first and second order derivates. First order derivatives are used to produce thicker edges in an image and are usually used for edge extraction.
Second order derivatives on the other hand, have a stronger response to fine detail and are usually better for image enhancement than the first order derivatives.

5.1 Laplacian Filtering Concept:

Laplacian filter is a second order or second derivative filter of enhancement. It is used to find areas of rapid change (edges) in images. Any feature with a sharp discontinuity (like noise) is enhanced by a Laplacian filter [30].
Since derivative filters are very sensitive to noise, it is common to smooth the image (e.g., using median filter, Gaussian filter…) before applying the Laplacian. If the Laplacian filter is used with a Gaussian filter, this process is called the Laplcaian of Gaussian (LoG).

The Laplacian is a linear operator; it forms an isotropic filter and is one of the simplest sharpening filters. In order to get a sharpened image, typically, the resulting Laplacian filtered image is added to the original image [31].

The Laplacian is given by (7)

$$L(x,y) = \nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \quad (7)$$

Where the partial 1st order derivative in the x direction is defined as follows (an approximation):

$$\frac{\partial^2 f}{\partial^2 x} = f(x+1,y) + f(x-1,y) - 2f(x,y) \quad (8)$$

and in the y direction as follows:

$$\frac{\partial^2 f}{\partial^2 y} = f(x,y+1) + f(x,y-1) - 2f(x,y) \quad (9)$$

Replacing (8) and (9) in (7) give the final approximation of the Laplacian filter:

$$\nabla^2 f(x,y) = [f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1)] - 4f(x,y) \quad (10)$$

Finally (10) can be represented with the following matrix that's used to implement the digital Laplacian:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (11)$$

There are other slightly different versions of Laplacian implementation [31].
Applying the kernel shown in (11) give the result showed in (Figure 7)

5.2 Laplacian Filtering Limitations:

Laplacian is a second derivative operator, which make it very sensitive to noise. Thus, it has the effect of enhancing noise as well as the structure. Therefore, it should be applied only in areas which have low noise, or areas which have been

subjected to noise reduction operator (such as image smoothing).

Fig. 7: left: the resulting image after Clahe enhancement, Gaussian Blurring and Unsharp Masking, Right: the resulting image after Clahe enhancement, Gaussian Blurring, Unsharp Masking and Laplacian filtering.

Fig. 8: left: The resulting image after Clahe enhancement, Gaussian Blurring, Unsharp Masking and Laplacian filtering, Right: the resulting image after subtracting the left image from the CLAHE enhancement image.

The filtered image should be subtracted from the original one to get the final sharpened image (Figure 8).

6 Performance evaluation

To prove the efficiency of our HBCCE approach on enhancing the features detection, we test our approach extensively following the evaluation method and datasets proposed by Mikolajczyk and Schmid [14]. We use the same dataset used in [14] and available online[2].

In this paper, the Open Computer Vision (OpenCV)[3] is used as the implementation framework for BRISK detector [4] (we use the last stable available version 2.4.9). HBCCE is implemented using C++ language.

We test our approach on a PC with CPU: INTEL(R) PENTIUM(R) 2.13 GHZ dual core, RAM: 3GO and windows 7 Ultimate Edition (32 bits) as an operating system.

Each of the datasets contains a sequence of six images presenting an increasing amount of transformation.

Our proposed method is evaluated against six image transformations: scale changes, view changes (Graffiti and Wall), zoom and rotation changes (Bark and Boat), illumination changes (Leuven), image blur (Bikes and Trees) and JPEG compression (Ubc).

6.1 Repeatability enhancement:

(Figure 11) shows the result of applying HBCCE stack effect on enhancing the repeatability of the BRISK [4] detector. As defined in [20], the repeatability is the ratio between the corresponding keypoints and the minimum total number of keypoints visible in both images.

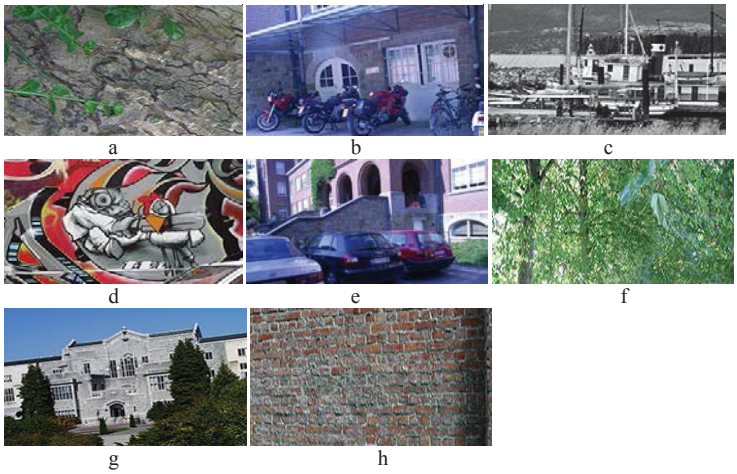


Fig. 9: Examples of images used for the evaluation: viewpoint change (Graffiti (d) and Wall (h)), zoom and rotation (Bark (a) and Boat (c)), JPEG compression (Ubc (e)), brightness change (Leuven (g)), and blur (Bikes (b) and Trees (f)).

As we can see in (fig. 11), the HBCCE stack drastically enhance the repeatability of the BRISK [4] detector.
We have also tested HBCEE on AGAST [2] and FAST [1] with similarly great success.

6.2 Results:

Results show that the proposed algorithm, Histo-Binary Combined Corner Enhancement (HBCCE), improves drastically the quality of the features detected while delivering comparable computation time. (Figure 10) illustrates the quality image enhancement introduced by the algorithm. HBCCE corrected the lighting effect, enhanced the appearance of details by increasing small-scale acutance, reduced the Gaussian noise and highlighted the fine details with second order spatial derivative to produce the final enhanced image. (Figure 11) shows the repeatability scores for 50% overlap error of the BRISK [4] and the BRISK-HBCCE detector. It is clearly shown that the proposed algorithm enhances the repeatability of the detector from 10% to 40%.

Figure 10: Left: The original image taken from dataset proposed in [20], Right: The resulting image after applying the HBCCE enhancement method.

[2]http://www.robots.ox.ac.uk/~vgg/data/data-aff.html
[3]http://docs.opencv.org/index.html

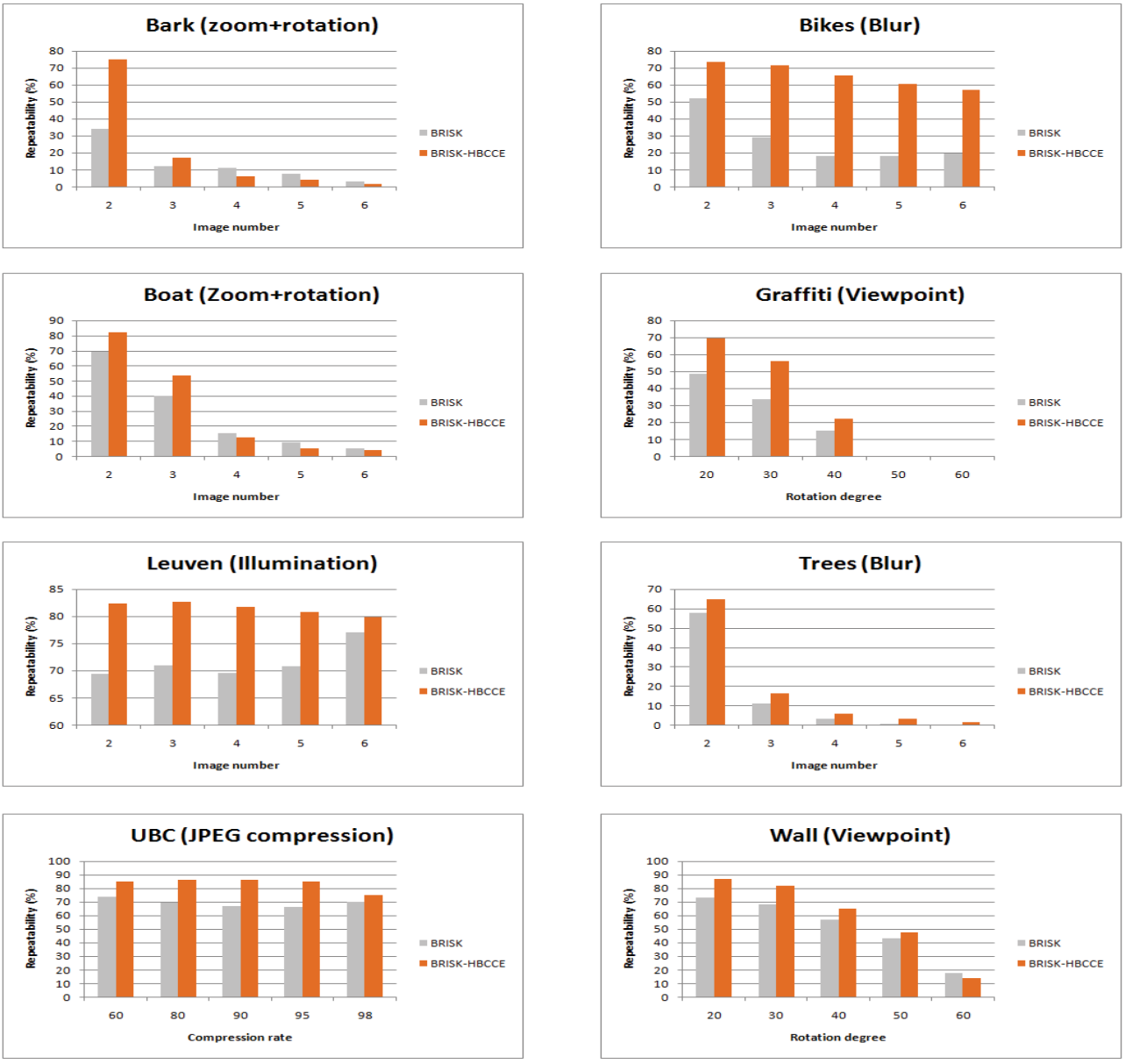


Fig. 11: Repeatability scores for 50% overlap error of the BRISK and the BRISK-HBCCE detector

7 Conclusions:

We have presented a novel approach named HBCCE, which Effectively combines different preprocessing algorithms to produce a very efficient preprocessing stack.

We have demonstrated the efficiency of our approach by exhaustively testing it on a standards dataset, and shown the improvement introduced by the proposed method on detectors repeatability.

8 Acknowledgements

This research was supported by the INVENTIVE Technologies laboratory[4] , a part of the Creargie MediaScan[5] in CASABLANCA Morocco and the TT&MIA (Télédétection spatiale-Traitement signal/image & Maths appliquées-Informatique-Aide à la decision) from university Abdelmalek Essaâdi. We are grateful to Mr. Dominique Schwartz the CEO of INVENTIVE Technologies and Creargie MediaScan for his valuable inputs, as well as to many other colleagues at TT&MIA and INVENTIVE Technologies laboratory for very helpful discussions.

[4]http://www.inventive-technologies.com/

[5]http://www.creargie.com/

9 References

[1] E. Rosten, R. Porter and T. Drummond, FASTER and better: A machine learning approach to corner detection, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 32, pp.105-119, 2010.

[2] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger. "Adaptive and generic corner detection based on the accelerated segment test". In Proceedings of the European Conference on Computer Vision (ECCV), 2010.

[3] Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: an efficient alternative to SIFT or SURF. In ICCV'11, 15th IEEE International Conference on Computer Vision, pages 2564–2571.

[4] Stefan Leutenegger, Margarita Chli and Roland Siegwart: BRISK: Binary Robust Invariant Scalable Keypoints," in IEEE International Conference on Computer Vision (ICCV), 2011: 2548-2555.

[5] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-2, no. 2, pp. 165–168, Feb. 1980.

[6] S. M. Smith and J. M. Brady, "SUSAN—A new approach to low level image processing," Int. J. Comput. Vis., vol. 23, no. 1, pp. 45–78, 1997.

[7] P. Saint-Marc, J. S. Chen, and G. Medioni, "Adaptive smoothing: A general tool for early vision," IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 6, pp. 514–529, Jun. 1991.

[8] M. Hillebrand and C. H. Müller, "On outlier robust corner-preserving methods for reconstructing noisy images," Ann. Statist., vol. 35, no. 1, pp. 132–165, 2007.

[9] W. K. Pratt, Digital Image Processing. New York: Wiley, 1978.

[10] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in Proc. Int. Conf. Comput. Vision, 1998, pp. 839–846.

[11] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in Proc. IEEE Int. Conf. Computer Vision, Kerkyra, Greece, 1999, pp. 1197–1203.

[12] L. I. Rudin, S. O. Stanley, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Phys. D, pp. 250–268, 1992.

[13] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 7, pp. 629–639, Jul. 1990.

[14] G. Winkler, K. Hahn, and V. Aurich, "A brief survey of recent edgepreserving smoothers," in Proc. 5th German-RussianWorkshop on Pattern Recognition and Image Understanding, B. Radig, H. Niemann, Y. Zhuravlev, I. Gourevitch, and I. Laptev, Eds., Herrsching, Germany, Sep. 21-25, 1998, pp. 62–69.

[15] John Canny. A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-8(6):679–698, Nov. 1986

[16] C. Harris and M.J. Stephens. A combined corner and edge detector. In Alvey Vision Conference, pages 147–152, 1988

[17] Pizer S, Zimmerman JB, Staab EV: Adaptive grey level assignment in CTscan display. J comput Assist Tomogr 8:300-305, 1984

[18] Zimmerman, JB, SM Pizer, EV Staab, JR Perry, W McCartney, BC Brenton, "An Evaluation of the Effectiveness of Adaptive Histogram Equalization for Contrast Enhancement", IEEE Trans. Med. Imaging, 7(4): 304-312, 1988.

[19] Suprijanto, Gianto, E. Juliastuti, Azhari, and Lusi Epsilawati, "Image Contrast Enhancement for Film-Based Dental Panoramic Radiography," in International Conference on System Engineering and Technology, Bandung, Indonesia, 2012.

[20] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", In Proceedings of Computer Vision and Pattern Recognition, pp. 257-264, 2003.

[21] N. M. Kwok, H. Y. Shi, G. Fang, and Q. P. Ha, "Intensity-based gain adaptive unsharp masking for image contrast enhancement," Image and Signal Processing (CISP), 2012 5th International Congress on, On page(s): 529-533.

[22] F. Y. M. Lure, P. W. Jones and R. S. Gaborski, "Multiresolution unsharp masking technique for mammogram image enhancement," Proc. SPIE Med. Imag., pp.830 -839 1996.

[23] J. Bednar and T.L. Watt, 'Alpha-trimmed means and their relationship to median filters', IEEE Trans. Acoust., Speech, Signal Processing, Vol. 32, No.1, pp.145-153. 1984

[24] P. Perona and J. Malik, "Scale -space and edge detection using anisotropic diffusion," IEEE Trans. Pattern Anal. Machine Intell., vol. 12, pp. 629–639, 1990.

[25] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," Proc. IEEE Int. Conf. Computer Vision, pp. 839–846., 1998.

[26] Tamer Rabie, 'Robust Estimation Approach for Blind Denoising', IEEE Trans. on Image Processing, Vol. 14, No. 11, pp.1755-1765, 2005.

[27] R. Garnett , Timothy Huegerich and Charles Chui, 'A Universal Noise Removal Algorithm with an Impulse Detector' IEEE Trans. on Image Processing, Vol. 14, No.11, pp.1747-1754, 2005.

[28] A. Chakrabarti, T. Zickler, and W. T. Freeman, "Analyzing spatially varying blur," in Proc. IEEE CVPR, Jun. 2010, pp. 2512–2519.

[29] A.Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In IEEE Conference on Computer Vision and Pattern Recognition, 2012. CVPR 2012 Open Source Award Winner

[30] X. Yan, M. Zhou, L. Xu, W. Liu and G. Yang, "Noise Removal of MRI data with Edge Enhancing", IEEE 5th International Conference on Bioinformatics and Biomedical Engineering, (iCBBE), (2011) May 10-12, pp. 1-4; Wuhan, China.

[31] T.Q. Phan, P. Shivakumara and C.L. Tan, "A Laplacian Method for Video Text Detection", In Proc. ICDAR 2009, pp. 66-70.

[32] A. El Harraj and N. Raissouni: "Toward Indoor and Outdoor Surveillance Using an Improved Fast Background Subtraction Algorithm" International Journal of Computer, Control, Quantum and Information Engineering, vol 9, no. 4, pp 595-600, April 2015

Small-scale objects extraction in digital images

V. Volkov[1,2] **S. Bobylev**[1]
[1]Radioengineering Dept., The Bonch-Bruevich State Telecommunications Univ., Saint-Petersburg, Russia
[2]Radioengineering Dept., State Univ. of Aerospace Instrumentation, Saint-Petersburg, Russia

Abstract *- Detection and localization problem of extended small-scale objects with different sizes and shapes appears in radio observation systems which use SAR, infra-red, lidar and television camera. Intensive non-stationary background is the main difficulty for processing. The problem of extraction small-scale objects is solved here on the basis of directional filtering, adaptive thresholding and morphological analysis. An advanced method of dynamical adaptive threshold setting is investigated which is based on isolated fragments extraction after thresholding. Hierarchy of isolated fragments on binary image is proposed for the analysis of segmentation results. The method uses extraction of isolated fragments in binary image and counting points in these fragments. Number of points in extracted fragments is normalized to the total number of points exceeding threshold and is used as effectiveness of extraction for these fragments. New method for adaptive threshold setting and control maximises effectiveness of extraction. It has optimality properties for objects extraction in normal noise field and shows effective results for real SAR images.*

Keywords: Filtering, Segmentation, Extraction, Automatic threshold control

1 Introduction

The task of detection and localization of small extended objects in noisy images occur in the electronic surveillance systems using radars with SAR, infrared and laser systems, as well as television cameras [1,2]. This task is relevant, because these facilities typically have an artificial origin and are of Prime interest.

Upon detection, extraction and localization of such objects had substantial difficulties in obtaining effective algorithms and structures processing, as in taken images there is intense and non-stationary background, also contains elements that are structurally similar to the signals, the signal/background is usually small, and the registered digital image has a low quality, small number of quantization levels, patency character and fuzzy borders of natural and artificial structures (rivers, roads, bridges, buildings). Statistics background is very different from a Gaussian, the distribution is clearly asymmetric, and the tails of the distributions like lognormal density normal or mixed (contaminated-normal), and when small numbers of samples are identified with difficulty. Such a character background virtually eliminates the use of the known methods of thresholding, since improper formation thresholds can cause loss of useful objects at a very early stage of processing. We cannot use traditional methods of detecting contours in images in order to highlight natural features (rivers, roads, borders, windbreaks, etc.), which are based on the spatial derivatives (gradients and Laplacians), because the result will be a significant growth impulse noise without visible effect for selection of quality circuit.

The basic principles that allow solving this difficult problem, are the location-based filtering, adaptive thresholding and selection of useful sites on the connectivity of neighboring pixels given the length of the useful structures [2].

2 Problem statement and method of object detection

Suppose there is a classification problem, so that objects have to be extracted belong to certain classes. But these classes have no precise description and this is the object of investigation. One of the main features for object segmentation is its extension. Useful objects are usually extensive, after binarization they consist of connected points, and look like lines or blobs.

Solving of segmentation problem includes thresholding and selection of objects with different extensions. Threshold processing gives binary image and the following segmentation is realized with special method of extraction. Dynamical threshold setting is the problem which should be solved with taking into account results of segmentation.

The aim of this paper is to develop method for segmentation and extraction of extensive objects with unknown sizes and orientations. Method proposed includes extraction of isolated fragments after binarization of pre-filtered image. Method includes logical filtering with corresponding masks and allows extracting isolated fragments with different sizes and orientations.

Different isolated fragments differ in their sizes and orientation. Hierarchy of isolated fragments is proposed which relates to their extensions and orientations and *characterizing mask* of fragment is inserted.

Setting of threshold is the main problem for qualitative segmentation. Low threshold level gives much noise, and the following processing becomes inefficient, too high level results in destroying of useful objects which may be split up to small fragments. The best threshold should be set after

analysis of segmented fragments with the use of some quality indicator for extraction and segmentation.

It is desirable to get some attributes which characterize the quality of segmentation. The simplest attribute is proposed here as the number of points at each step of extraction. It should be normalized to the initial points in binary image and represents the *effectiveness of extraction* at corresponding step.

It is worth noting that threshold setting is dependent on pre-filtering processing. The general idea of threshold setting and control via results of segmentation is admissible for different pre-filtering algorithms. It also may be used for local thresholds in sliding windows.

Consider there is an image in digital form, containing useful small-scale objects, which have a relatively small length in relation to the size of the entire image and an arbitrary orientation. Shape of objects of interest can be linear or speckle, and their length is specified by specifying a maximum size or length of the object in pixels, and set the minimum and maximum bounds on the length of objects. The problem feature is that the emergences of small-scale objects of interest practically no effect on the integral characteristics of the image.

There are two examples of extended small-scaled objects in SAR images which are shown in Fig. 1 and Fig. 2. These objects are artificially highlighted with a white oval. First image contains two oriented linear objects and the second image contains two blob-like objects inside corresponding white ovals.

The general structure of digital image processing includes a pre-filter, binary quantization (threshold processing), and subsequent morphological processing (Fig. 3). The input image after registration is submitted in digital form (two-dimensional array on a rectangular grid of points).

The problem of automatic setting of the threshold in autonomous information and control systems is very important for segmentation [3,4].

Well-known installation methods of global and local thresholds typically use histograms or local properties of the pixels in the image [2-4]. In our case, the threshold processing should depend on the results of binarization [5,6].

The purpose of this paper is to study adaptive method threshold setting for the detection and selection of objects based on structural decomposition of a binary image into elementary, isolated objects, analysis of the impact of the threshold on the results of the decomposition, and algorithm development for installation and changes the threshold in accordance with the results of the decomposition.

3 Preliminary filtering

Pre-filtering aims to improve the image and highlight the differences and boundaries. It is assumed that the useful objects always have a higher intensity relative to the background; otherwise it is necessary to invert the image. In this case we applied the differentiating filters (Laplacian type), which permit to use a global threshold for binary quantization exceeding the intensity threshold quantization.

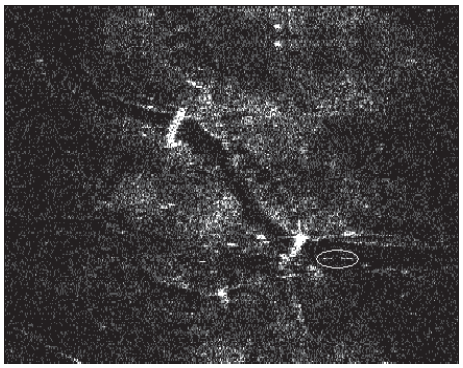


Fig. 1. Small-scaled oriented objects to be extracted

Fig. 2. Small-scale blob-like objects to be extracted

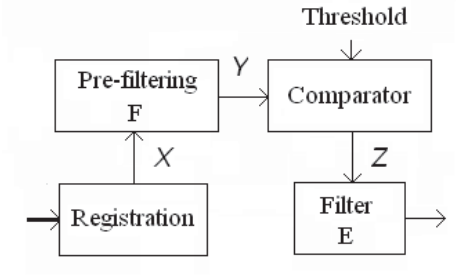


Fig. 3. Image processing structure for objects extraction

When filtering oriented linear objects we used space-oriented mask filter of the following form (Fig. 4), which would have effectively allocate endpoints of the segments of unknown length. In this particular case the coefficients are $a = 1$, $b = -1$. Several channels should be organized if objects have arbitrary orientation. They may use such masks with different orientations for searching all possible small-scale objects.

For blob-like small-scaled objects we used non-oriented (Laplacian-like) filter mask which is presented in Fig. 5.

Along with averaging in differentiating filters used other operators such as the sample median and selection the maximum value (in cells with coefficients *b*). The results of using the averaging filters are shown in Fig. 6 and Fig.7 below.

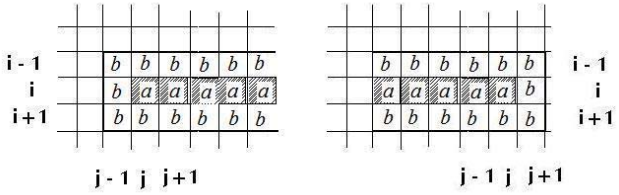


Fig. 4. Space-oriented filter masks

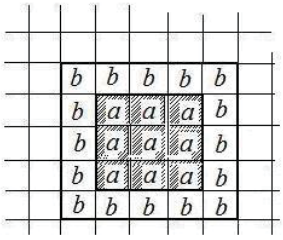


Fig. 5. Non-oriented filter mask

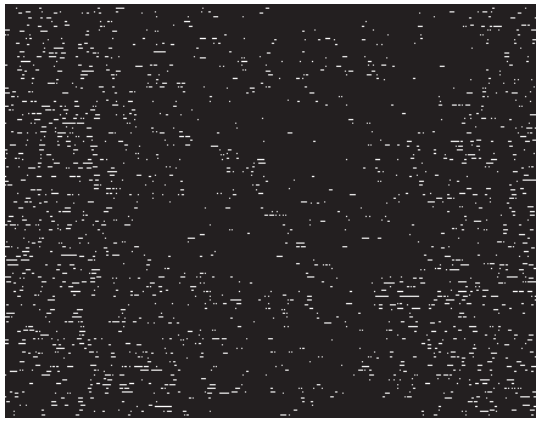


Fig. 6. The output of pre-filtering of image in Fig. 1 with space-oriented mask

4 Threshold processing

This stage is very important. Incorrect threshold often results in irreversible losses of information. Suppose we are interested in objects with high intensity, and processing results in high level for pixels if threshold level is exceeded, and zero level otherwise. Fig. 8 – Fig. 10 illustrate changes of binary images upon threshold variations from low to high levels for image shown in Fig.7. As it may be easily observed from Fig.8, low threshold (it was determined by Otsu method) is not appropriate because of segmentation problems: useful and noisy objects do not differ in extensions.

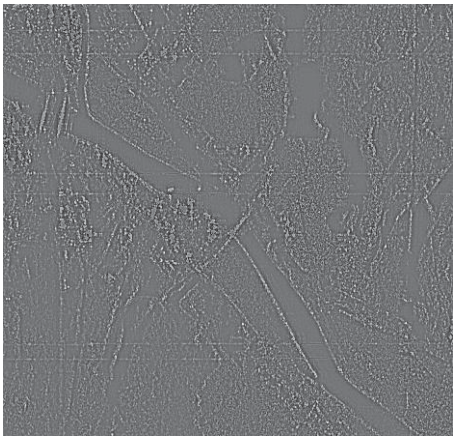


Fig. 7. The output of pre-filtering of image in Fig. 2 with non-oriented mask

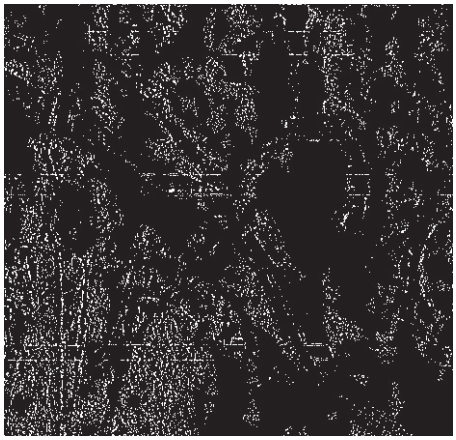


Fig. 8. Binary image after low threshold level

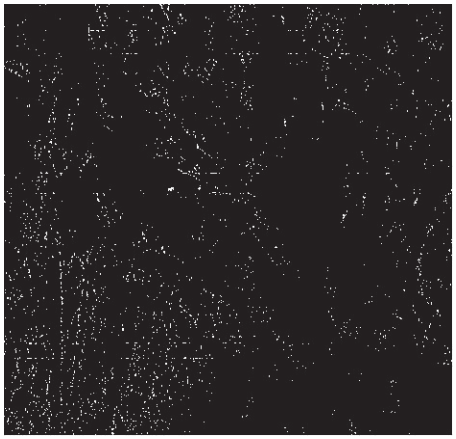


Fig. 9. Binary image after intermediate threshold level

Fig. 10. Binary image after high threshold level

As far as threshold level rises, differences between extensions of objects start to appear, but at very high levels we can see destroying of useful extensive objects. The best threshold level gives maximal differences in extensions between useful and noise objects.

The main idea is to set threshold level according to segmentation results. For this purpose hierarchy of isolated fragments is proposed, and effectiveness of extraction is inserted as indicator of segmentation degree. It may be used for threshold setting and control.

5 Hierarchy of isolated fragments in binary image

Our aim is to find out attributes of image which characterize extension properties of objects and allow us to control threshold level for qualitative segmentation. Suppose we have binary image after threshold processing. The concept of *characterizing mask* for isolated fragments is introduced for analysis of segmentation results. It relates to definition of continuity and adjacency of pixels, here usual definitions are used [2,6,7].

Fragments are isolated if they have no mutual pixels. Suppose isolated fragment consists of several adjacent pixels on the binary image. Extension properties of isolated fragment may be characterized by size of minimal rectangular mask which entirely covers this fragment. Objects may include several isolated fragments. Extensive objects usually contain extensive fragments.

Fig. 11 contains results of threshold binarization for Gaussian noise field with zero mean and standard deviation equals to unity with threshold level equals to 1.5. Top picture in Fig.12 represents isolated points. After deleting isolated points there are lots of connected fragments with different sizes and shapes (bottom picture in Fig. 12).

It is possible to represent hierarchy of small isolated fragments which is shown in Fig. 13 up to characterizing mask 3x3. Isolated point has characterizing mask 1x1, it is shown at high left corner of the picture.

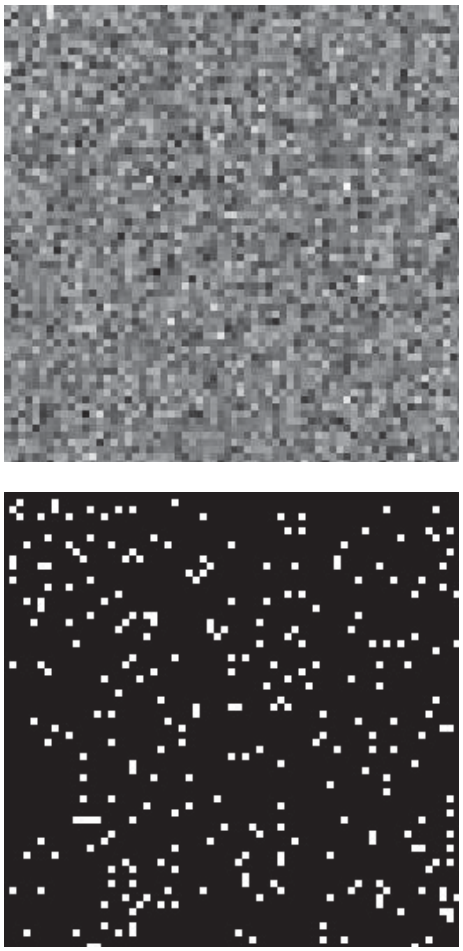


Fig. 11. Gaussian noise field (top picture) and the result of binarization (bottom picture) with threshold level equals to 1.5

Isolated pairs may have characterizing masks 1x2, 2x1 or 2x2 corresponding to their orientations. Triple of points may form horizontal line (at high right corner of figure), vertical line (at low left corner), or diagonal lines (slash and inverted slash). Other fragments with three points may have different characterizing masks.

Triple of points may look like “corner”, and have characterizing mask 2x2. Extensive fragment with horizontal orientation has characterizing mask 2x3 if it is more in sizes than 2x2 but entirely covers by mask 2x3. Similar fragment with vertical orientation covers by characterizing mask 3x2, and so on. “Gate” or “scoop” with three points has 2x3 or 3x2 masks. Four points give large variety of isolated fragments. Squares 2x2 and 3x3 do not consider fragment orientations.

Hierarchy of fragments is their ordering and may be obtained by the choice of characterizing masks to be considered. Masks 2x3, 3x5, 3x7, 5x7 and other similar masks characterize horizontal fragments; masks 3x2, 5x3, 7x3, 7x5 describe vertical fragments.

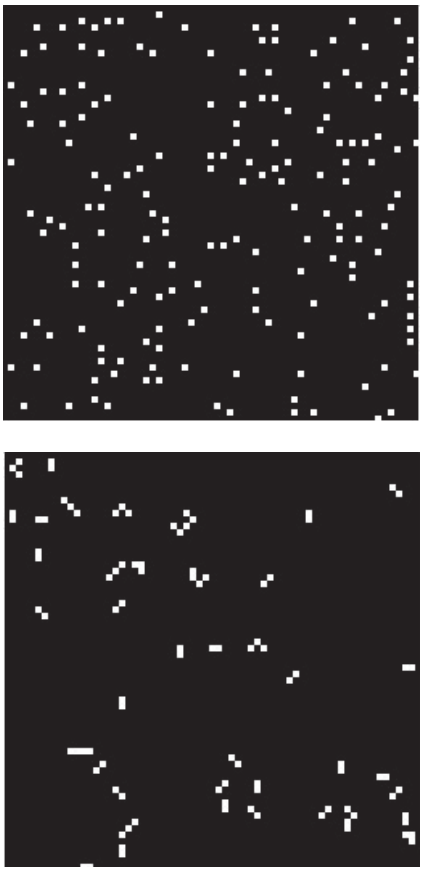


Fig. 12. Isolated points extracted (top picture) and remaining small-scale objects (bottom picture) after thresholding of Gaussian noise field

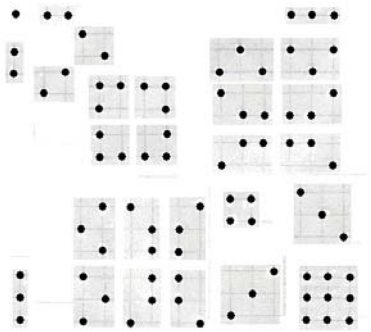


Fig. 13. Hierarchy of small isolated fragments

The choice of set of characterizing mask allows us to obtain different attributes for describing fragments with different extensions and orientations.

The simplest hierarchy uses only square characterizing masks. Then we have 1x1, 2x2, 3x3, 5x5, 7x7 and so on, as characterizing masks of extensive fragments with increasing extensions. This hierarchy does not take into account orientations of fragments.

6 Threshold setting and control by the use of connected fragments in noise image

It is easy to verify that the number of extracted small-scale objects of the given type in binary image is small at both very low and high thresholds. Thus, with some intermediate value of the threshold number of such objects is maximal. However, along with the decrease in the number of the objects by increasing the threshold also decreases the total number of points in which intensities exceed the threshold. Let after binary quantization with a predetermined threshold value the image contains only $N(T)$ pixels exceeding the threshold. At each step of extraction we use corresponding characterizing masks 1x1, 2x2 and so on. Then at each step of the extraction binary image loses $N_1(T)$, $N_2(T)$ ...etc. points. Since all these numbers depend on the quantization threshold, it is necessary to perform the normalization, and to consider the relative values $N_1(T)/N(T)$, $N_2(T)/N(T)$ and so on.

These values can be considered as the estimates of *effectiveness of extraction* of fragments at the appropriate step with given quantization threshold. For homogeneous Gaussian noise field we can calculate probability that intensity in each pixel exceeds the threshold

$$P(T) = 1 - \Phi(T), \tag{1}$$

where $\Phi(T)$ is a cumulative distribution function.

Probability to occur an isolated point in the square 3x3 will be

$$P_1(T) = \Phi^8(T)[1 - \Phi(T)], \tag{2}$$

and probability for any fragment with characterizing mask 2x2 is

$$\begin{aligned} P_2(T) &= 2\Phi^{10}(T)[1-\Phi(T)]^2[1+\Phi^2(T)] \\ &+ \Phi^{12}(T)[1-\Phi(T)]^3[1+3\Phi] \end{aligned}, \tag{3}$$

Normalized values $E_1 = P_1(T)/P(T)$, $E_2 = P_2(T)/P(T)$ and so on were calculated and represented on the top of Fig. 14. Numbers of lines correspond to characterizing masks 1x1, 2x2, 3x3, etc. Lines for connected fragments have evident maxima at the threshold value 1.3. This is the best threshold for extracting small connected fragments from Gaussian noise field. These results are confirmed by simulation curves represented on the bottom of Fig. 14.

7 Adaptive thresholding structure

Processing structure containing adaptive threshold setting and control is shown in Fig. 15. Filter F produces pre-filtering, filter E realizes morphological operation for extraction isolated fragments with given characterizing masks. Maximum selector obtains the threshold value which

maximizes efficiency of extraction, length selector extracts small-scale objects with prescribed lengths.

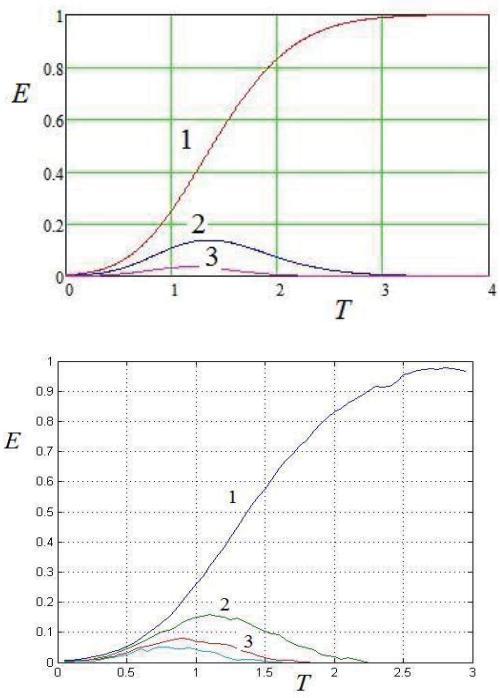


Fig. 14. Extraction efficiency for isolated points (line 1), fragments with characterizing mask 2x2 (line 2), and mask 3x3 (line 3)

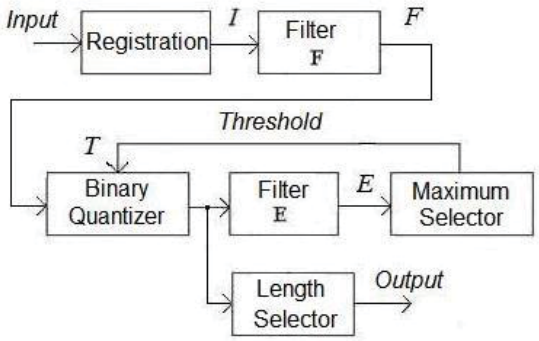


Fig. 15. Processing structure with adaptive threshold setting and control

8 Extraction of small-scale objects from real SAR images

The results of the extraction and allocation of small-scale objects on real radar images presented in Fig. 16 and Fig 17 for the two images from Fig. 1 and Fig. 2 accordingly.
Top pictures represent outputs of binary quantizers which threshold were set for maximum extraction efficiency for small connected fragments. Objects of interest are highlighted by the ovals. Dependency plots of extraction efficiency upon threshold values are presented in the middle row of pictures. Numbers 1, 2 and 3 correspond to connectivity of pixels in fragments. Pictures on the bottom show objects of interest extracted by the use of length selection.

9 Conclusions

New method of extraction and allocation small-scale objects is described which is based on pre-filtering, adaptive thresholding and morphological selection. It allows extracting small-scale objects with different extension and orientation.

Hierarchy of isolated fragments is proposed for analysis of connected fragments. Analysis of small connected fragments is useful for obtaining indicators for threshold setting and control. Extraction efficiency is introduced as the relative number of points in fragments with respect to all points exceeding threshold. The best threshold level should give maximal extraction efficiency for a given size of characterizing mask.

Adaptive method for threshold setting and control has optimal property which was checked by modeling Gaussian field with extensive object region. Thresholds obtained are settled near the value of optimal maximal likelihood threshold for detection of shift on Gaussian field.

10 References

[1] Gui Gao. "Statistical modeling of SAR images: A Survey"; Sensors, Vol. 10, 775—795, 2010.

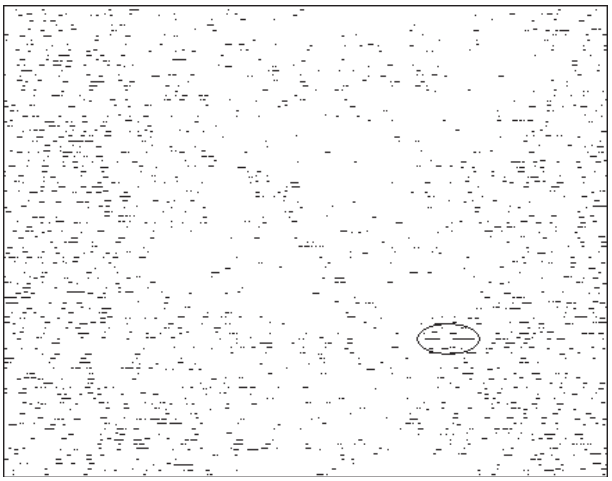
[2] R.C. Gonzalez, R.E. Woods, S.L. Eddins. "Digital Image Processing using MATLAB". Englewood Cliffs, NJ: Prentice-Hall, 2004.

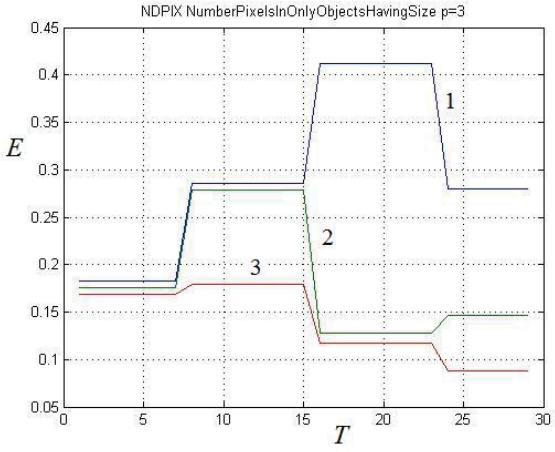
[3] H. Akcay, S. Aksoy. "Morphological Segmentation of Urban Structures"; Urban Remote Sensing Joint Event, 1—6, 11-13 April, Paris, 2007.

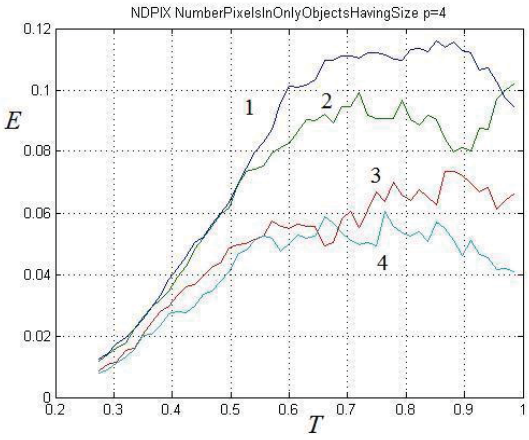
[4] M. Sezgin, B. Sankur. 2004. "Survey over image thresholding techniques and quantitative performance evaluation"; Journal of Electronic Imaging, Vol. 13(1), 146—165, 2004.

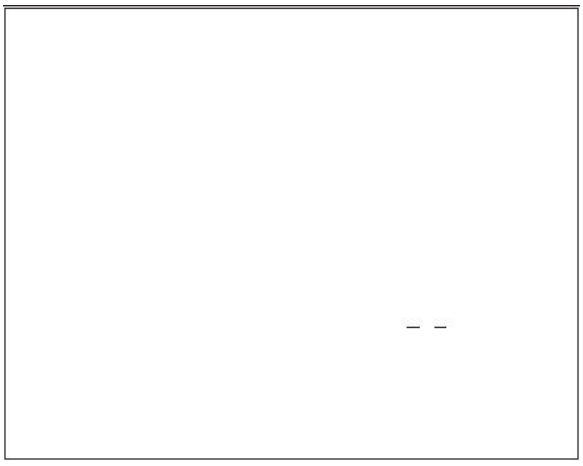
[5] V. Volkov. "Segmentation and extraction of extensive objects on digital Images"; Proceedings of the 2009 International Conference on Image Processing, Computer Vision and Pattern Recognition. IPCV'09, Vol. II. Las Vegas, Nevada, USA, CSREA Press, 656—662, 2009.

[6] V. Volkov. "Thresholding for segmentation and extraction of extensive objects on digital images"; Proceedings 32 Annual German Conference on Artificial Intelligence. KI 2009. Paderborn, Germany, Springer, 623—630; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5803 LNAI , P. 623—630. http://www.springer.com/computer/ai/book/978-3-642-04616-2.









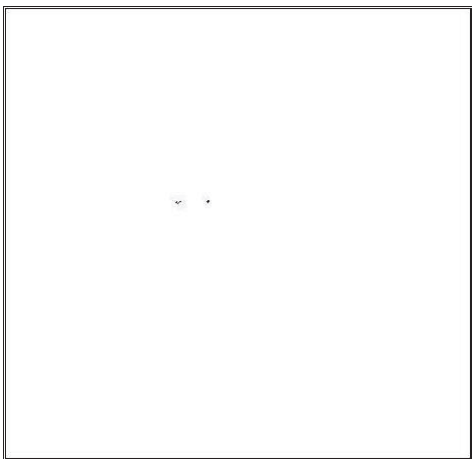


Fig. 16. Extraction of small-scale objects from the image Fig. 1

Fig. 17. Extraction of small-scale objects from the image Fig. 2

Low Level Processing for Digital Images

Indiara B. Vale[1], Géssica T. Marchiori[1], Dalton M. Tavares[1], Tércio A. Santos Filho[1], Antônio C. de Oliveira Jr.[1], Ivan S. Sendin[2], Sérgio F. da Silva[3], Marcos N. Rabelo[3] and Marcos A. Batista[3]

{indiarabarbosavale,gessicamarchiori,dmatsuo,tercioas}@gmail.com, sendin@ufu.br, sergio@ufg.br, rabelo@dmat.ufpe.br, marcos.batista@pq.cnpq.br

[1]IBiotec, Federal University of Goiás, Catalão, GO, Brasil

[2]FACOM, Federal University of Uberlândia, Uberlândia, MG, Brasil

[3]Graduate Program in Modeling and Optimization, Federal University of Goiás

Av. Dr. Lamartine Pinto de Avelar, 1120, CEP 75704-020

Abstract— *Image processing is a very important area of Computer Science and can be applied in many researches. In this paper we present some essential concepts about image processing. We also show the developed algorithms which helps in the understanding and application of the techniques. The presented results encourage new promising works.*

Keywords: Image Processing, Smoothing, Morphing, Characterization

1. Introduction

An image can be defined as a physical object representation. It is possible to store, manipulate and process the image according to the necessities [1]. These representation depends on the illumination source incident on the scene (*illumination*) and the amount of illumination reflected by the objects (*reflectance*). We denote images as a two-dimensional function $f(x, y)$, where x and y are spatial coordinates and f is the intensity in the point. So,

$$f(x, y) = i(x, y) \times r(x, y) \tag{1}$$

where i is the *illumination* $(0 < i < \infty)$ and r is the *reflectance* determined by the objects materials $(0 \leq r \leq 1)$, denoting 0 to total absorption and 1 to total reflectance.

In this paper, methods are applied in converted images from continuous to digital [2]. Digital images are represented as a discrete vector, commonly a 2-D matrix containing m rows and n columns:

$$f(x, y) = \begin{pmatrix} f(0,0) & \dots & f(0, n-1) \\ f(1,0) & \dots & f(1, n-1) \\ \vdots & \ddots & \vdots \\ f(m-1, 0) & \dots & f(m-1, n-1) \end{pmatrix}$$

Each point of the image has a color value, that can be a tuple of 3 values in RGB (Red, Green and Blue) space or the intensity of the image at that point. The intensity (gray level) of a monochrome image f at the coordinate (x, y) is denoted by

$$L = f(x, y) \tag{2}$$

From Eq. 2,

$$(L_{min} \leq L \leq L_{max}) \tag{3}$$

where L_{min} and L_{max} are positive and finite values.

The interval $h = [L_{min}, L_{max}]$ is called *gray scale*. Normally, the numerical representation is a integer interval $[0, W)$, where $L = 0$ means black and $L = W - 1$ is white [3]. All intermediate values varies from black to white.

Thus, mathematical representation of an image allow us to manipulate its content, apply geometric transformations or extract important information. These huge set of operations that can be performed in the image is called *image processing*. Then process an image consists of upgrade its appearance and/or prepare it to be measured [4]. Each element of the matrix-image is called *picture element* or *pixel*. Pixel size depends on spatial resolution of image acquisition. The picture element is the smallest part where operations are performed.

One of the most common operations performed on pixel is the *neighborhood operations*. A pixel p has horizontal and vertical neighbors whose can be denoted by a squared grid with 4-neighbors or 8-neighbors, and a diagonal grid with 6-neighbors (See Fig. 1). These set of pixels is denoted by $N(p)$ [5].

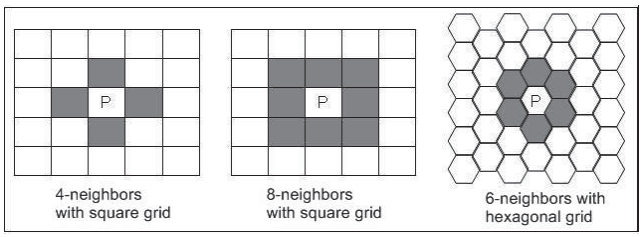


Fig. 1: Neighborhood relations of a pixel p

Digital images are manipulated in two different domains: frequency or spatial domain. In the frequency domain, methods are applied on Fourier transform, where it is possible to

obtain fast processing, but with high complexity. Considering that, the presented techniques are in the spatial domain, in other words, transformations are directly performed on the image pixel.

To manipulate images in spatial domain, we are going to use *templates* [6], [7]. Templates are represented by a matrix $3x3$, where the middle element is the main pixel and the others are the neighborhood. The result of an operation performed on all matrix elements replaces the main pixel value.

This paper is organized as follow. Section 2 describes all the proposed methods. Section 3 shows the results of performed experiments. And Section 4 presents the conclusions.

2. Methods

Low level image processing aims an image transformation that improves the final result of an application. We show here some methods that provides image manipulation, in the way to obtain informations about image content. Using these informations, it is also possible to get better results when some techniques are performed.

2.1 Image Smoothing

The first proposed method is *image smoothing*, also known as *blurring*. These method is capable of minimize unexpected effects resulted from image acquisition (e.g., noises). We use a weighted average filter, which is applied a template with mask $3x3$ that provides the 4-neighbors ($N_4(p)$) of middle pixel p. The obtained set of neighbors is used to calculate the weighted mean, considering all the pixels of the mask. The final value replaces the pixel p value. So, it works like a "sliding window" that goes through the image modifying pixels, based on the neighborhood.

The results are obtained by weighted average $\bar{p}$, denoted in Eq. 4, where x_i represents one pixel and p_i its respective weight.

$$\bar{p} = \frac{x_1 p_1 + x_2 p_2 + \ldots + x_n p_n}{p_1 + p_2 + \ldots + p_n} \tag{4}$$

The proposed algorithm is showed below. I denotes the original image, and the weights are integer numbers provided according to the used template.

Algorithm 1

```
copy ← I
for all p ∈ I do
    n ← [ ]
    n ← p neighbors ∪ p
    average = p̄, as Eq. 4.
    copy[p] = average
end for
return copy
```

2.2 Morphing

In this section, we present a basic technique of image morph that gradually transforms (or morphs) one image into another through a seamless transition, decreasing a opacity rate (α) from 1 to 0. During the process, intermediate images are obtained, and in the end, it is possible to put them together and see the morphing effect.

The operations takes two different images, the initial I_i one and the final I_f one, both with the same size. At each step, a third image T_α is obtained according to Eq. 5.

$$T_\alpha = (\alpha * I_i) + ((1 - \alpha) * I_f) \tag{5}$$

As α decrease, I_i is gradually distorted while I_f starts to appear. Thus, the first images are similar to the initial image, the middle images are between the initial and the final, and the last ones are close to the final image.

2.3 Characterization

Characterization process extracts from images useful features. These characteristics are utilized in others image processing applications. We present here an usual technique that consists in "count" the number of pixels p belonging to each gray level L of the image I. The process gives us a vector v containing informations about the image.

The Algorithm 2 explains how this process works.

Algorithm 2

```
v ← zeros
for all p ∈ I do
    v[L(p)] ← v[L(p)]+1
end for
return v
```

2.4 Normalization

In an attempt to optimize the final results, we applied the normalization process. To obtain the normalized vector n (i.e. convert to unit vector) each component is divided by the image size (See Algorithm 3).

Algorithm 3

```
v ← characteristic vector
s ← image size
for i ∈ length(v) do
    n[i] ← v[i] / s
end for
return n
```

This process in very important considering that normalized vectors are generally applied in images that are going to be posteriorly compared.

2.5 Comparison

Image comparison is used to obtain and analyze the relation between two images, in order to see their equalities and differences.

There are a lot of techniques to compare images, however, we made it by Euclidean distance. Given two vectors x and $y \in \mathbb{R}^n$, the distance is denoted by

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (6)$$

where $x = (x_1, x_2, ..., x_n), x_i \in \mathbb{R}$ e $y = (y_1, y_2, ..., y_n), y_i \in \mathbb{R}$.

Through these process, the reference image is compared with other images from the image bank. Thus, we conclude that smaller distances means more similarity with the reference.

The obtained results from the experiments are presented in Section 3.

3. Experiments

In order to perform the proposed experiments, we chose the programming language Python associated with the libraries OpenCV (Open Source Computer Vision Library) and PIL (Python Imagin Library).

Python was conceived in the late 1980s by Van Rossum. It is a general-purpose, interpreted and high-level programming language focused on code readability. Python uses dynamic typing and a combination of reference counting and a cycle-detecting garbage collector for memory management.

The biggest reason of our choice is that the language provides constructs intended to enable clear programs on both a small and large scale, and it is available for many operating systems.

OpenCV was started at Intel in 1999, and now it supports a multitude of algorithms related to Computer Vision and Machine Learning. It is also possible to combine OpenCV with variety of programming languages. So, OpenCV-Python gives the advantage that the code is as fast as the original C code (since it is the actual C++ code working in background). OpenCV-Python is a Python wrapper for the original OpenCV C++ implementation.

Python Imaging Library (PIL) is a free library for Python that adds image processing capabilities to the interpreter, such as support for opening, manipulating, and saving many different image file formats. It is also available for Windows, Mac OS X and Linux.

The libraries was independent used in each method and the obtained results are equivalent. However, the runtime was different for same methods, considering that the algorithms was processed by a Core 2 Duo with 3GB RAM, running Windows 7 32 bits.

This fact encouraged us to analyze the situation and create an evaluative graphic, presented in the end of this section.

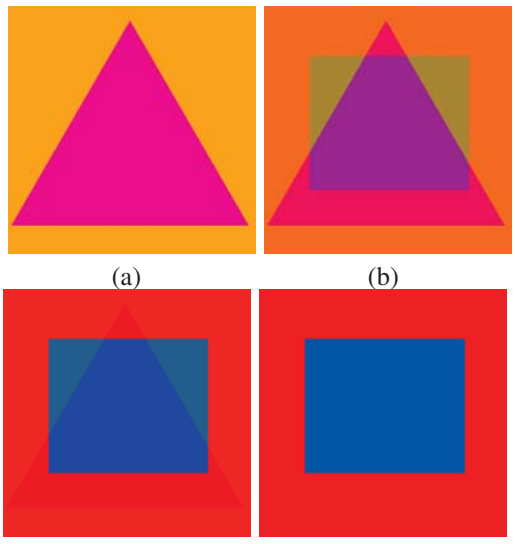
The first experiment is a blurred image, like was described in Section 2.1. The results are presented in Fig. 2, where it is possible to see the difference between the original and the final image.

(a) (b)

Fig. 2: Images from image smoothing process: (a) Original image; (b) Final blurred image.

As was described in Section 2.2, morphing process generates an image sequence that starts in the inicial imagem (Fig. 3(a)) and ends in the final image (3(d))).

Some of the images obtained during the process are shown on Figs. 3(b) and 3(c).



(a) (b)

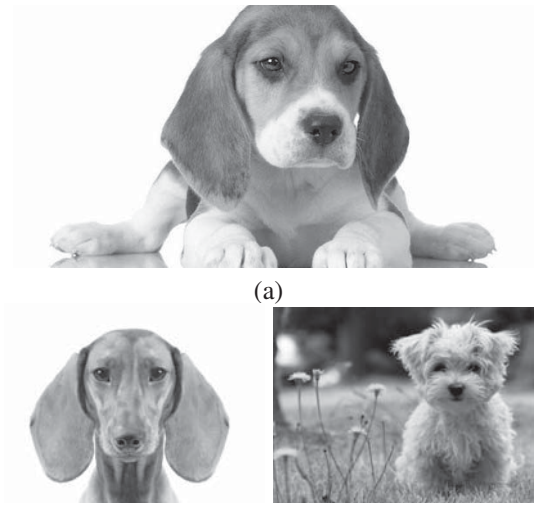
(c) (d)

Fig. 3: Images from morphing technique: (a) Original image; (b) Intermediate image from the beginning of the process; (c) Intermediate image from the end of the process; (d) Final image.

The main problem about these experiment is that the used images doesn't have enough details, what provides some intermediate images with non-visible transformations. Nevertheless, we obtained relevant results that contributes with the experiments.

The final experiment was about image comparison using Euclidean Distance, as we describe in Section 2.5.

It is also important to highlight that before comparison, the images underwent characterization and normalization process (Sections 2.3 and 2.4, respectively).



(a)

(b) (c)

Fig. 4: Compared images: (a) Reference image; (b) The most similar image, comparing to the reference; (c) The most different image, comparing to the reference.

The reference image is represented in Fig. 4(a). In the Fig. 4(b) we notice a lot of gray level similarities when it is compared with Fig. 4(a), what doesn't happen with Fig. 4(c).

Therefore, to ensure the presented results, we show a performance graphic from OpenCV and PIL for each method. (Note: Image smoothing method was performed only with OpenCV library).

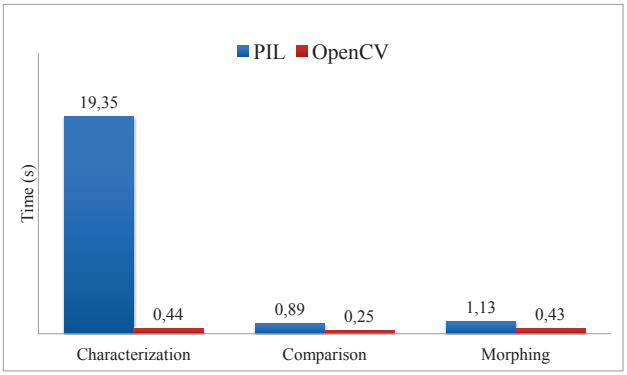


Fig. 5: Used libraries performance graphic

The graphic confirms the expected result: OpenCV-Python library is much faster than PIL, the native image library from Python. It happens due to the fact that, actually, C++ code works in background of OpenCV, providing faster runtimes.

Thus, we conclude that OpenCV is more appropriate to the proposed methods.

4. Conclusions

In this article we presented some basic and very important image processing techniques, that may be used as support for other researches.

All the methods (Image smoothing, Characterization, Comparison and Morphing process) showed satisfactory results during the experiments.

We also presented a graphic comparing the runtime of the used image processing libraries, OpenCV and PIL, that led us to the conclusion that OpenCV-Python library is more efficient.

Even considering the huge amount of complex image processing techniques existing in literature, the proposed methods provide an essential learning about image processing theory and its applications.

Finally, our results encourage new researches considering its consistency and relevancy, and we still working to achieve better solutions.

Acknowledgments

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) and Research Support Foundation of Goiás State (FAPEG) for the financial support.

References

[1] D. H. Ballard and C. M. Brown, *Computer Vision*, 1st ed. Prentice Hall Professional Technical Reference, 1982.

[2] W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3rd ed. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[3] O. Marques Filho and H. V. Neto, *Processamento digital de imagens*. Rio de Janeiro, Brasil: Brasport, 1999.

[4] J. C. Russ, *The Image Processing Handbook*. Boca Raton, FL, USA: CRC Press, Inc., 1992.

[5] F. Meyer, "Topographic distance and watershed lines," *Signal Process.*, vol. 38, no. 1, pp. 113–125, Jul. 1994. [Online]. Available: http://dx.doi.org/10.1016/0165-1684(94)90060-4

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*, 3rd ed. Prentice Hall, 2007. [Online]. Available: http://www.worldcat.org/isbn/013168728X

[7] F. L. S. NUNES, "Introdução ao processamento de imagens médicas para auxílio ao diagnóstico," in *Karin Breitman; Ricardo Anido. (Org.). Atualizações em Informática. 1 ed.*, F. L. S. NUNES, Ed. Rio de Janeiro: Elsevier Science, 2006, ch. 2, pp. 73–126.

Two Fast Alternating Direction Optimization Methods for Multiphase Segmentation

Lu Tan [1], **Weibo Wei** [2], **and Zhenkuan Pan** [2]
[1,2] College of Information Engineering, Qingdao University, Qingdao, Shandong, China
luna086@163.com, njustwwb@163.com

Abstract - *In this paper, two new methods associated with alternating direction optimization, fast alternating direction method of multipliers(FastADMM) and fast alternating minimization algorithm(FastAMA), are proposed for image segmentation using the multiphase Chan-Vese model, which is on the basis of piecewise constant optimal approximations. For these methods, we incorporate the variable splitting approach and a 'reset' condition in order to update the Lagrange multiplier and make sure the value of energy functional is always positive. The Osher and Stethian level set method, binary level set functions, thresholding method and projection formula are applied in the implementation. Finally, numerical results with rapid convergence are obtained by our methods, which are also compared with those of some other fast variational methods to demonstrate better effectiveness of our methods.*

Keywords: Multiphase segmentation, fast alternating direction method of multipliers, fast alternating minimization algorithm, active contours, level sets.

1 Introduction

In image segmentation, major advances were made in two-phase image segmentation[1-3] in the early days. Mumford-Shah model proposed by Mumford D and Shah J [4] is regarded as the most significant region-based model. It has been extended to a great deal of applications. In 2001, Tony F. Chan and Luminita A. Vese proposed the Chan-Vese model [5] for active contours to detect objects in a given image. It is one of the simplified variants of Mumford-Shah model. Nevertheless, as the complexity of the images increases, 2-phase image segmentation is not able to meet the actual needs. Therefore, multiphase segmentation is applied to satisfy the demands. On the basis of Potts model [6][7] from statistical mechanics, Zhao et al. [8] started to study multiphase motion segmentation by using the level set method and proposed a model which can represent n different regions by n level set functions. In order to reduce the number of level set functions, Chan et al. continued their work and proposed multiphase segmentation model [9] which is a generalization of Chan-Vese model. Their scheme can naturally avoid "overlap" and "leakage" problem. But there are still some problems about solving the global optimization, accuracy, stability and speed.

Alternating direction method of multipliers(ADMM) was first described by Glowinski and Marocco [10] and alternating minimization algorithm(AMA) was presented by Tseng [11]. These techniques are commonly known as the Split Bregman Method [12], and are known to be an efficient solver for problems involving the total-variation norm [13]. These methods can be accelerated using optimal first order methods, of the type first proposed by Nesterov [14]. And accelerated variants of ADMM and AMA can be called FastADMM and FastAMA.

In our paper, we design methods (FastADMM and FastAMA) that can achieve computational efficiency and own even faster convergence to solve the functional of multiphase Chan-Vese model based on binary level sets framework [15,16]. The gradient descent method (GDM) [17], Chambolle's dual method (DM) [18], alternating direction method of multipliers(ADMM) [19] i.e. the augmented Lagrangian method (ALM) [20], and alternating minimization algorithm(AMA) [11] are used to compare with our methods. But results of these methods which are used in multiphase segmentation model will be obtained in a slow convergence. Tom Goldstein et al. [19] introduced FastADMM and applied it to solve the TV model as an example and some other strongly convex problems. A predictor-corrector type acceleration step is used in this method.

The remaining of this paper is organized as follows. In Section 2, the binary level set formulation of the functional of multiphase Chan-Vese model used in our paper is reviewed along with its four traditional solution methods. Our proposed methods are discussed in Section 3 and its iterative discrete formulas for implementation will be presented in detail. In Section 4, some numerical experiments are given to illustrate the effectiveness of our method by comparing with other methods. Finally a conclusion is given in section 5.

2 The multiphase Chan-Vese model and its four traditional methods

2.1 The binary level set based formulation

In order to separate an image domain Ω into n subdomains with $\Omega = \bigcup_{i=1}^{n} \Omega_i$ and $\Omega_i \bigcap_{i \neq j} \Omega_j = \varnothing$. Vese and Chan defined up to $n = 2^m$ phases and m level set functions. This way makes sure that each pixel $(x, y) \in \Omega$ will belong to one, and only one phase.

The level set method proposed by Osher and Sethian [21] is an effective representation for evolving curves and surfaces because of automatic change of topology. The main idea of the level set formulation is to implicitly represent a given interface $\Gamma(t)$ as the zero level set of a Lipschitz continuous function $\phi(R^2 \to R)$. ϕ is defined as follows:

$$\begin{cases} \phi(x,t) > 0, \ if \ x \ is \ inside \ \Gamma(t) \\ \phi(x,t) = 0, \ if \ x \ is \ at \ \Gamma(t) \\ \phi(x,t) < 0, \ if \ x \ is \ outside \ \Gamma(t) \end{cases} \quad (1)$$

It is normal to define ϕ as a signed distance function in order to keep stability in numerical implementation. The distance function ϕ obeys the Eikonal equation.

$$\left|\nabla \phi(x,t)\right| = 1 \quad (2)$$

The variational level set method [13] gives a way to apply the level set function to the energy functional. For a given open region Ω with smooth boundary $\Gamma(t)$, simple facts can be got as follows:

$$length(\Gamma) = \int_\Omega \left|\nabla H(\phi)\right| dx = \int_\Omega \delta(\phi)\left|\nabla \phi\right| dx \quad (3)$$

$$area(\Omega) = \int_\Omega H(\phi)\, dx \quad (4)$$

where $H(x)$ and $\delta(x)$ are Heaviside function and Dirac delta function respectively. According to their work, for i = 1,2...,n, let $\left(b_{i-1}^1 b_{i-1}^2 ... b_{i-1}^m\right)$ be the binary representation of i-1, where $b_{i-1}^k = 0 \vee 1$. The characteristic function $\chi_i(x)$ of Ω_i can be written as the following general expression:

$$\chi_i(x) = \prod_{j=1}^{m} \left[b_{i-1}^j + (-1)^{b_{i-1}^j} H(\phi_j) \right]. \quad (5)$$

Then energy functional for n phases is obtained:

$$E(\phi) = \sum_{j=1}^{m} \int_\Omega \gamma \left|\nabla H(\phi_j)\right| dx + \sum_{i=1}^{n} \alpha_i \int_\Omega Q_i \chi_i dx \ , \quad (6)$$

where γ and $(\alpha_1, \alpha_2, ..., \alpha_n)$ is positive parameters. The function Q_i is defined as $(c_i - f)^2$, c_i is a constant vector which can be obtained by the mean intensity value of f inside Ω_i as follows:

$$c_i = \frac{\int_\Omega f \chi_i dx}{\int_\Omega \chi_i dx} \ . \quad (7)$$

On this basis, a new approach called a binary level set function is introduced by Johan Lie et al. [15] and Bresson et al. [16], which has a simpler definition about initialization of level set function. Firstly, assume that the interface is enclosing $\Omega_1 \subset \Omega$. A discontinuous level set function ϕ is used instead. It is defined as follows:

$$\phi(x) = \begin{cases} 1, \ if \ x \in int(\Omega_1) \\ 0, \ if \ x \in ext(\Omega_1) \end{cases} \quad (8)$$

If the level set function $\phi(x)$ satisfy $\phi(x)^2 = 1$, then we can use the basis function $\phi(x)$ to calculate the length of the boundary of Ω_1, and the area inside Ω_1.

$$length(\partial\Omega_1) = \int_\Omega \left|\nabla \phi(x)\right| dx \quad (9)$$

$$area(\Omega_1) = \int_\Omega \phi(x)\, dx \quad (10)$$

Wang Qi et al. [22] proposed a multiphase Chan-Vese model based on a plurality of binary level set functions and alternating convex optimization in 2010. They rewrite the multiphase Chan-Vese model based on binary level set as:

$$E(\phi) = \sum_{j=1}^{m} \int_\Omega \gamma \left|\nabla \phi_j\right| dx + \sum_{i=1}^{n} \alpha_i \int_\Omega Q_i \chi_i dx \ , \quad (11)$$

where ϕ_j is the binary level set function and the characteristic function $\chi_i(x)$ should be restated as follows:

$$\chi_i(x) = \prod_{j=1}^{m} \left[b_{i-1}^j + (-1)^{b_{i-1}^j} \phi_j \right]. \quad (12)$$

2.2 GDM, DM, ADMM or ALM and AMA for minimizing multiphase Chan-Vese model

2.2.1 Gradient descent method (GDM)

The energy functional minimization problem associated with Equation (11) can be solved by computing the evolution equation of ϕ via gradient descent flow as:

$$\begin{cases} \dfrac{\partial \phi_j}{\partial t} = \gamma \nabla \cdot \left(\dfrac{\nabla \phi_j}{\left|\nabla \phi_j\right|} \right) - \sum_{i=1}^{n} \alpha_i Q_i \dfrac{\partial \chi_i}{\partial \phi_j} \quad in \ \ \Omega \\ \dfrac{\partial \phi_j}{\partial \vec{n}_j} = 0 \quad on \ \ \partial\Omega \end{cases} . \quad (13)$$

Where the second formula of (13) is the boundary condition. But (13) includes fourth order derivatives need to be discretized using complex finite difference formulas and the integration steps of time marching depend on right hand terms heavily.

2.2.2 Dual method (DM)

In order to speed up the calculation, Dual formula of TV norm $\int_\Omega \left|\nabla \phi\right| dx = \sup_{|\vec{p}| \le 1} \int_\Omega \phi \nabla \cdot \vec{p}\, dx$ proposed by Chambolle [18] can be applied in the energy functional.

$$E(\phi,\vec{p})=\sum_{j=1}^{m}\int_{\Omega}\gamma\phi_j\nabla\cdot\vec{p}_j dx+\sum_{i=1}^{n}\alpha_i\int_{\Omega}Q_i\chi_i dx \quad (14)$$

A way of alternating optimization is used to compute ϕ_j and dual variable $\vec{p}_j$:

$$\frac{\partial\phi_j}{\partial t}=-\gamma\nabla\cdot\vec{p}_j-\sum_{i=1}^{n}\alpha_i Q_i\frac{\partial\chi_i}{\partial\phi_j} \quad (15)$$

$$\frac{\partial\vec{p}_j}{\partial t}=-\left|\nabla\phi_j\right|\vec{p}_j-\nabla\phi_j \quad (16)$$

Though dual method can speed up the processing to some degree, it can not obtain the ideal convergence rate.

2.2.3 Alternating direction method of multipliers (ADMM)

The basic idea of [23] is to use low-order variables instead of high-order variables and obtain an approximate result. The constraint $\vec{w}_j=\nabla\phi_j$ is added in energy functional.

$$E(\phi,\vec{w})=\sum_{j=1}^{m}\int_{\Omega}\gamma\left|\vec{w}_j\right|dx+\sum_{i=1}^{n}\alpha_i\int_{\Omega}Q_i\chi_i dx+\sum_{j=1}^{m}\int_{\Omega}\vec{\lambda}_j\left(\vec{w}_j-\nabla\phi_j\right)dx+\frac{\mu}{2}\sum_{j=1}^{m}\int_{\Omega}\left(\vec{w}_j-\nabla\phi_j\right)^2 dx \quad (17)$$

where $\vec{\lambda}_j$ is called the Lagrange multiplier and μ is a penalization parameter. Then the variables are optimized as:

$$\begin{cases}\sum_{i=1}^{n}\alpha_i Q_i\frac{\partial\chi_i}{\partial\phi_j}+\nabla\cdot\vec{\lambda}_j^k+\mu\nabla\cdot\left(\vec{w}_j^k-\nabla\phi_j\right)=0 \ \ in\ \Omega\\ \left(\mu\left(\nabla\phi-\vec{w}_j^k\right)-\vec{\lambda}_j^k\right)\cdot\vec{n}=0 \ \ on\ \partial\Omega\end{cases} \quad (18)$$

$$\vec{w}_j^{k+1}=max\left(\left|\nabla\phi_j^{k+1}-\frac{\vec{\lambda}_j^k}{\mu}\right|-\frac{\gamma}{\mu},0\right)\frac{\nabla\phi_j^{k+1}-\frac{\vec{\lambda}_j^k}{\mu}}{\left|\nabla\phi_j^{k+1}-\frac{\vec{\lambda}_j^k}{\mu}\right|} \quad (19)$$

2.2.4 Alternating minimization algorithm (AMA)

The computing framework of AMA is similar to ADMM. Both of them adopt the alternating direction method while AMA is simpler. The only significant difference is about the calculation of ϕ_j. Here Gradient descent method is used to ensure its convergence of iterative method. Then we can get ϕ_j^{k+1} through the following iteration:

$$\begin{cases}\frac{\partial\phi_j}{\partial t}=-\nabla\cdot\vec{\lambda}_j^k-\sum_{i=1}^{n}\alpha_i Q_i\frac{\partial\chi_i}{\partial\phi_j} \quad in\ \ \Omega\\ \frac{\partial\phi_j}{\partial\vec{n}_j}=0 \quad on\ \ \partial\Omega\end{cases}. \quad (20)$$

3 Our proposed methods

Though the GDM, DM, ADMM or ALM and AMA have been successfully extended to the multiphase Chan-Vese model, they cannot get results with exact values as well as in rapid speed. Through introducing a 'restart rule'- i.e. the acceleration parameters are reset when certain conditions are met. So both expensive computing process and complex item appearance in the evolution equations are able to be avoided by our proposed fast method.

3.1 Fast alternating direction method of multipliers (FastADMM)

First the basic idea of using FastADMM is proposed. We introduce auxiliary variables $\vec{v}_j\ (j=1,2,...,m)$ to replace the $\nabla\phi_j$, so the high-order variables can be simplified by low-order variables. Then the variables will be optimized respectively. Following conventions, the symbol $'\rightarrow'$ is going to be used to denote vector functions. The energy functional is rewritten as follows:

$$E\left(c,\phi,\vec{w},\vec{\lambda}^{\wedge}\right)=\sum_{j=1}^{m}\int_{\Omega}\gamma\left|\vec{w}_j\right|dx+\sum_{i=1}^{n}\alpha_i\int_{\Omega}Q_i\chi_i dx+\sum_{j=1}^{m}\int_{\Omega}\vec{\lambda}_j^{\wedge}\left(\vec{w}_j-\nabla\phi_j\right)dx+\frac{\mu}{2}\sum_{j=1}^{m}\int_{\Omega}\left(\vec{w}_j-\nabla\phi_j\right)^2 dx \quad (21)$$

Where $\vec{\lambda}_j^{\wedge}$ should be carefully calculated by the intermediate variable $\vec{\lambda}_j$. Please notice that ϕ_j should be updated by $\vec{v}_j$. If the value of $\vec{w}_j$ is obtained by solving the Euler-Lagrange equation, $\vec{v}_j$ will be updated by $\vec{w}_j$. Detailed implementation of FastADMM is shown in Algorithm 1.

Algorithm 1: FastADMM for multiphase Chan-Vese model

1.Initialization: $\vec{w}_j^0=\vec{v}_j^0$, $\vec{\lambda}_j^0=\vec{\lambda}_j^{\wedge 0}$, $(j=1,2,...,m)$, $\alpha^0=1$, $\mu>0$.

2. For $k\geq 1$, solve the following problems alternatively:

2.1. Subproblem 1 about c_i^{k+1}, $i=1,2,...n$:

$$c_i^{k+1}=argmin\left\{\varepsilon_1(c_i)=E\left(c_i,\phi_j^k,\vec{w}_j^k;\vec{\lambda}_j^k\right)\right\}. \quad (22)$$

2.2. Subproblem 2 about ϕ_j^{k+1}:

$$\tilde{\phi}_j^{k+1}=argmin\left\{\varepsilon_2(\phi_j)=E\left(c_i^{k+1},\phi_j,\vec{v}_j^k;\vec{\lambda}_j^{\wedge k}\right)\right\}, \quad (23)$$

$$\phi_j^{k+1}=\prod\nolimits_{\Omega}\left(\tilde{\phi}_j^{k+1}\right). \quad (24)$$

2.3. Subproblem 3 about $\vec{w}_j^{k+1}$:

$$\vec{w}_j^{k+1}=argmin\left\{\varepsilon_3(\vec{w}_j)=E\left(c_i^{k+1},\phi_j^{k+1},\vec{w}_j;\vec{\lambda}_j^{\wedge k}\right)\right\}, \quad (25)$$

2.4. Update Lagrange multiplier $\vec{\lambda}$:

$$\vec{\lambda}_j^{k+1} = \vec{\lambda}_j^{\wedge k} + \mu\left(\vec{w}_j^{k+1} - \nabla\phi_j^{k+1}\right)\Big). \quad (26)$$

2.5. if $E^k > 0$, then a 'restart rule' and a relaxation factor are used to update $\vec{v}_j$ and $\vec{\lambda}_j^{\wedge}$.

$$\alpha^{k+1} = \frac{1+\sqrt{1+4\left(\alpha^k\right)^2}}{2}. \quad (27)$$

$$\vec{v}_j^{k+1} = \vec{w}_j^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}}\left(\vec{w}_j^{k+1} - \vec{w}_j^{k}\right). \quad (28)$$

$$\vec{\lambda}_j^{\wedge k+1} = \vec{\lambda}_j^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}}\left(\vec{\lambda}_j^{k+1} - \vec{\lambda}_j^{k}\right). \quad (29)$$

else

$$\alpha^{k+1} = 1, \quad \vec{v}_j^{k+1} = \vec{w}_j^{k+1}, \quad \vec{\lambda}_j^{\wedge k+1} = \vec{\lambda}_j^{k+1}$$

2.6. ϕ^{k+1} need to be processed by threshold:

$$\phi_j^{k+1} = \begin{cases} 1 & \phi_j^{k+1} > a \\ 0 & otherwise \end{cases}. \quad (30)$$

3. The overall loop will be terminated if the stopping criterions (described in section 4) are satisfied.

The projection $\prod_{\Omega}(\cdot)$ in equation (24) is a simple truncation of ϕ_j^{k+1} to the interval $[0,1]$. For $k = 0,1,\ldots$, the minimizers of variables c_i, ϕ_j, $\vec{w}_j$ in subproblems 1-3 can be obtained by minimizing the following energy functionals:

$$\varepsilon_1(c_i) = \sum_{i=1}^{n}\alpha_i\int_{\Omega} Q_i\chi_i dx \quad (31)$$

$$\varepsilon_2(\phi_j) = \sum_{i=1}^{n}\alpha_i\int_{\Omega} Q_i\chi_i dx + \int_{\Omega}\vec{\lambda}_j^{\wedge}\left(\vec{v}_j - \nabla\phi_j\right)dx + \frac{\mu}{2}\int_{\Omega}\left(\vec{v}_j - \nabla\phi_j\right)^2 dx \quad (32)$$

$$\varepsilon_3(\vec{w}_j) = \int_{\Omega}\gamma\left|\vec{w}_j\right|dx + \int_{\Omega}\vec{\lambda}_j^{\wedge}\left(\vec{w}_j - \nabla\phi_j\right)dx + \frac{\mu}{2}\int_{\Omega}\left(\vec{w}_j - \nabla\phi_j\right)^2 dx \quad (33)$$

Thus the energy functional is simplified which can be computed by easier iterative algorithm and avoid the subproblem which occurs in ϕ_j associating with no convergence successfully. Next, (31) to (33) will be solved respectively by different iterative methods.

3.1.1 Estimations of piecewise constant parameters

We can obtain c_i^{k+1}, $i = 1,2,\ldots,n$ as in Equation (7).

3.1.2 Computing of the binary level set function

When ϕ_j is being computed, semi-implicit Gauss-Seidel iterative scheme can be used because it can ensure its fast convergence. The $(k+1)$th value of c_i^{k+1} and the kth auxiliary variable $\vec{v}_j^{k}$ should be fixed. This concept is also applied in the following paragraphs. The corresponding Euler-Lagrange equation of (32) is:

$$\mu\nabla\cdot\left(\vec{v}_j^{k} - \nabla\phi_j\right) + \nabla\cdot\vec{\lambda}_j^{\wedge k} + \sum_{i=1}^{n}\alpha_i\int_{\Omega} Q_i\frac{\partial\chi_i}{\partial\phi_j} = 0. \quad (34)$$

In experiments, we find that two or three iterative steps are enough to achieve a good minimizer of ϕ_j. It is a powerful guarantee of the energy functional minimization.

Now $\tilde{\phi}_j^{k+1}$ becomes the nonstandard binary level set function. It must be projected to [0,1] as in Equation (24):

$$\phi_j^{k+1} = Max\left(Min\left(\tilde{\phi}_j^{k+1},1\right),0\right). \quad (35)$$

3.1.3 Calculation of the auxiliary variable

The soft thresholding formula [25-27] is used in this part to calculate the variable $\vec{w}_j^{k+1}$. It is one of the most classical algorithms which has been widely used to obtain minimizers of the variables in this kind of equation. The variables c_i^{k+1} and ϕ_j^{k+1} are fixed. The calculation result is shown as :

$$\vec{w}_j^{k+1} = Max\left(\left|\nabla\phi_j^{k+1} - \frac{\vec{\lambda}_j^{\wedge k}}{\mu}\right| - \frac{\gamma}{\mu}, 0\right)\frac{\nabla\phi_j^{k+1} - \frac{\vec{\lambda}_j^{\wedge k}}{\mu}}{\left|\nabla\phi_j^{k+1} - \frac{\vec{\lambda}_j^{\wedge k}}{\mu}\right|}. \quad (36)$$

After the minimizers of subproblems 1-3 are found, Lagrange multipliers should be updated according to (26). Now let us introduce the 'restart rule' which is able to guarantee convergence for weakly-convex problems. If the judging criteria E^k is greater than 0, the parameter sequence $\{\alpha^k\}$ is used to over-relax the sequence of iteration and help update $\vec{v}_j$ and $\vec{\lambda}_j^{\wedge}$. The definition of E^k is described in [19]. At the end of the implementation, we can work out the threshold a for binarization of ϕ^{k+1} on the basis of the histogram of its result (as described in (30)). In section 4, the stopping criterions which can be used to terminate the overall loop are presented.

3.2 Fast alternating minimization algorithm (FastAMA)

The basic idea of using FastAMA is similar to FastADMM. In this algorithm, only one intermediate variable

$\vec{\lambda}_j^{\wedge}$ is used to accelerate the convergence. Detailed implementation of FastAMA is shown in Algorithm 2.

Algorithm 2: FastAMA for multiphase Chan-Vese model

1. Initialization: $\vec{w}_j^0 = \vec{v}_j^0$, $\vec{\lambda}_j^0 = \vec{\lambda}_j^{\wedge 0}$, $(j = 1,2,...,m)$, $\alpha^0 = 1$, $\mu > 0$.
2. For $k \geq 1$, solve the following problems alternatively:
   2.1. Subproblem 1 about c_i^{k+1}, $i = 1,2,...n$:

$$c_i^{k+1} = argmin\left\{\varepsilon_1(c_i) = E\left(c_i, \phi_j^k, \vec{w}_j^k; \vec{\lambda}_j^k\right)\right\}. \quad (37)$$

   2.2. Subproblem 2 about ϕ_j^{k+1}:

$$\tilde{\phi}_j^{k+1} = argmin\left\{\varepsilon_2(\phi_j) = E\left(c_i^{k+1}, \phi_j, \vec{w}_j^k; \vec{\lambda}_j^{\wedge k}\right)\right\}, \quad (38)$$

$$\phi_j^{k+1} = \prod\nolimits_{\Omega}\left(\tilde{\phi}_j^{k+1}\right). \quad (39)$$

   2.3. Subproblem 3 about $\vec{w}_j^{k+1}$:

$$\vec{w}_j^{k+1} = argmin\left\{\varepsilon_3(\vec{w}_j) = E\left(c_i^{k+1}, \phi_j^{k+1}, \vec{w}_j; \vec{\lambda}_j^{\wedge k}\right)\right\}. \quad (40)$$

   2.4. Update Lagrange multiplier $\vec{\lambda}$:

$$\vec{\lambda}_j^{k+1} = \vec{\lambda}_j^{\wedge k} + \mu\left(\vec{w}_j^{k+1} - \nabla\phi_j^{k+1})\right). \quad (41)$$

   2.5. Update the relaxation parameter:

$$\alpha^{k+1} = \frac{1 + \sqrt{1 + 4\left(\alpha^k\right)^2}}{2}. \quad (42)$$

   2.6. Update the intermediate variable $\vec{\lambda}_j^{\wedge}$:

$$\vec{\lambda}_j^{\wedge k+1} = \vec{\lambda}_j^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}}\left(\vec{\lambda}_j^{k+1} - \vec{\lambda}_j^k\right). \quad (43)$$

   2.7. ϕ^{k+1} need to be processed by threshold:

$$\phi_j^{k+1} = \begin{cases} 1 & \phi_j^{k+1} > a \\ 0 & otherwise \end{cases}. \quad (44)$$

3. The overall loop will be terminated if the stopping criterions (described in section 4) are satisfied.

The minimizers of variables c_i, ϕ_j, $\vec{w}_j$ in subproblems 1-3 can be obtained by minimizing the following functionals:

$$\varepsilon_1(c_i) = \sum_{i=1}^{n} \alpha_i \int_{\Omega} Q_i \chi_i dx \ , \quad (45)$$

$$\varepsilon_2(\phi_j) = \sum_{i=1}^{n} \alpha_i \int_{\Omega} Q_i \chi_i dx + \int_{\Omega} \vec{\lambda}_j^{\wedge}\left(\vec{v}_j - \nabla\phi_j\right)dx \ , \quad (46)$$

$$\varepsilon_3(\vec{w}_j) = \int_{\Omega} \gamma\left|\vec{w}_j\right|dx + \int_{\Omega} \vec{\lambda}_j^{\wedge}\left(\vec{w}_j - \nabla\phi_j\right)dx + \frac{\mu}{2}\int_{\Omega}\left(\vec{w}_j - \nabla\phi_j\right)^2 dx \ . \quad (47)$$

Where equation (45) can be minimized as presented in (7). ϕ_j^{k+1} of (46) can be obtained as in equation (20), please notice that $\vec{\lambda}_j$ should be replaced by the intermediate variable $\vec{\lambda}_j^{\wedge}$. And $\vec{w}_j^{k+1}$ of (47) can be obtained as in (36). After the minimizers of subproblems 1-3 are found, Lagrange multipliers should be updated according to (41). The parameter sequence $\left\{\alpha^k\right\}$ is used to over-relax the sequence of iteration and help update $\vec{\lambda}_j^{\wedge}$. At the end of the implementation, binarization of ϕ^{k+1} is required as well. We use the same stopping criterions to terminate the overall loop as presented in section 4.

4 Numerical experiments

In this section, the numerical results of our proposed methods are applied on some real cases and they will be compared with different methods (GDM, DM, ADMM or ALM, AMA) to demonstrate the effectiveness and efficiency of our methods. All the experiments are operated on the same platform (Matlab7.8) on a PC (Intel (R), CPU 2.60GHz). The same initial contours and initiations of variables for all the methods in each experiment are used in order to have a relatively neutral criterion for comparison. To clarify this, the initial values of variables are shown as follows:

GDM: $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$,

DM: $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$, $\vec{p}_j^0 = 0$,

ADMM or ALM: $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$, $\vec{w}_j^0 = 0$, $\vec{\lambda}_j^0 = 0$,

AMA: $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$, $\vec{w}_j^0 = 0$, $\vec{\lambda}_j^0 = 0$,

FastADMM: $\alpha^0 = 1$, $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$, $\vec{w}_j^0 = 0$, $\vec{\lambda}_j^0 = 0$, $\vec{v}_j^0 = 0$, $\vec{\lambda}_j^{\wedge 0} = 0$

FastAMA: $\alpha^0 = 1$, $c_i^0 = 0$, $\phi_j^0 \in \{0,1\}$, $\vec{w}_j^0 = 0$, $\vec{\lambda}_j^0 = 0$.

As described in [28], the iterations need to be terminated when the following criterions are satisfied. In this paper, the same stopping criterions can be used in the proposed two methods.

〈i〉 We need to monitor the constraints errors in iterations:

$$R_{\vec{w}_j}^k = \frac{\left\|R_{\vec{w}_j}^k\right\|_{L^1}}{\left\|R_{\vec{w}_j}^0\right\|_{L^1}} \ \left(j = 1,2,...,m\right), \quad (48)$$

with

$$R_{\vec{w}_j}^k = \vec{w}_j^k - \nabla\phi_j^k, \quad (49)$$

where $\left\|\cdot\right\|_{L^1}$ denotes the L^1 norm on image domain Ω. If $R_{\vec{w}_j}^k < \varepsilon$ (ε is a small enough parameter), iteration of outer repeat k will be stopped. These good numerical indicators are also used to determinate the values of μ, which can be the basis of penalty parameter adjustment.

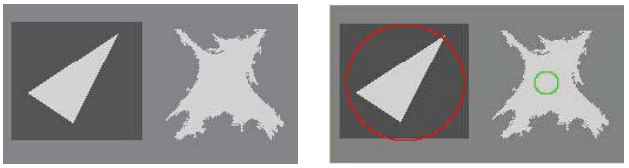
〈ii〉 In iterations, the relative errors of Lagrange multipliers and the solution ϕ_j^k should be noticed. They should reduce to a sufficiently small level:

$$L_{\vec{\lambda}_j}^k = \frac{\left\|\vec{\lambda}_j^k - \vec{\lambda}_j^{k-1}\right\|_{L^1}}{\left\|\vec{\lambda}_j^{k-1}\right\|_{L^1}} \quad (j = 1,2,...,m), \tag{50}$$

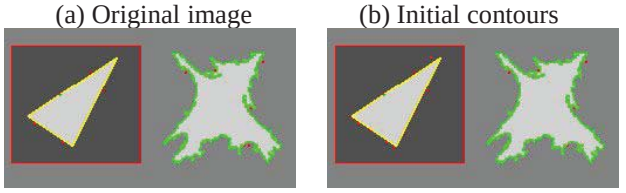
$$\frac{\left\|\phi_j^k - \phi_j^{k-1}\right\|_{L^1}}{\left\|\phi_j^{k-1}\right\|_{L^1}}. \tag{51}$$

〈iii〉 The convergence of energy functional $E(\phi)$ need to be guaranteed. $\frac{\left|E(\phi^{k+1}) - E(\phi^k)\right|}{\left|E(\phi^k)\right|} \le \varepsilon$ should be satisfied.

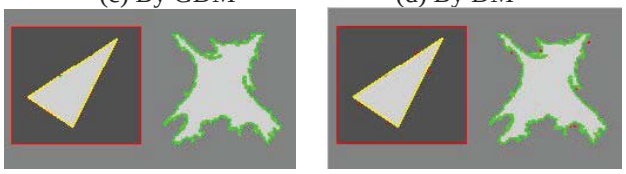
Experiment 1. Synthetic image of size 250×136 is used as the test image. In this experiment, two binary level set functions are used to detect three different subdomains (*m*=2). In Fig.1, some results of GDM, DM, ADMM or ALM, AMA and proposed two methods are firstly presented respectively so that we can make visual comparisons with the segmented images. Fig. 1(a) shows the original image. The initial contours are shown in Fig.1(b). Fig.1 (c)-(f) shows the segmentation results of GDM, DM, ADMM or ALM and AMA respectively. The segmented images shown in Fig. 1 (g) and (h) are from our proposed two methods. The parameters used in FastADMM for Fig.1 (g) are given as follows: $\gamma = 0.5,\ \mu = 0.4,\ \alpha_1 = 2,\ \alpha_2 = 1,\ \alpha_3 = 1,\ \alpha_4 = 2$. And the parameters used in FastAMA for Fig.1 (h) are : $t = 0.1,\ \gamma = 5,\ \mu = 0.4,\ \alpha_1 = 3,\ \alpha_2 = 2,\ \alpha_3 = 1,\ \alpha_4 = 2$



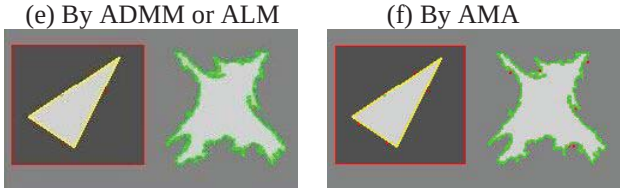
(a) Original image (b) Initial contours



(c) By GDM (d) By DM



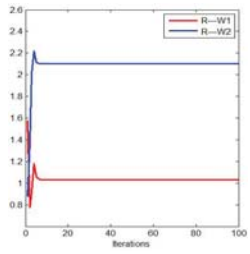
(e) By ADMM or ALM (f) By AMA



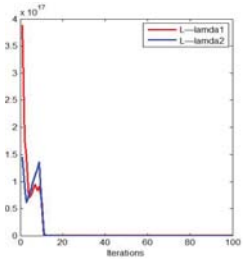
(g) By FastADMM (h) By FastAMA

Fig.1. The effects of GDM, DM, ADMM or ALM, AMA and proposed two methods. The first row: original image and the initial contours. The second and third row: segmentation results of GDM, DM, ADMM or ALM, AMA. The last row : results of our proposed two methods.

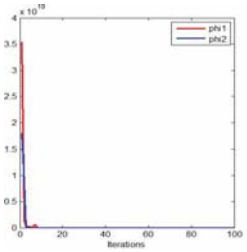
From left to right, we illustrate relative residuals (48), relative errors of Lagrange multipliers (50), relative error of ϕ_j^k (51) and energy curve along the outer repeat k in Fig.2. The graphs come from Fig. 1 (g) and (h) respectively. It can be observed that the algorithm has converged long before 100 iterations. They also give important information about how to choose penalty parameter μ . In order to guarantee convergence as well as the speed of convergence, the constraint errors $R_{\vec{w}_j}^k$ should converge to zero with nearly the same speed. If $R_{\vec{w}_j}^k$ goes to zero quicker than others, then decrease μ . $R_{\vec{w}_j}^k$ will converge to zero with the same speed as the iteration proceeds and the energy will decrease to a steady constant value when μ are chosen properly. This experiment points out that the selection of parameter γ has no obvious effect on the results.



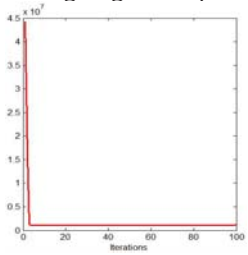
(i) Relative residuals



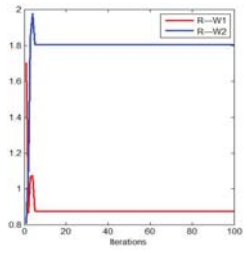
(j) Relative errors of Lagrange multipliers



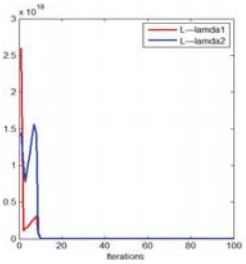
(k) Relative error of ϕ_j^k



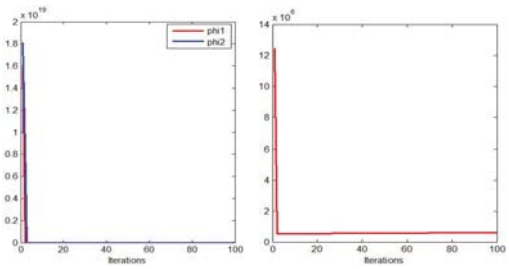
(l) The energy



(m) Relative residuals



(n) Relative errors of Lagrange multipliers



(o) Relative error of ϕ_j^k (p) The energy

Fig.2. The plots of parametric errors and the energy curve. (i)-(l) are obtained by FastADMM from Fig.1 (g). (m)-(p) are obtained by FastAMA from Fig.1 (h).

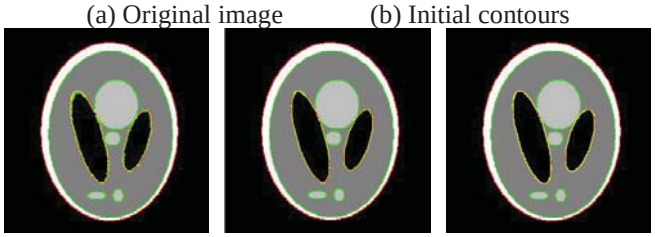
In the aspect of algorithm efficiency, iterations and computational time of methods presented in this experiment are given. It is easy to see that FastADMM and FastAMA have the faster convergence rate.

TABLE 1.
Comparisons of iterations and computational time

Approaches	Iterations	Time (sec)
Fig. 1-(c): GDM	36	0.198
Fig. 1- (d): DM	20	0.163
Fig. 1- (e): ADMM	9	0.094
Fig. 1- (f): AMA	8	0.112
Fig. 1- (g): FastADMM	4	0.085
Fig. 1- (h): FastAMA	4	0.081

Experiment 2. In this experiment, our methods will be compared with GDM, DM ADMM and AMA by using them on a image of size 256×256. The original image is presented in Fig. 3(a) (m=2) and the same initial contours are used in Fig. 2(b). Three parts contained by the methods mentioned above are given in Fig. 3(c)-(h). We can see all these methods can obtain almost the same segmentaion effects. Parameters used in FastADMM and FastAMA are given: $\gamma = 0.5,\ \mu = 0.4,$ $\alpha_1 = 2,\ \alpha_2 = 1,\ \alpha_3 = 1,\ \alpha_4 = 2.$ t=0.1, γ=5, μ=0.4, α_1=3, α_2=2, α_3=1, α_4=2.

(a) Original image (b) Initial contours

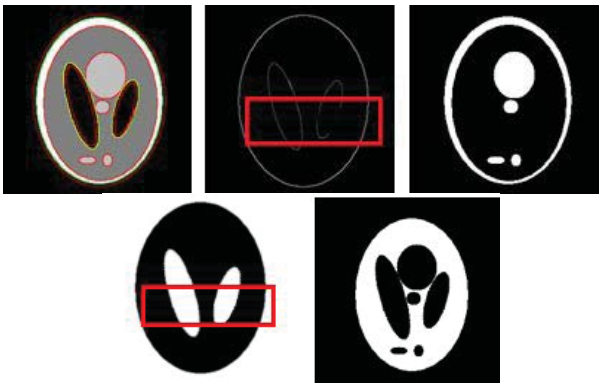


(c) By GDM (d) By DM (e) By ADMM

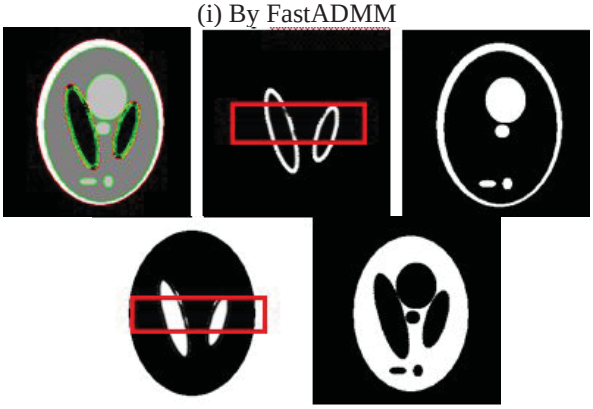
(f) By AMA (g) By FastADMM (h) By FastAMA

Fig.3. The effects of GDM, DM, ADMM, AMA and proposed two methods. The first row: original image and the initial contours. The second and third row: segmentation results of these six methods.

Here a threshold method should be used to realize the binaryzation of ϕ_j^{k+1}. It is an important way to help find the accurate results. Non-threshold solutions of the proposed methods are shown as follows. It can be observed that non-threshold often results in fuzzy edges (red rectangles).



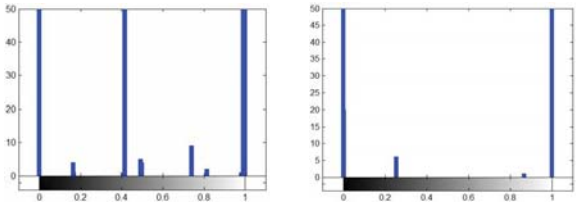
(i) By FastADMM



(j) By FastADMM

Fig.4. Non-threshold solutions of the proposed methods. (i) comes from FastADMM. (j) comes from FastAMA.

Next, the histograms of non-threshold solutions from FastADMM are given in Fig. 5. It gives us a good way to choose the threshold of ϕ_j^{k+1}. In this experiment, we find the threshold $a = 0.5$ could be applicable.



(k) histogram of ϕ_1^{k+1} (non-threshold) (l) histogram of ϕ_2^{k+1} (non-threshold)

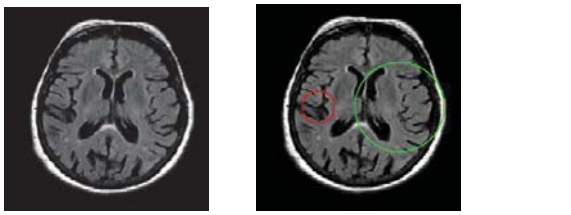
Fig.5. Histograms of ϕ_i^{k+1} (non-threshold).

From Table 2, it is obvious that the total computational cost required by our methods is much less than other four methods from the comparison.

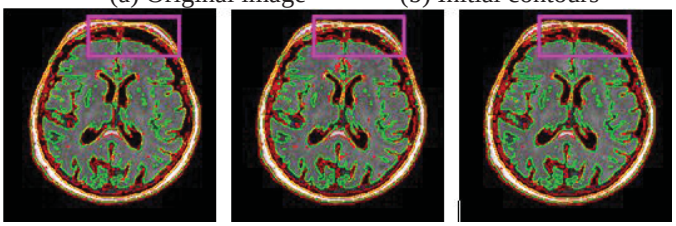
TABLE 2.
Comparisons of iterations and computational time

Approaches	Iterations	Time (sec)
Fig. 3-(c): GDM	29	0.298
Fig. 3- (d): DM	18	0.224
Fig. 3- (e): ADMM	7	0.162
Fig. 3- (f): AMA	6	0.158
Fig. 3- (g): FastADMM	4	0.126
Fig. 3- (h): FastAMA	3	0.132

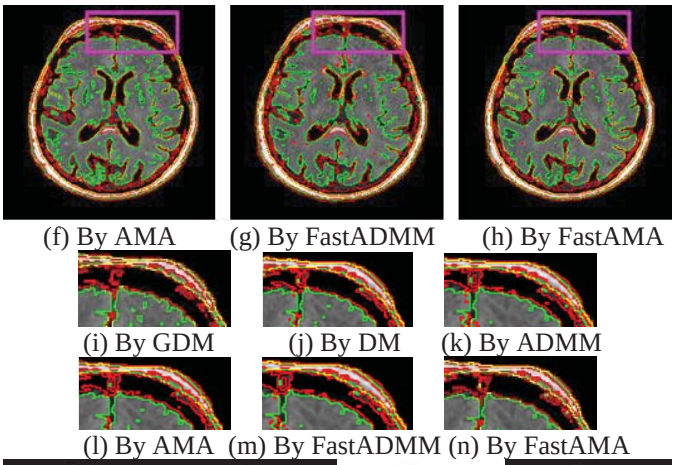
Experiment 3. The results of all these methods are shown on a brain magnetic resonance image (MRI). From the original image of size 256×256 in Fig. 6(a), there are four parts need to be segmented. Fig. 6(b) shows the initial contours. The segmentation results from different methods are given in Figs. 6(c)-(h) and local enlarged results of (c)-(h) are shown in Figs. 6(i)-(n). Those subdomains separated from Fig. 6(g) and (h) with proposed methods are respectively presented in Figs. 6(o)-(p). The parameters used in FastADMM for Fig. 6 (g) are: $\gamma = 0.5,\ \mu = 0.4,\ \alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 2,\ \alpha_4 = 1$. And the parameters used in FastAMA for Fig.1 (h) are : $t = 0.1,\ \gamma = 0.5,\ \mu = 0.4,\ \alpha_1 = 3,\ \alpha_2 = 2,\ \alpha_3 = 1,\ \alpha_4 = 2$.



(a) Original image (b) Initial contours



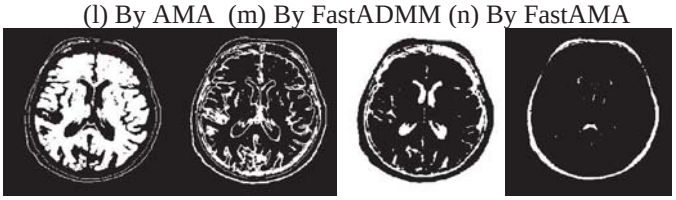
(c) By GDM (d) By DM (e) By ADMM



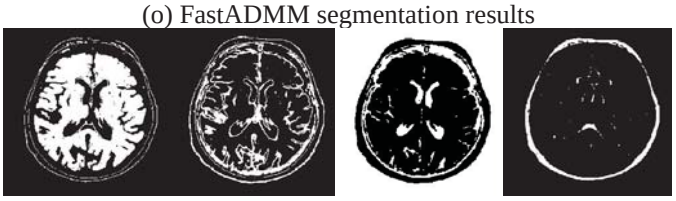
(f) By AMA (g) By FastADMM (h) By FastAMA

(i) By GDM (j) By DM (k) By ADMM

(l) By AMA (m) By FastADMM (n) By FastAMA



(o) FastADMM segmentation results



(p) FastAMA segmentation results

Fig.6. The comparison between other methods and our methods on a MRI. The first row: original image and the initial contours. The second and Third row: results of other methods and our methods. The fourth and fifth row: zoomed small subregions (purple rectangles). The last two row: four different phases of (g) and (h) obtained by proposed methods.

In Table 3, comparisons of iterations and computational time using different methods are given.

TABLE 3.
Comparisons of iterations and computational time

Methods	Iterations	Time (sec)
Fig. 6-(c): GDM	25	2.58
Fig. 6-(d): DM	14	1.84
Fig. 6-(e): ADMM	9	0.69
Fig. 6-(f):AMA	9	0.86
Fig. 6-(g) and (o): FastADMM	5	0.41
Fig. 6-(h) and (p): FastAMA	4	0.33

5 Conclusions

In this paper, by using the relevant concepts of Nesterov's accelerated algorithm, convex optimization and multiphase Chan-Vese model, we propose FastADMM and FastAMA for multiphase image segmentation. Our proposed accelerated methods have been validated by several numerical experiments. The comparison of results obtained by some

other approaches and our proposed approach indicate that our approach owns good enough effects and it is a good way to efficiently minimize the difficult functional. Our method can also be applied into surface segmentation, 3D reconstruction and image denoising models etc. in the future work. It is supposed to yield shorter runtime than the traditional methods, while the quality of results is identical.

Acknowledgements

The work has been partially supported by the National Natural Science Foundation of China (nos.61305045, 61170106, and 61303079).

References

[1] Kass M, Witkin A, Terzopoulos D, Snakes: active contour models. Int. J. Comput. Vis. 4(1), 321–331, 1987.

[2] Aubert G, Barlaud M, Faugeras O, Jehan-Besson S, Image segmentation using active contours: calculus of variations or shape gradient. SIAM J. Appl. Math. 63(6), 2128–2154, 2003.

[3] Jinming D, Zhenkuan P, Xiangfeng Yin, Weibo Wei, Guodong Wang, Some fast projection methods based on Chan-Vese model for image segmentation. EURASIP Journal on Image and Video Processing. (10): 1687-5281, 2014.

[4] Mumford D, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems[J]. Communications on pure and applied mathematics, 42(5): 577-685, 1989.

[5] Chan T F, Vese L A. Active contours without edges[J]. Image processing, IEEE transactions on, 10(2): 266-277, 2001.

[6] Potts R B. Some generalized order-disorder transformations[C]//Proceedings of the Cambridge Philosophical Society. 48(1): 106-109, 1952.

[7] Gilles Celeux, Florence Forbes, Nathalie Peyrard. EM-based image segmentation using Potts models with external field. [Research Report] RR-4456, 2002.

[8] Zhao H K, Chan T F, Merriman B, Osher S. A variational level set approach to multiphase motion [J]. Journal of Computational Physics, 127: 179-195, 1996.

[9] Luminita A. Vese, Tony F. Chan, "A multiphase level set framework for image segmentation using the mumford and shah model", International Journal of Computer Vision, vol.50, no.3, pp.271-293, 2002.

[10] R. Glowinski and A. Marrocco Inf. Rech. Oper., vol. R-2, pp. 41–76, 1975.

[11] Paul T. Applications of splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim, 29:119-138, 1991.

[12] T. Goldstein and S. Osher, "The split bregman method for '1regularized problems," UCLA CAM Report 08-29, 2008.

[13] T. Goldstein, X. Bresson, and S. Osher, "Geometric applications of the split bregman method: Segmentation and surface reconstruction," J. Sci. Comput., vol. 45, pp. 272–293, October 2010.

[14] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o(1/k^2).," Soviet Math. Dokl., vol. 27, pp. 372–376, 1983.

[15]Lie J, Lysaker M, Tai X C. A binary level set model and some applications to Mumford-Shah image segmentation [J]. IEEE Transactions on Image Processing, 15(5): 1171-1181, 2006.

[16]X. Bresson, S. Esedoglu, P. Vandergheynst, et al. Fast global minimization of the active contour/snake model. Journal of Mathematical Imaging and Vision, 28(2):151-167, 2007.

[17]L. I. Rudin, S. Osher, E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1): 259-268, 1992.

[18]A. Chambolle. An algorithm for total variation minimization and applications. Journal of Mathematical Imaging and Vision, 20(1-2): 89-97, 2004.

[19]T. Goldstein, B. O'Donoghue, S. Setzer and R. Baraniuk, Fast alternating direction optimization methods, SIAM Journal on Imaging Sciences, 7(3):1588-1623, 2014.

[20] W. Zhu, X.-C. Tai, and T. F. Chan, Image Segmentation Using Euler's Elastica as the Regularization, Journal of Scientific Computing, 57(2):414-438, 2013.

[21] Stanley Osher, James A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation", Journal of Computational Physics, vol.79, no.1, pp.12-49, 1988.

[22] Qi Wang, Zhenkuan Pan, Weibo Wei, "Split-Bregman method and dual method for multiphase image segmentation", Journal of Computer-Aided Design & Computer Graphics, vol.22, no.9,pp.1561-1569, 2010.

[23]Lie J, Lysaker M, Tai X C. A variant of the level set method and applications to image segmentation[J]. Mathematics of computation, 75(255): 1155-1174, 2006.

[24]Wu C, Zhang J, Tai X C. Augmented Lagrangian method for total variation restoration with non-quadratic fidelity [J]. Inverse Problems and Imaging, 5: 237-261, 2011.

[25]P. K. Sahoo, S. Soltani, AND A. K. C. Wong, A survey of thresholding techniques, Computer vision, graphics, and image processing, (41):233-260, 1988.

[26] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences, 2(1):183-202, 2009.

[27]J. Yang, W. Yin, Y. Zhang, Y. Wang, A Fast Algorithm for Edge-Preserving Variational Multichannel Image Restoration, SIAM Journal on Imaging Sciences 2(2): 569-592, 2009.

[28]X.-C. Tai, Fast numerical schemes related to curvature minimization: a brief and elementary review, UCLA CAM Report 14-40, May 2014.

SESSION

WAVELET TRANSFORMATION, ANALYSIS, APPLICATIONS + WATERMARKING METHODS, SECURITY, PRIVACY, AND ENCRYPTION METHODS

Chair(s)

TBA

Image Blur Detection with 2D Haar Wavelet Transform and Its Effect on Skewed Barcode Scanning

Vladimir Kulyukin
Department of Computer Science
Utah State University
Logan, UT, USA
vladimir.kulyukin@usu.edu

Sarat Andhavarapu
Department of Computer Science
Utah State University
Logan, UT, USA
sarat.andhavarapu@aggiemail.usu.edu

***Abstract*—An algorithm is presented for image blur detection with the 2D Haar Wavelet transform (2D HWT). The algorithm classifies an image as blurred or sharp by splitting it into *N* x *N* tiles, applying several iterations of the 2D HWT to each tile, and grouping horizontally, vertically, and diagonally connected tiles with pronounced changes into tile clusters. Images with large tile clusters are classified as sharp. Images with small tile clusters are classified as blurred. If need be, the blur extent can be estimated as the ratio of the total area of the connected tile clusters and the area of the image. When evaluated on a sample of five hundred images, the algorithm performed on par or better than two other blur detection algorithms found in the literature. The effect of blur detection on skewed barcode scanning is investigated by integrating the presented blur detection algorithm into a skewed barcode scanning algorithm. The experimental results indicate that blur detection had a positive effect on skewed barcode scanning rates.**

Keywords—computer vision; image blur detection; Haar wavelets, 2D Haar wavelet transform, barcode scanning

I. Introduction

In our previous research [1, 2], we developed an algorithm for in-place vision-based skewed barcode scanning with relaxed pitch, roll, and yaw camera alignment constraints. The skewed barcode scanning experiments were conducted on a set of 506 video recordings of common grocery products. Our experiments showed that the scanning results were substantially higher on sharp images than on blurred ones. A limitation of that algorithm was that it did not filter the blurred frames out of the barcode localization and scanning process.

The same limitation was experimentally discovered in another algorithm that we developed for mobile vision-based localization of skewed nutrition labels (NLs) on grocery packages that maximizes specificity, i.e., the percentage of true negative matches out of all possible negative matches [3]. The NL localization algorithm works on frames captured from the smartphone camera's video stream and localizes NLs skewed up to 35-40 degrees in either direction from the vertical axis of the captured frame.

The NL localization algorithm uses three image processing methods: edge detection, line detection, and corner detection. We experimentally discovered that the majority of false negative matches were caused by blurred images. Both the Canny edge detector [4] and dilate-erode corner detector [5] used in the algorithm require rapid and contrasting changes to identify key points and lines of interest. These data cannot be readily retrieved from blurred images, which results in run-time barcode scanning and NL localization failures. Consequently, effective image blur detection methods will likely improve both skewed barcode scanning and NL localization rates.

Toward this end, in this paper, an algorithm is presented for image blur detection based on the 2D HWT [6]. The algorithm classifies an image as blurred or sharp by splitting it into *N* x *N* tiles, applying several iterations of the 2D HWT to each tile, and grouping the horizontally, vertically, and diagonally connected tiles with pronounced changes into clusters. Images with large clusters are classified as sharp whereas images with small tile clusters are classified as blurred. If need be, the blur extent can be estimated as the ratio of the total area of the connected tile clusters and the area of the image. The effect of blur detection on skewed barcode scanning is investigated by integrating the blur detection algorithm into our in-place vision-based skewed barcode scanning with relaxed pitch, roll, and yaw camera alignment constraints [1, 2]. The experimental results indicate that blur detection improves skewed barcode scanning rates.

The remainder of our paper is organized as follows. In Section II, we present and analyze related work. In Section III, we outline the details of our image blur detection algorithm. In Section IV, we present our experiments with the blur detection algorithm and experimentally compare the algorithm's performance with two other blur detection algorithms found in the literature [7, 8]. In Section V, the results of the experiments are discussed. Section VI summarizes our findings, presents our conclusions, and outlines some research venues we would like to pursue in the future.

II. Related Work

A. Blur Detection

Mallat and Hwang [9] mathematically prove that signals carry information via irregular structures and singularities. In particular, they show that the local maxima of the wavelet transform detect the locations of irregularities. For example,

the 2D wavelet transform maxima indicate the locations of edges in images. The Fourier analysis [10], which has been traditionally used in physics and mathematics to investigate irregularities, is not always suitable to detecting the spatial distribution of such irregularities.

According to Tong et al. [7], image blur detection methods can be broadly classified as direct or indirect. Indirect methods characterize image blur as a linear function $I_B = B \cdot I_O + N$, where I_O is the original image, B is an unknown image blur function, N is a noise function, and I_B is the resulting image after the introduction of the blur and noise.

Indirect methods consider B unknown and use various techniques to estimate it. Rooms et al. [11] propose a wavelet-based method to estimate the blur of an image by looking at the sharpness of the sharpest edges in the image. The Lipschitz exponents [12] are computed for the sharpest edges and a relation between the variance of a Gaussian point spread function and the magnitude of the Lipschitz exponent is shown to be dependent on the blur present in the image and not on the image contents.

Venkatakrishnan et al. [13] show that the wavelet transform modulus maxima (WTMM) detect all the singularities of a function and describe strategies to measure their regularity and propose an algorithm for characterizing singularities of irregular signals. The researchers present a method for measuring the Lipschitz exponents that uses the area between the straight line satisfying specific properties and the curve of the WTMM in a finite scale interval in the log-log plot of scales versus WTMM as the objective function.

Pavlovic and Tekalp [14] propose a formulation of the maximum likelihood (ML) blur identification based on parametric modeling of the blur in the continuous spatial coordinates. Unlike ML blur identification methods based on discrete spatial domain blur models, their method finds the ML estimate of the extent and the parameters of arbitrary point spread functions that admit a closed form parametric description in the continuous coordinates. Experiments show significant results for the cases of 1D uniform motion blur, 2D out-of-focus blur, and 2D truncated Gaussian blur at different signal-to-noise ratios.

Panchapakesan et al. [15] present an indirect method for image blur identification from vector quantizer encoder distortion. The method takes a set of training images from all candidate blur functions. These sets are used to train vector quantizer encoders. The a-priori unknown blur function is identified from a blurred image by choosing among the candidate vector quantizer encoders the encoder with the lowest distortion. The researchers investigated two training methods: the generalized Lloyd algorithm and a non-iterative discrete cosine transform (DCT)-based approach.

Direct methods estimate blur extent on the basis of some distinctive features directly found in images such as edges, corners, or discrete cosine transform (DCT) coefficients. Marichal et al. [16] estimate image blur based on the histogram computation of non-zero DCT coefficients computed from MPEG or JPEG compressed images. The proposed method takes into account the DCT information from the entire image. A key assumption is that any edge type will likely cross some 8 x 8 blocks at least once in the image. The camera and motion blur is estimated through the globalization among all DCT blocks.

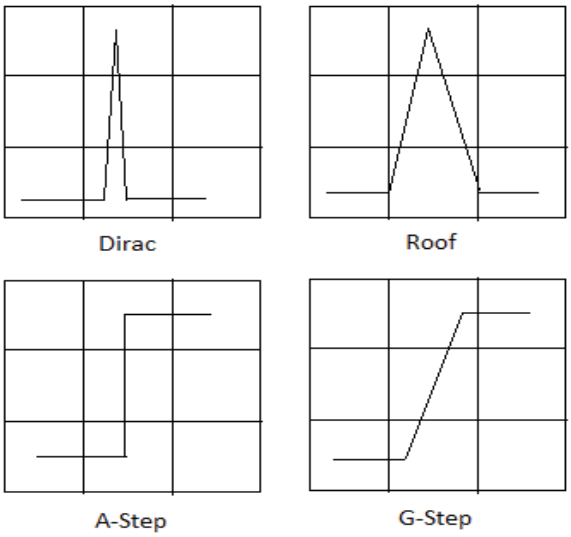


Figure 1. Edge classification

Tong et al. [7] propose a direct method similar to the one proposed in this paper in that it also uses the 2D Haar Wavelet Transform (2D HWT). Their method is based on the assumption that the introduction of blur has different effects on the four main types of edges shown in Figure 1: Dirac, A-Step, G-Step, and Roof. It is claimed that in blurred images the Dirac and A-Step edges disappear while G-Step and Roof edges lose their sharpness. The method classifies an image as blurred on the basis of the presence or absence of Dirac and A-Step edges and estimates the blur extent as the percentage of G-Step and Roof edges present in the image.

The algorithm presented in this paper is also based on the 2D HWT. However, it does not extract any explicit morphological features such as edges or corners from the image. Instead, it uses the 2D HWT to detect regions with pronounced changes and combines those regions into larger segments without explicitly computing the causes of those changes. Since the algorithm is based on the 2D HWT, the regions are square tiles whose side is an integral power of 2. This algorithm continues our investigation of vision-based barcode and nutrition label scanning on mobile phones with relaxed pitch, yaw, and roll constraints [1, 2]. As we have previously reported, most false negatives in our experiments were caused by blurred images. While newer models of smartphones will likely have improved camera stability and focus, software techniques to detect blurred images can still make vision-based skewed barcode scanning and nutrition information extraction more reliable and efficient. Since our barcode scanning and nutrition information extraction algorithms are cloud-based, eliminating blurred images from processing will likely improve the network throughput and decrease data plan consumption rates.

B. 2D Haar Transform

Our implementation of the 2D HWT is based on the approach taken in [6] where the transition from 1D Haar wavelets to 2D Haar wavelets is based on the products of basic wavelets in the first dimension with basic wavelets in the second dimension. For a pair of functions f_1 and f_2 their tensor product is defined as $(f_1 \times f_2)(x,y) = f_1(x) \cdot f_2(x)$. Two 1D basic wavelet functions are defined as follows:

$$\varphi_{[0,1[}(r) = \begin{cases} 1 \; if \; 0 \le r < 1, \\ 0 \; otherwise. \end{cases}$$

$$\psi_{[0,1[}(r) = \begin{cases} 1 \; if \; 0 \le r < \frac{1}{2}, \\ -1 \; if \; \frac{1}{2} \le r < 1, \\ 0 \quad otherwise. \end{cases}$$

The 2D Haar wavelets are defined as tensor products of $\varphi_{[0,1[}(r)$ and $\psi_{[0,1[}(r)$: $\Phi_{0,0}^{(0)}(x,y) = \left(\varphi_{[0,1[} \times \varphi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{h,(0)}(x,y) = \left(\varphi_{[0,1[} \times \psi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{v,(0)}(x,y) = \left(\psi_{[0,1[} \times \varphi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{d,(0)}(x,y) = \left(\psi_{[0,1[} \times \psi_{[0,1[}\right)(x,y)$. The superscripts *h*, *v*, and *d* indicate the correspondence of these wavelets with horizontal, vertical, and diagonal changes, respectively. The horizontal wavelets detect horizontal (left to right) changes in 2D data, the vertical wavelets detect vertical (top to bottom) changes in 2D data, and the diagonal changes detect diagonal changes in 2D data.

In practice, the basic 2D HWT is computed by applying a 1D wavelet transform of each row and then a 1D wavelet transform of each column. Suppose we have a 2 x 2 pixel image

$$\begin{bmatrix} s_{0,0} & s_{0,1} \\ s_{1,0} & s_{1,1} \end{bmatrix} = \begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix}.$$

Applying a 1D wavelet transform to each row results in the following 2 x 2 matrix:

$$\begin{bmatrix} \frac{s_{0,0}+s_{0,1}}{2} & \frac{s_{0,0}-s_{0,1}}{2} \\ \frac{s_{1,0}+s_{1,1}}{2} & \frac{s_{1,0}-s_{1,1}}{2} \end{bmatrix} = \begin{bmatrix} \frac{11+9}{2} & \frac{11-9}{2} \\ \frac{7+5}{2} & \frac{7-5}{2} \end{bmatrix} = \begin{bmatrix} 10 & 1 \\ 6 & 1 \end{bmatrix}.$$

Applying a 1D wavelet transform to each new column is fetches us the result 2 x 2 matrix:

$$\begin{bmatrix} \frac{10+6}{2} & \frac{1+1}{2} \\ \frac{10-6}{2} & \frac{1-1}{2} \end{bmatrix} = \begin{bmatrix} 8 & 1 \\ 2 & 0 \end{bmatrix}.$$

The coefficients in the result matrix obtained after the application of the 1D transform to the columns express the original data in terms of the four tensor product wavelets $\Phi_{0,0}^{(0)}(x,y)$, $\Psi_{0,0}^{h,(0)}(x,y)$, $\Psi_{0,0}^{v,(0)}(x,y)$, and $\Psi_{0,0}^{d,(0)}(x,y)$:

$$\begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix} = 8\Phi_{0,0}^{(0)}(x,y) + 1\Psi_{0,0}^{h,(0)}(x,y) + 2\Psi_{0,0}^{v,(0)}(x,y) + 0\Psi_{0,0}^{d,(0)}(x,y).$$

The value 8 in the upper-left corner is the average value of the original matrix: (11+9+7+5)/4= 8. The value 1 in the upper right-hand corner is the horizontal change in the data from the left average, (11+7)/2=9, to the right average, (9+5)/2=7, which is equal $1 \cdot \Psi_{0,0}^{h,(0)}(x,y) = 1 \cdot -2$. The value 2 in the bottom-left corner is the vertical change in the original data from the upper average, (11+9)/2=10, to the lower average, (7+5)/2=6, which is equal to $2 \cdot \Psi_{0,0}^{v,(0)}(x,y) = 2 \cdot -2 = -4$. The value 0 in the bottom-right corner is the change in the original data from the average along the first diagonal (from the top left corner to the bottom right corner), (11+5)/2=8, to the average along the second diagonal (from the top right corner to the bottom left corner), (9+7)/2=8, which is equal to $0 \cdot \Psi_{0,0}^{d,(0)}(x,y)$. The decomposition operation can be represented in terms of matrices:

$$\begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix} = 8\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + 1\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} + 2\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} + 0\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

III. Blur Detection Algorithm

The first stage of the blur detection algorithm is to find image regions that have pronounced horizontal, vertical, or diagonal changes. A captured frame is divided into *N* x *N* windows, called *tiles*, where $N = 2^k, k \in Z$. Figure 2 shows square tiles of size $N = 64$. The border pixels at the right and bottom margins are discarded when captured frames are not evenly divisible by *N*. A candidate tile must have a pronounced change along at least one of the three directions: horizontal, vertical, or diagonal. Whether a change is pronounced or not is determined through a threshold.

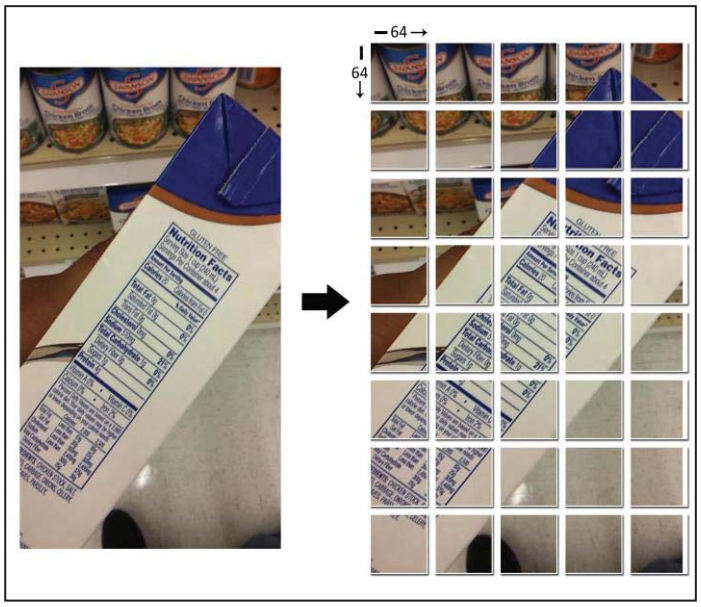


Figure 2. Tile splitting

Each tile is processed by four iterations of the 2D Haar transform. The number of iterations and the tile size are parameters and can be made either smaller or larger. The values reported in this paper were experimentally found to work well in our domain of vision-based skewed barcode scanning and nutrition information extraction [1, 2, 3].

Let *HC* be the horizontal change between the left half of the tile and the right half of the tile. Let *VC* be the vertical change between the upper half of the tile and the lower half of the tile. Let *DC* be the change between the first diagonal (top left to bottom right) and the second diagonal (top right to bottom left). If at least one of the values *HC*, *VC*, and *DC* is above the corresponding thresholds HC_θ, VC_θ, and DC_θ, respectively, the tile is marked as having a pronounced change. Figure 3 (left) shows the tiles with pronounced changes marked as squares.

Figure 3. Squre tiles with pronounced changes (left); Tile clusters found by DFS (right)

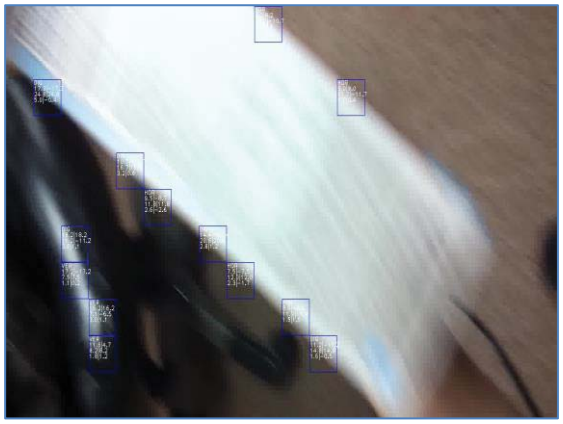


Figure 4. Tiles with pronounced changes in a blurred image

After the tiles with pronounced changes are found, the depth-first search (DFS) is used to combine them into tile clusters. The DFS starts with an unmarked tile with a pronounced horizontal, vertical, or diagonal change and proceeds by connecting to its immediate horizontal, vertical, and diagonal tile neighbors if they have pronounced changes. If such tiles are found, they are marked with the same cluster number and the search continues recursively. The search stops when no other tiles can be added to the current cluster. The algorithm continues to look for another unmarked tile to which it can apply the DFS. If no such tile is found, the algorithm terminates. Figure 3 (right) shows five DFS-found tile clusters.

Figure 5. DFS-found tile clusters

```
1. DetectBlur(Img, N, NITER, HC_θ, VC_θ, DC_θ, CSZ, A)
2. {
3.   FOR each N x N tile T in image Img {
4.       [AVRG, HC, VC, DC] = 2DHWT(T, NITER);
5.       AH=Avrg(HC); AV=Avrg(VC); AD=Avrg(DC);
6.       IF (AH > HC_θ or AV > VC_θ or AD > DC_θ)
7.          Mark T as having pronounced change;
8.   }
9.   FOR each N x N tile T in image Img {
10.       IF ( T is not in any tile cluster )
11.          Run DFS(T) to find and mark all tiles in the
12.          same cluster;
13.  }
14.  TotalArea = 0;
15.  FOR each tile cluster TC {
16.      IF ( TC's size > CSZ ) {
17.          ClusterArea = TC's size * N x N;
18.          TotalArea += ClusterArea;
19.      }
20.  }
21.  IF ( TotalArea/Area(Img) < A ) Return True;
22.  ELSE Return False;
22. }
```

Figure 6. Pseudocode of the blur detection algorithm

After the iterative applications of the DFS have found the tile clusters, two cluster-related rules are used to classify a whole image as sharp or blurred. The first rule is the percentage of the total area of the image covered by the clusters. The second rule uses the number of the tiles in each cluster to discard small clusters.

The first rule captures the intuition that in a sharp image there are many tiles with pronounced changes. The second rule discards small clusters whose size, i.e., the number of tiles in the cluster, is below a given threshold. This cluster weeding rule is currently based on a threshold of 5. In other words, any

cluster with fewer than five tiles is rejected by the second rule and does not contribute to the area of the image with pronounced changes. Thus, in Figure 3, only two clusters are left after the application of the second rule: the cluster with 18 tiles and the cluster with 6 tiles. Since both clusters are large, the image is classified as sharp by the first rule. The three singletons are discarded. On the other hand, all clusters found in the blurred image of Figure 4 are shown in Figure 5. Since all of them are small, they are discarded by the second rule.

Figure 6 gives the pseudocode of our blur detection algorithm. The first argument to the **DetectBlur** function is the image, the second argument is the size of the square tile into which the image is split, as shown in Figure 2. In our experiments presented in the next section, *N*=64. The next argument, *NITER*, is the number of iterations of the **2DHWT** function runs in each square tile in line 4. In our current implementation, *NITER*=4. Our Java source of the **2DHWT** function is available at [17]. This function returns an array of four matrices: *AVRG*, *HC*, *VC*, and *DC*. The matrix *AVRG* contains the average numbers after all iterations and the matrices *HC*, *VC*, and *DC* contain the horizontal, vertical, and diagonal wavelet coefficients, respectively.

If at least one of the averages of the *HC*, *VC*, or *DC* matrices after all *NITER* iterations is above a corresponding threshold, which is computed in lines 5 and 6, then the appropriate tile is marked as having pronounced change. In lines 9-13, the DFS is used to find all tile clusters in the image, as shown in Figure 3 (right).

In lines 14-20, the two rules described above to compute the total area occupied by the clusters greater than the value of the cluster threshold parameter *CSZ*. In lines 21-22, if the percentage of the total area occupied by the clusters with pronounced changes is smaller than the threshold value specified by the last parameter *A*, the image is classified as blurred. If need be, the algorithm can be modified to return the blur extent as the ratio of the total area occupied by the large tile clusters and the total area of the image. The smaller the ratio, the more blur exists in the image.

The outlined algorithm is based on the assumption that in blurred images square tiles with pronounced changes do not form large clusters but scatter across the image as singletons or form clusters whose combined area is small relative to the size of the image.

For example, Figure 4 is a blurred image where 64 x 64 tiles with pronounced changes are marked. Figure 5 shows the tile clusters found by the iterative applications of the DFS to the image in Figure 4. As can be seen in Figure 5, most clusters are either singletons or are of size 2. The largest cluster in the bottom left of the image is of size 4.

IV. Experiments

We took five hundred random RGB images from a set of 506 video recordings of common grocery products that we made publicly available in our previous field investigations of skewed barcode scanning [18]. The videos have a 1280 x 720 resolution and an average duration of 15 seconds. All videos were recorded on an Android 4.2.2 Galaxy Nexus smartphone in a supermarket in Logan, UT. All videos were taken by an operator who held a grocery product in one hand and a smartphone in the other. The videos covered four different categories of products: bags, boxes, bottles, and cans.

Three human volunteers were recruited to classify each of the five hundred images as blurred or sharp. An image was classified as blurred if at least two volunteers classified it as blurred. It was otherwise classified as sharp. The human evaluation resulted in 167 blurred images and 333 sharp images. These results were used as the ground truth.

We compared our algorithm with two other image blur detection algorithms frequently cited in the literature [7, 8]. Since we could not find the source code of [7] publicly available online, we implemented it ourselves in Python. Our Python source code is available at [19]. We found a MATLAB implementation of the other algorithm at [20] and used it for the experiments.

Table 1 gives the numbers of true and false positives for all three algorithms. The columns TPB and FPB give the numbers of true and false positives, respectively, for the blurred images. The columns TPS and FPS give the numbers of true and false positives, respectively, for the sharp images. The row Algo 1 gives the statistics for our algorithm implemented in Java. The rows Algo 2 and Algo 3 give the statistics for the Python implementation of [19] and the MATLAB implementation of [8], respectively.

Table 1. True and false positives

Algorithm	TPB	FPB	TPS	FPS
Algo 1	163	4	254	79
Algo 2	167	0	183	150
Algo 3	81	86	268	65

To compare the performance numbers of each algorithm with the ground truth, we used the relative difference percentage, which is a unitless measure that compares two quantities while taking into account their magnitudes. Table 2 gives the relative difference percentages computed as $|x-y|/\max(|x|,|y|)\cdot 100$, where y is the humanly estimated number of blurred or sharp images, i.e., the ground truth, and x is the number of sharp or blurred images found by a given algorithm. For example, for Algo 1, the first relative difference is computed as $|163-167|/\max(|163|,|167|)\cdot 100=2.39$, where 163 is the number of blurred images found by Algo 1 and 167 is the number of blurred images found by the human evaluators.

Table 2. Relative differences

Algorithm	Blurred	Sharp
Algo 1	2.39	23.72
Algo 2	0.00	45.05
Algo 3	51.50	19.52

Table 3. Effect of blur on barcode scanning I

Sample	Sharp	Blurred	Barcode in sharp	Barcodes in blurred
1	15	15	12	1
2	13	17	11	0
3	16	14	12	0

We investigated the effect of image blur on skewed barcode scanning. We chose three random samples of 30

images from the 500 images classified by the three human evaluators. In each sample, 15 images were classified as blurred and 15 as sharp. We integrated our blur detection algorithm into our cloud-based barcode scanning algorithm and estimated the effect of accurate image blur detection on skewed barcode scanning. Tables 3 and 4 give the results of our experiments.

In Table 3, the first column gives the sample numbers. The column *Sharp* gives the number of sharp images classified as sharp by our algorithm. The column *Blurred* gives the number of images classified as blurred by our algorithm. The column *Barcode in sharp* gives the number of barcodes correctly scanned in the images classified as sharp. The column *Barcodes in blurred* gives the number of barcodes correctly scanned in the images classified by our algorithm as blurred. Thus, in sample 1, all blurred and sharp images were classified accurately. However, in the 15 sharp images, the barcode scanner accurately scanned 12 barcodes whereas in the 15 blurred images, the barcode scanner accurately scanned only 1 barcode.

In sample 2, 13 out of 15 images were accurately classified as sharp with 2 false negatives and 17 images were classified as blurred with 2 false positives. In 11 images classified as sharp, the barcodes were accurately scanned. No barcodes were accurately scanned in the images classified as blurred.

In sample 3, 16 images were classified as sharp with 1 false positive and 14 images were classified as blurred with 1 false negative. Barcodes were successfully scanned in 12 images classified as sharp. No barcodes were scanned in the images classified as blurred.

Table 4. Effect of blur on barcode scanning II

Sample	Blurred	Sharp	Total	Gain
1	1/15	12/15	13/30	0.37
2	0/17	11/13	11/30	0.50
3	0/16	12/14	12/30	0.46

Table 4 gives the results of the effect of blur detection on skewed barcode scanning. The first column gives the numbers of the random samples. The second column records the ratio of accurately scanned barcodes in the images classified as blurred. The third column records the ratio of accurately scanned barcodes in the images classified as sharp. The fourth column gives the ratio of recognized barcodes in all images. The fifth column gives the gain measured as the difference between the ratio of the accurately recognized barcodes only in the sharp images and the ratio of the accurately recognized barcodes in all images, which estimates the effect of blur detection on barcode scanning. Thus, in sample 1, we increase the barcode scanning rate by 37 percent if we eliminate images classified as blurred from barcode scanning. In sample 2, if blurred images are eliminated from barcode scanning, we gain 50 percent in barcode scanning rates. In sample 3, the gain is 46 percent.

V. Results

In discussing the results of the experiments, we will again refer to our algorithm as *Algo* 1, to the algorithm by Tong et al. [7] as *Algo* 2, and to the algorithm by [8] as *Algo* 3. The experiments indicate (see Table 1) that, on our sample of images, in classifying images as blurred. Algo 1 performs as well as Algo 2 and outperforms Algo 3. In classifying images as sharp, Algo 1 performs as well as Algo 3 and outperforms Algo 2.

Table 2 confirms the observations recorded in Table 1. In image blur detection, there is almost no difference between Algo 1 and Algo 2 in that these algorithms do not deviate from the ground truth provided by the human evaluators on blurred images. On the other hand, Algo 3 shows a significant deviation from the ground truth on blurred images. On the other hand, in classifying images as sharp, Algo 1 and Algo 2 deviate from the ground truth by approximately 20 points while Algo 2 deviates from the ground truth by 45 points.

Tables 3 and 4 indicate that image blur detection has a pronounced positive effect on skewed barcode scanning. In all three random samples, the barcoding recognition gain was above thirty percent. While we ran these experiments only with our barcode scanning algorithm, we expect similar gains with other vision-based barcode scanning algorithms.

Our experiments indicate that direct methods provide a viable alternative to indirect methods. While indirect methods may be more accurate, they tend to be more computationally expensive due to complex matrix manipulations. Direct methods may not be as precise as their indirect counterparts. However, they compensate for it by increased efficiency, which makes them more suitable for mobile and wearable platforms.

Another observation that we would like to make is that in working with our samples of images we could not observe the blur effect on the edges observed by Tong et al. [7] in some images. The edge blur effect observed by these researchers is that the injection of blur in the images causes the Dirac and A-Step edges disappear or turn into Roof and G-Step edges, respectively and the G-Step and Roof edges to lose their sharpness.

Instead of using the 2D HWT to detect edge types, the algorithm proposed in this paper is based on the assumption that in blurred images square tiles with pronounced changes do not form larger clusters but scatter across the image as singletons or form clusters that are small in size relative to the overall size of the image.

Our approach is rooted in the research by Mallat and Hwang [9] who show that the 2D HWT can detect the location of irregularities in 2D images. In our algorithm, the 2D HWT is used to detect the location of changes via square tiles without explicitly identifying the causes of the detected changes, e.g., edges or corners.

VI. Summary

We have presented an algorithm for direct image blur detection with the 2D Haar Wavelet transform (2D HWT). The algorithm classifies an image as blurred or sharp by splitting it into N x N tiles, applying four iterations of the 2D HWT to each tile, and grouping the horizontally, vertically, and diagonally connected tiles with pronounced changes into tile clusters. Images with large tile clusters are classified as sharp. Images with small tile clusters are classified as blurred. If necessary, the blur extent can be estimated as the ratio of the total area of the large tile clusters and the area of the whole image.

Our experiments on a sample of 500 images indicate that our algorithm either performs on par or outperforms two other blur detection algorithms found in the literature. The experiments also indicate that image blur detection has a pronounced positive effect on skewed barcode scanning. One possible implication of the research presented in this paper is that it may be possible to estimate blurriness in the images without explicitly computing the explicit features in the image that caused the blurriness (e.g., edges) or using involved methods to find the best fitting blur function.

In our future work, we plan to investigate the effect of image blur detection on vision-based nutrition label scanning to improve the optical character recognition (OCR) rates. In our previous work, we proposed to a greedy spellchecking algorithm to correct OCR errors during nutrition label scanning on smartphones [21]. The proposed algorithm, called *skip trie matching*, uses a dictionary of strings stored in the trie data structure to correct run-time OCR errors by skipping misrecognized characters while going down specific trie paths.

We expect that eliminating blurred frames from the processing stream, if done reliably, will improve the OCR rates and will make it possible to use open source OCR engines such as Tesseract (http://code.google.com/p/tesseract-ocr) and GOCR (http://jocr.sourceforge.net) in vision-based nutrition label scanning.

Acknowledgment

We would like to thank Tanwir Zaman and Sai Rekka for volunteering their time to classify five hundred images as blurred or sharp.

References

[1] Kulyukin, V. and Zaman, T. "Vision-based localization and scanning of 1D UPC and EAN barcodes with relaxed pitch, roll, and yaw camera alignment constraints." *International Journal of Image Processing* (IJIP), vol. 8, issue 5, 2014, pp. 355-383.

[2] Kulyukin, V. and Zaman, T. "An Algorithm for in-place vision-based skewed 1D barcode scanning in the cloud." In *Proceedings of the 18th International Conference on Image Processing and Pattern Recognition* (IPCV 2014), pp. 36-42, July 21-24, Las Vegas, NV, USA, CSREA Press, ISBN: 1-60132-280-1.

[3] Kulyukin, V. and Blay, C. "An Algoritm for mobile vision-based localization of skewed nutrition labels that maximizes specificity." In *Proceedings of the 18th International Conference on Image Processing and Pattern Recognition* (IPCV 2014), pp. 3-9, July 21-24, 2014, Las Vegas, NV, USA, CSREA Press, ISBN: 1-60132-280-1.

[4] Canny, J.F. "A Computational approach to edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, 1986, pp. 679-698.

[5] Laganiere, R. *OpenCV 2 Computer Vision Application Programming Cookbook*. Packt Publishing Ltd, 2011.

[6] Nievergelt, Y. *Wavelets Made Easy*. Birkäuser, Boston, 2000, ISBN-10: 0817640614.

[7] Tong, H., Li, M., Zhang, H., and Zhang, C. "Blur detection for digital images using wavelet transform," In *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol.1, pp. 27-30, June 2004. doi: 10.1109/ICME.2004.1394114.

[8] Cretea,F., Dolmierea, T., Ladreta, P., Nicolas, M. "The Blur effect: perception and estimation with a new no-reference perceptual blur metric." In *Proceedings of SPIE 6492, Human Vision and Electronic Imaging* XII, 64920I, San Jose, CA, USA, January 28, 2007. doi:10.1117/12.702790.

[9] Mallat, S. and Hwang, W. L. "Singularity detection and processing with wavelets." *IEEE Transactions on Information Theory*, vol. 38, no. 2, March 1992, pp. 617-643.

[10] Smith, J.O. *Mathematics of the Discrete Fourier Transform with Audio Applications*, 2nd Edition, W3K Publishing, 2007, ISBN 978-0-9745607-4-8.

[11] Rooms, F., Pizurica, A., Philips, W. "Estimating image blur in the wavelet domain." In *Proc. of IEEE Int. Conf. on Acoustics and Signal Processing*, vol. 4, pp.4190-4195, IEEE, 2002.

[12] Wanqing, S., Qing, L., Yuming, W. "Tool wear detection using Lipschitz exponent and harmonic wavelet." *Mathematical Problems in Engineering*, August 2013, Article ID 489261, http://dx.doi.org/10.1155/2013/489261.

[13] Venkatakrishnan, P., Sangeetha, S., and Sundar, M. "Measurement of Lipschitz exponent using wavelet transform modulus maxima." *International Journal of Scientific & Engineering Research*, vol. 3, issue 6, pp. 1-4, June-2012, ISSN 2229-5518.

[14] Pavlovic, G., and Tekalp, M. "Maximum likelihood parametric blur identification based on a continuous spatial domain model." *IEEE Trans. on Image Processing*, vol. 1, issue 4, Oct. 1992, pp. 496-504.

[15] Panchapakesan, K., Sheppard, D.G., Marcellin, M.W., and Hunt, B.R. "Blur identification from vector quantizer encoder distortion." In *Proc. of the 1998 International Conference on Image Processing* (ICIP 98), pp. 751-755, 4-7 Oct. 1998, Chicago, IL., USA.

[16] Marichal, X., Ma, W., and Zhang, H.J. "Blur determination in the compressed domain using DCT information." in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1999, vol. 2, pp. 386–390.

[17] Java implementation of the 2DHWT procedure. https://github.com/VKEDCO/java/tree/master/haar.

[18] Mobile supermarket barcode videos of grocery packages. https://www.dropbox.com/sh/q6u70wcg1luxwdh/LPtUBdwdY1.

[19] Python implementation of the blur detection algorithm proposed in reference [7]. https://github.com/VKEDCO/PYPL/blob/master/haar_blur.

[20] MATLAB implementation of blur detection algorithm proposed in reference [8]. http://www.mathworks.com/matlabcentral/fileexchange/24676-image-blur-metric.

[21] Kulyukin, V., Vanka, A., Wang, W. "Skip trie matching: a greedy algorithm for real-time OCR error correction on smartphones." *International Journal of Digital Information and Wireless Communication* (IJDIWC): vol. 3, issue 3, pp. 56-65, 2013. ISSN: 2225-658X.

Utilizing Discrete Wavelet Decomposition as an Effective Alternative to Watermarking

C. Martin[1*] **and M. Allali**[1]
[1]School of Computational and Data Sciences, Chapman University, 1 University Drive, Orange, CA 92866, USA

Abstract - *In this paper we propose utilizing the discrete wavelet transform to create an image's unique digital fingerprint which can then be stored and compared with other image fingerprints as an economical and efficient aid in copyright protection. The appeal of digital fingerprinting rather than watermarking is freedom from introducing information into the original image in order to prove its authenticity. Fingerprinting allows the image's creator to preserve an unaltered version of their work and still be protected from copyright infringement. This paper offers a simple and efficient solution to copyright infringement by using wavelets to create and compare image fingerprints as a worthy alternative to watermarking.*

Keywords: Image processing, wavelets, digital fingerprinting

1 Introduction

Although audio detection algorithms may rely on fingerprinting alone, previous research has embraced using either a combination of fingerprinting and watermarking or solely watermarking to protect an image's copyright. A common technique is applying a two-dimensional discrete wavelet transform (DWT), typically Haar or Daubechies, to decompose an image into frequency coefficients, embedding a watermark, constructing the image's fingerprint, then taking the two-dimensional inverse DWT of the manipulated frequency coefficients to produce a watermarked version of the image. In "Combined Watermarking and Fingerprinting Technologies for Digital Image Copyright Protection," authors Chang et al. propose an algorithm that not only embeds a non-random binary watermark into the image but also constructs the original image's fingerprint [1]. Another insightful paper in the area of digital fingerprinting is "An Image Protection Scheme Using the Wavelet Coefficients Based on Fingerprinting Technique" by Shin et al. [2]. Authors describe and use content-associated information, which is generated by combining a unique copyright seal, called a copyright message, with the wavelet coefficients of the original image or design [2].

In 2011, Yoshitomi et al. introduced an authentication method for digital audio relying solely on the audio signal's fingerprint, constructed using the Discrete Wavelet Transform [3]. They aimed to protect digital audio copyright without inserting information into the original audio signal. This was done by authenticating the audio using features extracted from the DWT coefficients of the transformed signal. Experimental results showed that the DWT method for audio authentication proved supremely tolerant to compression. The algorithm is impressive due to its simplicity and capacity to be just as effective as a digital watermarking scheme without the risk of audio quality degradation, which can result from inserting external information, such as a watermark, into the original audio file. Although their algorithm implements a one dimensional Daubechies DWT and relies on the tendency of an audio signal's wavelet coefficients to be distributed around zero, we wondered if it could be adapted for digital image copyright protection.

Another common theme in pertinent journal articles refers to the type of image processing operations the watermarking and fingerprinting algorithms were designed to overcome. Many watermarking schemes claim robustness if they tend to be impervious to signal processing attacks such as compression, particularly lossy compression, additive noise, filtering, and histogram manipulation. Less common are algorithms that are designed to detect a watermark when the image has been cropped, re-sized, or rotated. An algorithm that performs well up against such geometric distortions experimentally is rarer still.

2 Wavelets and the Discrete Wavelet Transform

Fundamentally, wavelets are basis functions representing other, preferably more complicated, functions [4]. Wavelets are useful in signal and image processing because they allow a function, or signal, to be understood in terms of its overall shape as an approximation, as well as its details at various levels of decomposition [5]. This indicates the reason for the name multi-resolution analysis (MRA), which is used to construct wavelets. Wavelet decomposition is useful for describing signals, or images, at various levels of resolution. Orthonormal bases of wavelets can be constructed by a series of successive iterative approximations of L^2-functions. These approximations use a different resolution at each level [6]. It is important to note that

* Corresponding Author: Chloe Martin, marti192@mail.chapman.edu

MRA will always generate a wavelet but a wavelet does not lead to an MRA. More in-depth coverage of MRA can be found in [6].

A function, ψ is called an orthonormal wavelet if the family $\{\psi_{j,k}\}$, defined as:

$$\psi_{j,k}(x) \coloneqq \sqrt{2^j}\psi(2^j x - k), j,k \in Z, \quad (1)$$

is an orthonormal basis of $L^2(R)$, where $L^2(R)$ is the set of square integrable functions. Wavelets are described by the following equations:

$$\varphi(x) = \sum_{k\epsilon Z} c[k]\sqrt{2}\varphi(2x-k), \quad (2)$$

$$\psi(x) = \sum_{k\in Z}(-1)^k c[1-k]\sqrt{2}\varphi(2x-k), \quad (3)$$

where $\varphi(x)$ is the scaling function and $\psi(x)$ is the wavelet function.

We favored Daub4 experimentally because the associated set of filter coefficients; $\left\{c[0] = \frac{1+\sqrt{3}}{4\sqrt{2}},\ c[1] = \frac{3+\sqrt{3}}{4\sqrt{2}},\ c[2] = \frac{3-\sqrt{3}}{4\sqrt{2}},\ c[3] = \frac{1-\sqrt{3}}{4\sqrt{2}}\right\}$, where $c[k] = 0$ for $k \neq 0,1,2,3$, is the smallest set of coefficients that produces a continuous, compactly supported scaling function.

Daub4 is represented by the following equations:

$$\phi_m(x) = \sum_{k\in\mathbb{Z}} c_k\sqrt{2}\phi_{m-1}(2x-k), \quad (4)$$

$$\phi_0(x) = \mathbb{1}_{[0,1)}(x), \quad (5)$$

where $\mathbb{1}_{[0,1)}(x)$ is the indicator function:

$$\mathbb{1}_{[0,1)}(x) = \begin{cases} 1 & if\ x\epsilon[0,1) \\ 0 & otherwise \end{cases} \quad (6)$$

It is important to devise an algorithm to create an image's fingerprint that withstands at least two or three vigorous forms of image manipulation. It should not be necessary to amend or repair the algorithm to suit each fingerprint that passes through the database. The wavelet transform is powerful because it preserves spatial information as well as frequency information, unlike the Fourier transform. When a change is made locally to an image that has been transformed into the frequency domain by Fourier basis functions, that manipulation is applied to the entire image in the spatial domain [4]. Utilizing a wavelet transformation ensures any tampering done to the wavelet coefficients delivers results that are confined to the area they describe in the spatial domain. Additionally, images can be represented with considerably less wavelet basis functions than sine-cosine basis functions [4].

3 Experimental Setup and Algorithm

The experimental component of this paper consists of loading several images of logos into Matlab then creating and storing each image's unique digital fingerprint. Each fingerprint is saved as a matrix with pixel values belonging to the set {0, 1, ½}. A fingerprint is created using an algorithmically coded string of 0's, 1's, and ½'s representing specific discrete wavelet coefficients, which are later used to compare one image with another. If a potential match is detected, the fingerprint is reconstructed into a compressed, tri-gray-scale approximation of the original image. Although the compressed image is not stored in the database, it is easily constructed using the image's previously stored fingerprint. This saves on memory resources because storing any additional image information is unnecessary and saves on time because it is only constructed in the event of a potential match. Storing an image's tri-gray-scale digital fingerprint takes up less memory than storing the original rgb uint8 (unsigned integer, 8-bits per pixel) or double precision image while still preserving enough information from the original image to aid in copyright protection. Also, because information is not being inserted into the original image, the image maintains its intended appearance and integrity.

3.1 Creating an Image's Fingerprint

In order to protect an image's copyright, the image's fingerprint must be created and stored. The following steps detail how an image's fingerprint is made. We will use the image, *Rocket*, shown in Figure 1 as an example to illustrate the process [7].

Figure 1, *Rocket* [7]

1. An image, I is submitted as a three-dimensional color rgb uint8 matrix or two-dimensional gray-scale matrix with the stipulation that length and width must be equal and even.
2. This image is converted to a two-dimensional double precision matrix with gray-scale values in the range $[0, 255]$, where 0 is white and 255 is black.
3. The Daubechies wavelet transform is used to decompose the image into wavelet coefficients corresponding to an approximation and details of the image. Figure 2 is a plot representing all wavelet coefficients of *Rocket* at Daub4 and $j = 1$. The four main clusters in the wavelet coefficient plot represent the approximation, horizontal detail, vertical detail, and diagonal detail coefficients, respectively.

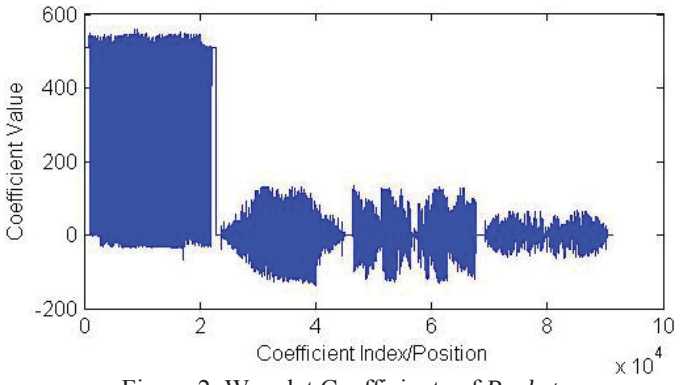


Figure 2, Wavelet Coefficients of *Rocket*

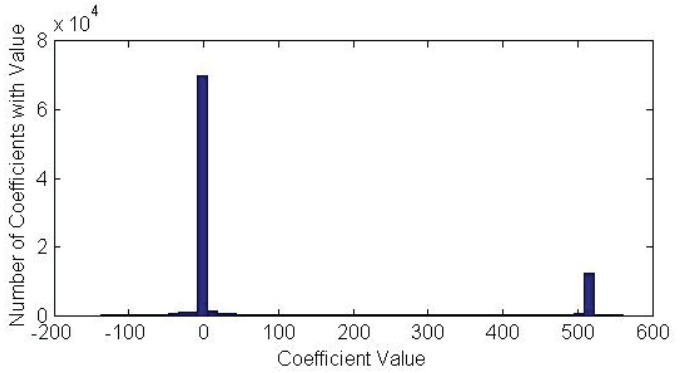


Figure 3, Wavelet Histogram

4. Next, the wavelet coefficients are divided into three groups depending on their frequency of occurrence. Figure 3 is a 54-bin histogram representation of the coefficients. Note that because the coefficient values are floating point numbers with double precision, bins are necessary for allowing us to see volume associated with specific ranges of coefficient values. For a 300 by 300 pixel image, the minimum number of coefficients is 90,000, and this special case occurs for Daub2, which is equivalent to the Haar Wavelet. For Daub4, there are 91,204 coefficients. The coefficients are rescaled according to an algorithm similar to the one utilized in Yoshitomi et al. The authors exploit the fact that a wavelet coefficient histogram is centered at approximately zero when the DWT is performed on audio data [3]. Although image wavelet coefficients do not share this property precisely, the wavelet coefficients from performing the DWT on an image do produce a histogram tending to have strongest modality at approximately zero as can be seen in Figure 3. Grouping an image's wavelet coefficients according to an algorithm that sorts based on the unimodal histogram characteristic does nonetheless lead to encouraging results. The algorithm used to code an image's fingerprint is as follows:
    1) Coefficient values within a certain distance from zero are assigned the value 0.
    2) Coefficients close to the minimum and maximum values are assigned the value 1.
    3) Coefficients not belonging to either of the sets above are assigned the value $\frac{1}{2}$.

This concept is represented in Figure 4. The size of each region is set depending on the desired ratio of coefficients falling into each category. Figure 5 is the histogram of the wavelet coefficients for *Rocket* with the reassigned value ranges indicated.

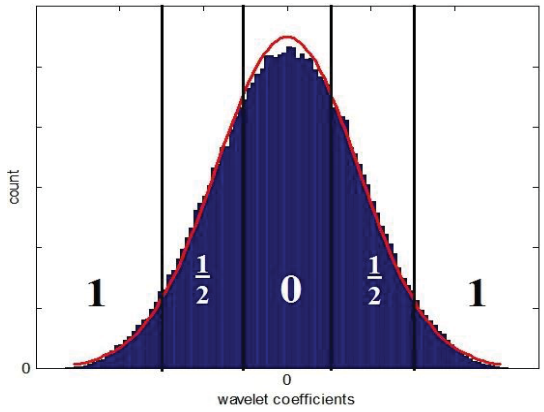


Figure 4, Histogram of DWT coefficients for an audio signal with value sets

5. The vector of values $\left\{0, 1, \frac{1}{2}\right\}$ is the fingerprint of the image. It may be stored as a vector or a matrix depending on preference. Figure 6 and Figure 7 represent the matrix form of the wavelet coefficients and the matrix form of the image's fingerprint respectively. The quality of the information does not appear to be compromised. The two vectors have the same length but rather than processing a vector with values ranging from -137.0511 to 560.8005 with double floating point precision, we are now working with a bit vector of 0's, 1's, and ½'s. In order to create the matrix representation of the wavelet coefficients, Figure 6, in Matlab, we rescaled the range of values to go from 0 to 255.

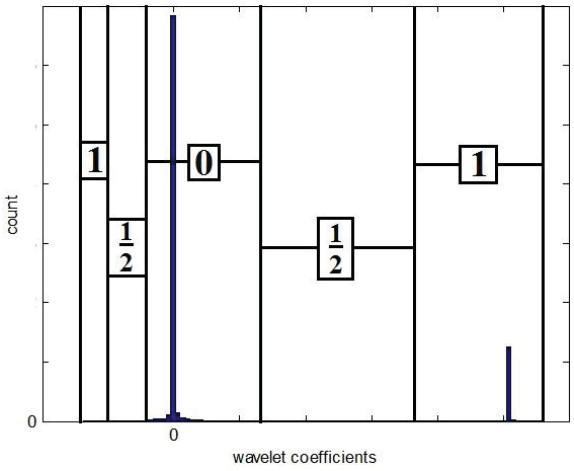


Figure 5, Bimodal histogram of DWT coefficients for *Rocket* with value sets

Figure 6, Matrix representation of the wavelet coefficients; pixel values [0,255].

Figure 7, Matrix representation of image's fingerprint vector; pixel values {0, 1, ½}.

3.2 Comparing Images Using Fingerprints

Before an image's fingerprint is created and added to the database, it is compared to all previously stored image fingerprints in the database. The local vector corresponding to the new image's wavelet coefficients is determined as follows:

1. DWT decomposition is applied to the new image and its coefficients are divided into two categories: $\{0,1\}$, depending on their distance from 0.
2. If the finger print (FP) we are comparing the new image with has a wavelet coefficient in section 0 or 1 at position (index) i, and the new image has a wavelet coefficient in section 0 at position i, then the new image's local vector will have a 0 at position i. If the FP has a wavelet coefficient in section 0 or 1 at position i and the new image has a wavelet coefficient in section 1 at position i, then the new image's local vector will be assigned a 1 at position i. If neither is the case, that is, if the FP has a $\frac{1}{2}$ at position i, then the new image's local vector will have a value of $\frac{1}{2}$ at position i, regardless of whether the new image's coefficient at position i is in section 0 or 1.

In Figure 8, I_c represents the matrix version of a previously stored image's fingerprint created using the algorithm above. NEW_c represents the matrix version of the wavelet coefficients of the new image that fall either within or outside of a specified threshold distance from 0. Finally, $LOCAL$ is the matrix version of the local vector based on the values of I_c and NEW_c. Notice that the elements of $LOCAL$ are identical to NEW_c sans elements corresponding to $I_c(i) = \frac{1}{2}$. The new image's local coefficient vector should not be confused with the image's fingerprint. The local vector is only valid for comparison with one fingerprint because it is dependent on the values of that fingerprint. The more similar two images are, the more similar the local vector will be to the previously stored fingerprint. The comparison process relies heavily on the wavelet coefficients to determine whether or not an image is a match.

$$I_c = \begin{bmatrix} .5 & 1 & .5 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & .5 & .5 & 0 \\ .5 & 0 & 0 & .5 \end{bmatrix}, \quad NEW_c = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad LOCAL = \begin{bmatrix} .5 & 0 & .5 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & .5 & .5 & 1 \\ .5 & 1 & 1 & .5 \end{bmatrix}$$

Figure 8

Initially, the assessment of whether two images are sufficiently similar occurs at the bit level with the correlation of the original image's fingerprint and the new image's local coefficient vector being examined. If the correlation is high enough, the original image's fingerprint and the new image's local coefficient vector are reprocessed using the inverse discrete wavelet transform at a pre-specified order and j-level. If the correlation between these two tri-gray-scale images is also high, we say that we have a potential match. If the correlation between the two fingerprints is low, we examine the correlation between the wavelet coefficients to make a final discerning decision about whether or not there is a potential copyright infringement.

4 Experimental Results

4.1 General Key Terms and Measurements

In the following experimental results tables, FPLV (Finger Print and Local Vector) correlation refers to the correlation between the original image's fingerprint and the new image's temporary local vector. APP (Approximation) correlation refers to the correlation between the two sets of approximation coefficients resulting from wavelet reconstruction being performed on the original image's fingerprint and the new image's local vector. AR (Authentication Ratio) refers to a percentage value associated with whether or not a new image is substantially similar to a previously stored image to be called a match. It is used for the authentication of digital audio in Yoshitomi et al. [3, p. 62]. We chose to keep track of the AR in our analyses and as a means of cross-checking our results. According to Yoshitomi et al., an authentication ratio of 95% or higher is indicative of a match [3, p. 64]. The authentication ratio is given by equation (7).

$$AR = \frac{100 \sum_{i=1}^{N} b_i(1-|p_i - v_i|)}{\sum_{i=1}^{N} b_i}, \qquad (7)$$

where b_i is a binary value indicating whether or not the original image's fingerprint has a value at position i in the 0 or 1 category, or conversely, in the $\frac{1}{2}$ category, p_i is the value of the original image's fingerprint at position i, and v_i is the value of the new image's temporary local vector at position i. Notice that the wavelet coefficients assigned a value of ½ are not included in the calculation of the AR since the corresponding b_i will equal 0 in those instances. FPLV $(b_i = 1)$ correlation is equivalent to FPLV with all fingerprint and temporary local vector values equal to ½ removed from the correlation calculation. For simplicity, unless stated otherwise each image is 300 by 300 pixels.

4.2 Setup I

Daubechies Order: Daub2, *db1*, Haar
Resolution Level: *j=2*
Lossy tri-gray-scale compressed image size: 75x75
Image: *Rocket*, Figure 1

Table 1

Manipulation Type	FPLV corr	APP corr	AR (%)	FPLV ($b_i = 1$)
Mean Filter	0.9994	1	99.9955	0.9994
Median Filter	1	1	100	1
Mean & Median Filters	0.9996	1	99.9966	0.9995
Laplacian Enhancement	0.9738	0.9911	99.7880	0.9716
Gaussian Noise (0,.01)[2], Figure 9, left	0.9528	0.9638	99.6154	0.9489
Gaussian Noise (.2,.5), Figure 9, right	0.7704	0.8428	97.4999	0.7572
Salt & Pepper[3]	0.9700	0.9674	99.7621	0.9675
Histogram Flattening	0.9923	0.9996	99.9391	0.9916
KCT.1[4]	0.6607	0.5082	97.6679	0.6246
KCT.2, Figure 10	0.7009	0.5138	97.9363	0.6692

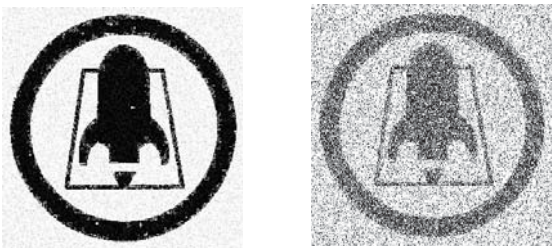


Figure 9, (left) *Rocket* with Gaussian noise, 0.0 mean, 0.01 variance, and (right) 0.2 mean, 0.5 variance.

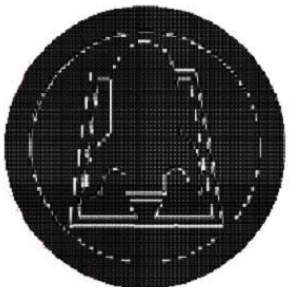


Figure 10, *Rocket* with KCT applied and key coefficients set to average value

4.3 Setup II

Daubechies Order: Daub4, db2
Resolution Level: j=1
Lossy tri-gray-scale compressed image size: 151x151
Image: *Rocket*
Secondary Image: *Roller* [8], Figure 11

Figure 11, *Roller*

Table 2

Manipulation Type	FPLV corr	APP corr	AR (%)	FPLV ($b_i = 1$)
Mean Filter	0.9882	1	99.6843	0.9880
Median Filter	1	1	100	1
Mean & Median Filters	0.9879	1	99.6765	0.9877
Laplacian	0.9514	0.9550	98.6869	0.9505
Gaussian Noise (0, .01)	0.9499	0.9388	98.6108	0.9491
Gaussian Noise (.2, .5)	0.7785	0.7216	93.2956	0.7754
Salt & Pepper	0.9131	0.9347	97.4902	0.9117
Histogram Flattening	0.9619	0.9930	98.9690	0.9612
KCT.3, Figure 12	0.6850	0.5352	92.3205	0.6778
Spliced: center 50% dead pixels	0.7044	0.5455	92.7202	0.6979
Spliced: left 50% *Roller*, Figure 13, left	0.8423	0.6638	95.8143	0.8393
Spliced: copy-paste, Figure 13, right	0.7774	0.4928	93.7680	0.7737
Compression Standard (Original:New)[5]				
JPG:BMP	1	1	100	1
BMP:JPG	1	1	100	1
JPG:GIF	1	1	100	1
GIF:JPG	1	1	100	1
JPG:PNG	1	1	100	1
PNG:JPG	1	1	100	1
BMP:GIF	1	1	100	1

[2] Gaussian white noise with constant mean and variance. Mean and variance parameters were varied during experimentation.
[3] Salt and pepper applies white and black pixels to approximately five percent of the pixels in an image. The noise density was varied during experimentation.
[4] Key Coefficient Tampering (KCT) is when an outside entity changes the DWT coefficients of a previously copyrighted image then applies wavelet reconstruction in order to create a very similar image, or counterfeit version, with the intention that it will not be detected by the proposed image copyright protection algorithm. This type of manipulation relies on knowledge of which wavelet type, wavelet order, and j-level the algorithm uses to create an original image's fingerprint along with the chosen ratio of coefficients within each set: $\left\{0, 1, \frac{1}{2}\right\}$.
[5] The compression results were the same for each wavelet order, j-level, and image tested and therefore compression results are

GIF:BMP	1	1	100	1
BMP:PNG	1	1	100	1
PNG:BMP	1	1	100	1
GIF:PNG	1	1	100	1
PNG:GIF	1	1	100	1

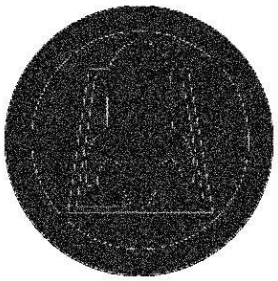


Figure 12, *Rocket* with KCT and key coefficients set to random values in range of DWT coefficient values.

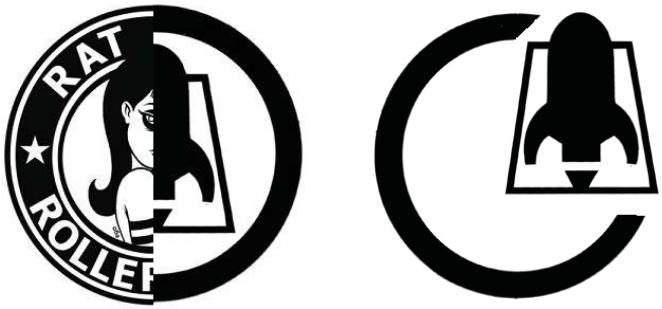


Figure 13, (left) Spliced *Rocket* with 50% *Roller* and (right) spliced copy-paste

4.4 Setup III

Daubechies Order: Daub4, db2
Resolution Level: *j*=2
Lossy tri-gray-scale compressed image size: 77x77
Image: Rocket

Table 3

Manipulation Type	FPLV corr	APP corr	AR (%)	FPLV ($b_i = 1$)
Mean Filter	0.9968	1	99.9735	0.9965
Median Filter	1	1	100	1
Mean & Median Filters	0.9968	1	99.9735	0.9965
Laplacian	0.9692	0.9942	99.7353	0.9666
Gaussian Noise (0, .01)	0.9554	0.9666	99.6139	0.9517
Gaussian Noise (.2, .5)	0.7674	0.8550	97.2963	0.7541
Salt & Pepper	0.9474	0.9670	99.5400	0.9431
Histogram Flattening	0.9423	1	99.4893	0.9377
KCT.2 (mean)	0.7531	0.5571	98.1678	0.7273
Spliced: center 50% dead pixels	0.7517	0.5499	98.1590	0.7258
Spliced: left 50% *Roller*, Figure 13, left	0.8695	0.6730	98.9079	0.8580
Spliced: copy-paste, Figure 13, right	0.8248	0.5172	98.4678	0.8103

solely listed in Table 2 and are not included in additional experimental results tables.

4.5 Comparing Similar and Dissimilar Images

The next phase of experimentation tested how the algorithm performed when previously stored original image fingerprints are compared with new or never introduced image information. The premise of the experiment is that a set of new images is submitted to the database for copyright protection purposes and therefore each new image must go through the comparison process in order to ensure that it is an original and its fingerprint can then be created and preserved. We wanted to see how images that are similar to and different from a specific original image effect the correlation values and authentication value at various DWT orders and levels. To save space and not distract from the body of the paper, we opted to describe the results rather than include all experimental data tables. Also, not all images experimented with or listed in the experimental results tables in the following sections are pictured.

4.6 Setup IV

Daubechies Order: Daub4, db2
Resolution Level: *j*=1
Lossy tri-gray-scale compressed image size: 151x151
Original Image: Rocket

Table 4

New Image	FPLV corr	APP corr	AR (%)	FPLV ($b_i = 1$)
Rocket (self)	1	1	100	1
Roller	0.6906	0.3223	91.7429	0.6848
Descendants	0.7033	0.0730	89.9148	0.7003
Swami	0.6757	0.1563	90.3066	0.6710
Arts	0.6541	0.2668	90.7488	0.6470
Dead	0.2041	-0.0728	84.6019	0.1663
Misfits	0.2945	-0.2115	84.0309	0.2752
Stitches	0.2823	-0.0648	84.5134	0.2600
Vandals	0.2783	-0.0085	85.2713	0.2508
RocketColor	0.9991	1	99.9754	0.9991
RocketRocket	0.8033	0.4682	93.4209	0.8012
RocketRocketSide	0.7301	0.1991	91.1686	0.7270
RocketRocketSideBack	0.7307	0.1985	91.2246	0.7276

4.7 Analysis

Based on the experimental results, we found that the higher the resolution level, the higher the threshold must be to distinguish a match from a non-match. For example, assuming the wavelet order is 2 (Daub4) in both cases, a reasonable

threshold for the FPLV correlation when j=1 is ≈0.70 whereas a reasonable threshold for the FPLV correlation when j=2 is ≈0.75. Additionally, we discovered that a higher wavelet order corresponds to only slightly higher FPLV correlation values. As long as there is a threshold that depends on the resolution level, there is a way of successfully categorizing an image as a match or not while minimizing the number of false positives and negatives. According to the graph in Figure 14, a FPLV correlation above 0.72 is indicative of a match. In the two cases where the FPLV correlation is less than 0.72, the APP correlation is above 0.50. We used this information to guide our criteria for determining potential matches.

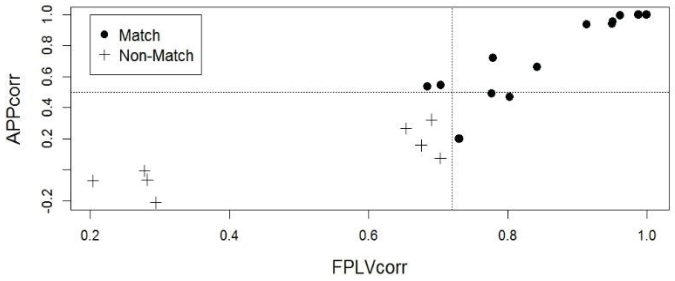


Figure 14, APPcorr vs FPLVcorr Data, Daub4, *j=1*

5 Breaking the System

In order to break the algorithm and copyright an image that has already been copyrighted, one would need to know which wavelet type is used, which order, and which j-level. Suppose a party knew that image copyright is preserved by taking the DWT using Daub4 at j=1. They could then perform the transform on the image to come up with the multi-resolution wavelet coefficients. Multiresolution representation (MRR), not to be confused with MRA, means all wavelet coefficients are used for the generation of an image's fingerprint, not solely the approximation coefficients.

They would also need to know which ratio of coefficients are assigned to each of the three categories {0, 1, ½} in order to create a binary vector of flags, equivalent to b described in equation (7). This would allow them to manipulate the appropriate coefficients in order to trick the system. We expected the AR and the FPLV ($b_i = 1$) correlation to be the only two match-discerning values severely affected by tampering with coefficient values because they rely solely on coefficients in categories 0 and 1. However, the other two correlation calculations, FPLV correlation and APP correlation use all coefficients. Experimentally, all values were affected by KCT.

The appearance of a logo with KCT applied is likely undesirable for marketing purposes because the logo is not clear or simple. Most importantly, it does not resemble the design the individual set out to steal. The less intrusive the KCT is, the less effective it is in fooling the detection algorithm. The proposed digital fingerprint generation and image comparison algorithm does not seem to be easily broken. As observed in 0, even the most aggressive KCT versions generally give higher correlation values than non-matches do, particularly APP correlation values.

6 Conclusion

Many times, a design is not tampered with directly. Rather, an individual sees the design then mimics what they see in a new instance using an arbitrary graphic design program. No trace of the original design is present in the new, plagiarized design. This is an area where digital watermarking fails but fingerprinting will not. Logos are unique because employing the most resilient, yet quiet watermark invented can nonetheless be a waste of time and money. Most logos do not have intricate detail and therefore their general look, shape, and feel can be easily rendered on paper or in a digital image design program.

In this paper we proposed and experimented with a simple algorithm based on the digital audio authentication method described in [3]. Rather than using the one dimensional discrete wavelet transform, we used the two dimensional discrete wavelet transform so that the authentication algorithm could be applied to images. We opted to focus on more tampering techniques than solely compression. We found the algorithm to be effective not only with compressed images, but also filtered images, noisy images, and degraded images.

7 References

[1] M.-C. Chang, D.-C. Lou and H.-K. Tso, "Combined Watermarking and Fingerprinting Technologies for Digital Image Copyright Protection," *The Imaging Science Journal,* vol. 55, no. 1, pp. 3-12, 2007.

[2] J.-W. Shin, J. C. Yang, S. Yoon and D.-S. Park, "An Image Protection Scheme Using the Wavelet Coefficients Based on Fingerprinting Technique," *Computational Intelligence and Security: International Conference, CIS,* pp. 642-51, 2007.

[3] Y. Yoshitomi, T. Asada, Y. Kinugawa and M. Tabuse, "An Authentication Method for Digital Audio Using a Discrete Wavelet Transform," *Journal of Information Security,* vol. 2, no. 2, pp. 59-68, 2011.

[4] B. Vidakovic and P. Mueller, "Wavelets for Kids: A Tutorial Introduction".

[5] E. J. Stollnitz, T. D. DeRose and D. H. Salesin, "Wavelets for Computer Graphics: A Primer Part 1," *IEEE Computer Graphics and Applications,* vol. 15, no. 3, pp. 76-84, 1995.

[6] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics,* vol. 41, no. 7, pp. 909-96, 1988.

[7] Pushead, *Rocket from the Crypt,* San Diego, California, 1989.

[8] S. Ego, *Rat City Roller Girls,* Seattle: Rat City Roller Girls, LLC, 2004.

Text Skew Angle Detection in Vision-Based Scanning of Nutrition Labels

Tanwir Zaman
Department of Computer Science
Utah State University
Logan, UT, USA
tanwir.zaman@aggiemail.usu.edu

Vladimir Kulyukin
Department of Computer Science
Utah State University
Logan, UT, USA
vladimir.kulyukin@usu.edu

***Abstract*— An algorithm is presented for text skew angle detection in vision-based scanning of nutrition labels on grocery packages. The algorithm takes a nutrition label image and applies several iterations of the 2D Haar Wavelet Transform (2D HWT) to downsample the image and to compute the horizontal, vertical, and diagonal change matrices. The values of these matrices are binarized and combined into a result set of 2D change points. The convex hull algorithm is applied to this set to find a minimum area rectangle containing all text pixels. The text skew angle is computed as the rotation angle of the minimum area rectangle found by the convex hull algorithm. The algorithm's performance is compared with the performance of the algorithms of Postl and Hull, two text skew angle algorithms frequently cited in the literature, on a sample of 607 nutrition label images whose text skew angles were manually computed by two human evaluators. The median text skew angle error of the proposed algorithm, Postl's algorithm, and Hull's algorithm are 4.62, 68.85, and 20.92, respectively.**

***Keywords*— computer vision; text skew angle detection; OCR; 2D Haar wavelet transform; wavelet analysis**

I. Introduction

Vision-based extraction of nutritional information from nutrition labels (NLs) available on most product packages is critical to proactive nutrition management, because it improves the user's ability to engage in continuous nutritional data collection and analysis. Many nutrition management systems underperform, because the target users find it difficult to integrate nutritional data collection into their daily activities due to lack of time, motivation, or training, which causes them to turn off or ignore such digital stimuli as emails, phone calls, and SMS's.

To make nutritional data collection more manageable and enjoyable for the users, we are currently developing a Persuasive NUTrition Management System (PNUTS) [1]. PNUTS seeks to shift current research and clinical practices in nutrition management toward persuasion, automated nutritional information processing, and context-sensitive nutrition decision support. PNUTS is inspired by the Fogg Behavior Model (FBM) [2], which states that motivation alone is insufficient to stimulate target behaviors. Even a motivated user must have both the ability to execute a behavior and a well-designed trigger to engage in that behavior at an appropriate place or time.

In our previous research, we developed a vision-based localization algorithm for horizontally or vertically aligned nutrition labels (NLs) on smartphones [3]. Our next NL processing algorithm [4] improved the algorithm proposed in [1] in that it handled not only aligned NLs but also NLs skewed up to 35-40 degrees from the vertical axis of the captured frame. A limitation of that algorithm was its inability to handle arbitrary text skew angles.

The algorithm presented in this paper continues our investigation of vision-based NL scanning. The algorithm's objective is to determine the text skew angle of an NL text in the image without constraining the angle's magnitude. If the skew angle is estimated correctly, the image can be rotated accordingly so that the standard optical character recognition (OCR) techniques can be used to extract nutrition information.

The algorithm takes an NL image and applies several iterations of the 2D HWT to downsample the image and to compute horizontal, vertical, and diagonal change matrices. The values of these matrices are binarized and combined into a result set of 2D points. The convex hull algorithm [5] is then applied to this set in order to find a minimum area rectangle containing all text pixels. The text skew angle is computed as the rotation angle of the minimum area rectangle found by the convex hull algorithm.

Our paper is organized as follows. In Section II, we give some background information and discuss related work. In Section III, we present our text skew angle detection algorithm and explain how it works. In Section IV, we describe the experiments we designed and conducted to test the algorithm's performance on a sample of NL images and to compare it with the text skew angle detection algorithms of Postl [6] and Hull [7], two classic algorithms frequently cited in the literature. In Section V, we analyze and discuss the results of the experiments. In Section VI, we present our conclusions and outline several directions for our future work.

II. Background

A. Related Work

A variety of algorithms have been developed to determine the text skew angle. Such algorithms typically use horizontal or vertical projection profiles. A horizontal projection profile is a 1D array whose size is equal to the number of rows in the image. Similarly, a vertical projection profile is a 1D array whose size is equal to the number of columns in the image. Each location in a projection profile stores a count of the number of black pixels associated with text in the

corresponding row or column of the image. Projections can be thought of as 1D histograms. A horizontal projection histogram is computed by rotating the input image through a range of angles and calculating black pixels in the appropriate bins. All projection profiles for all rotation angles are compared with each other to determine which one maximizes a given criterion function.

Postl's algorithm [6] uses the horizontal projection profile for text skew angle detection. The algorithm calculates the horizontal projection profiles for angles between 0 and 180 degrees in small increments, e.g., 5 degrees. The algorithm uses the sum of squared differences between adjacent elements of the projection profile as the criterion function and chooses the profile that maximizes that value.

Hull [7] proposes a text skew angle detection algorithm similar to Postl's. Hull's algorithm is more efficient, because it rotates individual pixels instead of rotating entire images. Specifically, the coordinates of every black pixel are rotated to save temporary storage and thereby to reduce the computation that would be required for a brute force implementation.

Bloomberg et al. [8] also use projection profiles to determine the text skew angle. Their algorithm differs from Postl's and Hull's algorithms in that the images are downsampled before the projection profiles are calculated in order to reduce computational costs. The criterion function used to estimate the text skew angle is the variance of the number of black pixels in a scan line.

Kanai et al. [9] present another text skew angle estimation algorithm based on projection profiles. The algorithm extracts fiducial points and uses them as points of reference in the image by decoding the lowest resolution layer of the JBIG compressed image. The JBIG standard consists of two techniques, a progressive encoding method and a lossless compression method for the lowest resolution layer. These points are projected along parallel lines into an accumulator array. The text skew angle is computed as the angle of projection within a search interval that maximizes alignment of the fiducial points. This algorithm detects a skew angle in the limited range from ±5 degrees to ±45 degrees.

Papandreou and Gatos [10] use vertical projection profiles for text skew angle detection. The criterion function is the sum of squares of the projection profile elements. The researchers argue that their method is resistant to noise and image warping and works best for the languages where most of the letters include at least one vertical line, such as languages with Latin alphabets.

Li et al. [11] propose a text skew angle detection algorithm based on wavelet decompositions and projection profile analysis. Document images are divided into sub-images using wavelet transform. The matrix containing the absolute values of the horizontal sub-band coefficients, which preserves the text's horizontal structure, is then rotated through a range of angles. A projection profile is computed at each angle, and the angle that maximizes a criterion function is regarded as the skew angle.

Shivakumara et al. [12] propose a document skew angle estimation approach based on linear regression. They use linear regression formula in order to estimate a skew angle for each text line segment of a text document. The part of the text line is extracted using static and dynamic thresholds from the projection profiles. This method is based on the assumption that there is space between text lines. The method loses accuracy for the documents having skew angle greater than 30 degrees and appears to work best for printed documents with well-separated lines.

B. 2D Haar Transform

Our implementation of the 2D HWT is based on the approach taken in [13] where the transition from 1D Haar wavelets to 2D Haar wavelets is based on the products of basic wavelets in the first dimension with basic wavelets in the second dimension. For a pair of functions f_1 and f_2 their tensor product is defined as $(f_1 \times f_2)(x, y) = f_1(x) \cdot f_2(x)$. Two 1D basic wavelet functions are defined as follows:

$$\varphi_{[0,1[}(r) = \begin{cases} 1 \; if \; 0 \le r < 1, \\ 0 \; otherwise. \end{cases}$$

$$\psi_{[0,1[}(r) = \begin{cases} 1 \; if \; 0 \le r < \frac{1}{2}, \\ -1 \; if \; \frac{1}{2} \le r < 1, \\ 0 \quad otherwise. \end{cases}$$

The 2D Haar wavelets are defined as tensor products of $\varphi_{[0,1[}(r)$ and $\psi_{[0,1[}(r)$: $\Phi_{0,0}^{(0)}(x,y) = \left(\varphi_{[0,1[} \times \varphi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{h,(0)}(x,y) = \left(\varphi_{[0,1[} \times \psi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{v,(0)}(x,y) = \left(\psi_{[0,1[} \times \varphi_{[0,1[}\right)(x,y)$, $\Psi_{0,0}^{d,(0)}(x,y) = \left(\psi_{[0,1[} \times \psi_{[0,1[}\right)(x,y)$. The superscripts h, v, and d indicate the correspondence of these wavelets with horizontal, vertical, and diagonal changes, respectively. The horizontal wavelets detect horizontal (left to right) changes in 2D data, the vertical wavelets detect vertical (top to bottom) changes in 2D data, and the diagonal changes detect diagonal changes in 2D data.

In practice, the basic 2D HWT is computed by applying a 1D wavelet transform of each row and then a 1D wavelet transform of each column. Suppose we have a 2 x 2 pixel image

$$\begin{bmatrix} s_{0,0} & s_{0,1} \\ s_{1,0} & s_{1,1} \end{bmatrix} = \begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix}.$$

Applying a 1D wavelet transform to each row results in the following 2 x 2 matrix:

$$\begin{bmatrix} \frac{s_{0,0}+s_{0,1}}{2} & \frac{s_{0,0}-s_{0,1}}{2} \\ \frac{s_{1,0}+s_{1,1}}{2} & \frac{s_{1,0}-s_{1,1}}{2} \end{bmatrix} = \begin{bmatrix} \frac{11+9}{2} & \frac{11-9}{2} \\ \frac{7+5}{2} & \frac{7-5}{2} \end{bmatrix} = \begin{bmatrix} 10 & 1 \\ 6 & 1 \end{bmatrix}.$$

Applying a 1D wavelet transform to each new column is fetches us the result 2 x 2 matrix:

$$\begin{bmatrix} \frac{10+6}{2} & \frac{1+1}{2} \\ \frac{10-6}{2} & \frac{1-1}{2} \end{bmatrix} = \begin{bmatrix} 8 & 1 \\ 2 & 0 \end{bmatrix}.$$

The coefficients in the result matrix obtained after the application of the 1D transform to the columns express the original data in terms of the four tensor product wavelets $\Phi_{0,0}^{(0)}(x,y)$, $\Psi_{0,0}^{h,(0)}(x,y)$, $\Psi_{0,0}^{v,(0)}(x,y)$, and $\Psi_{0,0}^{d,(0)}(x,y)$:

$$\begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix} = 8 \cdot \Phi_{0,0}^{(0)}(x,y) + 1 \cdot \Psi_{0,0}^{h,(0)}(x,y) + 2 \cdot \Psi_{0,0}^{v,(0)}(x,y) + 0 \cdot \Psi_{0,0}^{d,(0)}(x,y).$$

The value 8 in the upper-left corner is the average value of the original matrix: (11+9+7+5)/4= 8. The value 1 in the upper right-hand corner is the horizontal change in the data from the left average, (11+7)/2=9, to the right average, (9+5)/2=7, which is equal $1 \cdot \Psi_{0,0}^{h,(0)}(x,y) = 1 \cdot -2$. The value 2 in the bottom-left corner is the vertical change in the original data from the upper average, (11+9)/2=10, to the lower average, (7+5)/2=6, which is equal to $2 \cdot \Psi_{0,0}^{v,(0)}(x,y) = 2 \cdot -2 = -4$. The value 0 in the bottom-right corner is the change in the original data from the average along the first diagonal (from the top left corner to the bottom right corner), (11+5)/2=8, to the average along the second diagonal (from the top right corner to the bottom left corner), (9+7)/2=8, which is equal to $0 \cdot \Psi_{0,0}^{d,(0)}(x,y)$. The decomposition operation can be represented in terms of matrices:

$$\begin{bmatrix} 11 & 9 \\ 7 & 5 \end{bmatrix} = 8 \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} + 2 \cdot \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} + 0 \cdot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

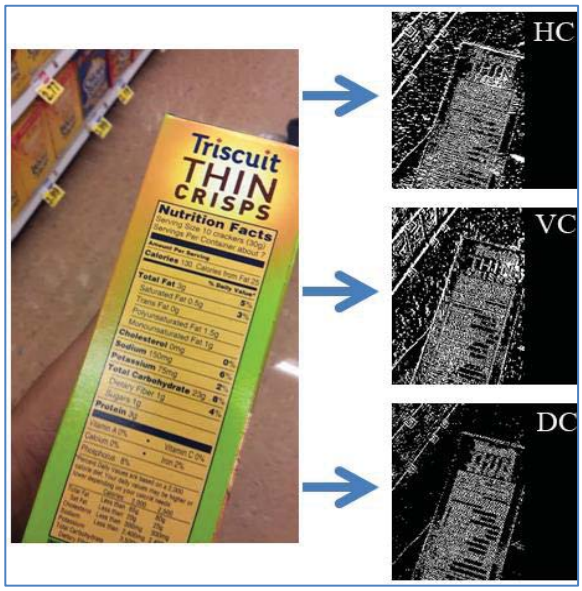


Figure 1. Horizontal, vertical, and diagonal changes

III. Text Skew Angle Detection

The proposed algorithm receives as input a frame with text or with a NL. Interested readers may refer to our previous research [4] on how images with text can be separated from images without text. In our current implementation, we work with frames of size 1,024 x 1,024. The 2D HWT is run for *NITER* iterations on the image to detect horizontal, vertical, and diagonal changes and store them in three corresponding *n* x *n* change matrices: HC (horizontal change), VC (vertical change), and DC (diagonal change), as shown in Figure 1.

In our current implementation, *NITER* = 2. Each *n* x *n* change matrix (*n* = 256 in our case, because *NITER* = 2) is binarized so that each pixel is set to one of the two values: v_1 and v_2, as shown in Figure 2. In our current implementation, v_1=0 and $v_2 = 255$. The binarized matrices are combined into a 256 x 256 result change set of 2D points $S = \{(i,j) | \alpha HC[i,j] + \beta VC[i,j] + \gamma DC[i,j] \geq \theta,$ where $\alpha + \beta + \gamma = 1$. In Figure 3 (right), the members of S are marked as white pixels.

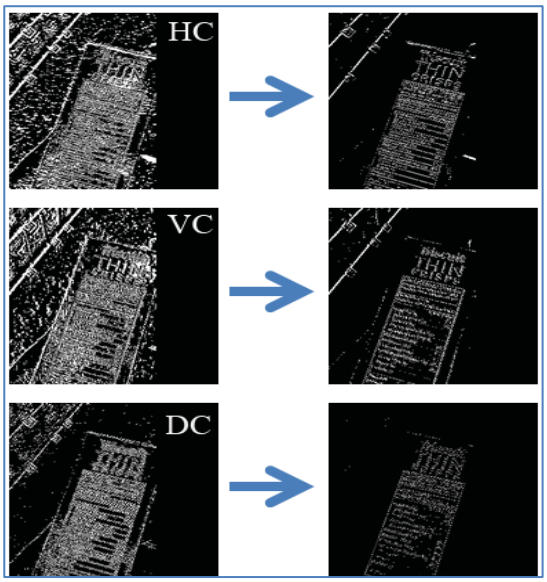


Figure 2. Binarization of HC, VC, and DC matrices

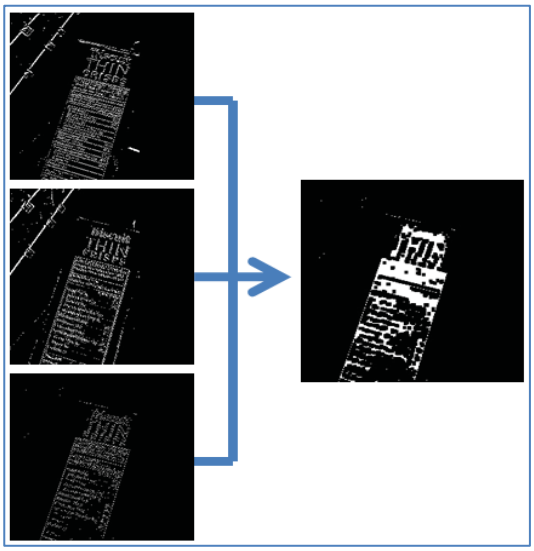


Figure 3. Combining wavelet matrices into result matrix

Once the result change set S is obtained, the convex hull algorithm [5] is used to find a minimum area rectangle bounding the polygon defined by S, as shown in Figure 4 (right). The text skew angle is computed as the skew angle of this rectangle, where the true north is 90 degrees.

We experimentally observed that the DC wavelets tend to detect the presence of text better than the HC and VC wavelets. This may be due to the fact that printed text has more diagonal edges than horizontal or vertical ones as compared to other objects in the image such as lines or

graphics. Consequently, in computing the 2D points of S we set $\alpha=\beta=0.2$ and $\gamma=0.6$.

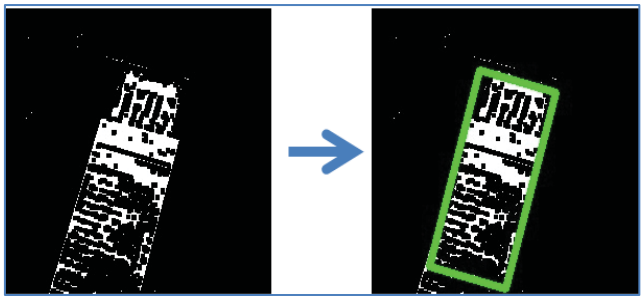


Figure 4. Text skew angle computation

```
1.  FUNCTION DetectTextSkewAngle(Img, N, NITER)
2.  [AVRG, HC, VC, DC] = 2DHWT(Img, NITER);
3.  Binarize(HC, n); Binarize(VC, n); Binarize(DC, n);
4.  FindSkewAngle(HC,VC,DC, (N/2 NITER));

5.  FUNCTION Binarize(Matrix, N, Thresh=5, v1=255, v2=0)
6.  For r = 0 to N
7.    For c=0 to N
8.      If Matrix[r][c] > Thresh Then
9.        Matrix[r][c]=v1;
10.     Else
11.       Matrix[r][c]=v2;
12.     End If
13.   End For
14. End For

16. FUNCTION FindSkewAngle(HC, VC, DC, n, α, β, γ, θ=255)
17.  S = {};
18.  For r = 1 to n Do
19.    For  c = 1 to n Do
20.      PV = α *HC[r][c]+ β*VC[r][c]+ γ*DC[r][c];
21.      If PV ≥ θ Then
22.        S = S ∪ (r, c);
23.      End If
24.    End For
25.  End For
26.  return TextSkewAngle(FindMinAreaRectangle(S));
```

Figure 5. Algorithm's pseudocode

Figure 5 gives the pseudocode of our algorithm. The algorithm takes as input a 2D image of size N x N. If the size of the image is not equal to an integral power of 2, as required by the 2DHWT, the image is padded with 0's. The third argument, *NITER*, specifies the number of iterations for the 2D HWT.

In Line 2, we apply the 2D HWT to the image for *NITER*, which, as stated above, in our current implementation is equal to 2. Our Java source of the **2DHWT** procedure is publicly available at [14]. The **2DHWT** returns an array of four n x n matrices *AVRG*, *HC*, *VC*, and *DC*. The first matrix contains the averages while *HC*, *VC*, and *DC* record horizontal, vertical and diagonal wavelet coefficients.

On line 3, the matrices *HC*, *VC*, and *DC* are binarized in place. Lines 5-14 give the code for the **Binarize** procedure. On line 4, a call to **FindSkewAngle** is made. As shown in lines 16-26, **FindSkewAngle** takes three n x n matrices *HC*, *VC*, and *DC* and the α, β, γ parameter values used in computing the 2D points of S.

On line 17, the set of change points is initialized. On lines 18-20, three corresponding values from *HC*, *VC*, and *DC* are combined into one *PV* value (line 20) using the formula $\alpha HC[i,j] + \beta VC[i,j] + \gamma DC[i,j]$. If this value clears the threshold θ that defaults to 255, the 2D point (i, j) is added to the set of 2D points on line 22.

On line 26, the algorithm first calls the procedure **FindMinAreaRectangle** that uses the convex hull algorithm to find a minimal area rectangle around the set of points found in lines 18–24 [5] and then calls the procedure **TextSkewAngle** that returns the value of the text skew angle using the true north as 90 degrees.

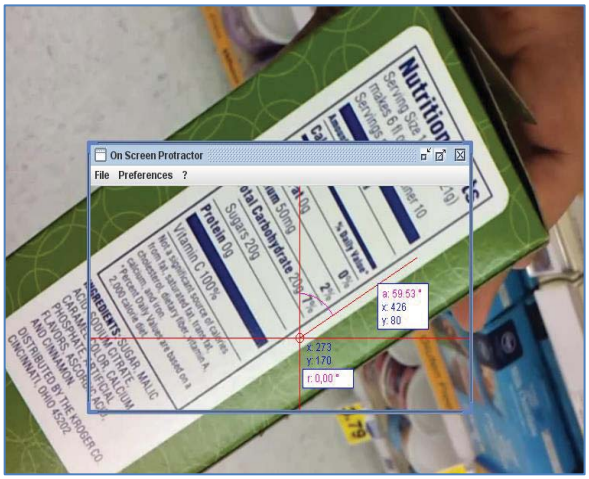


Figure 6. Ground truth text skew angle estimation

IV. Experiments

The text skew angle detection experiments were conducted on a set of 607 still images of nutrition labels from common grocery products. To facilitate data sharing and the replication of our results, we have made our images publicly available [15]. The still images used in the experiments were obtained from the 1280 x 720 videos of common grocery packages with an average duration of 15 seconds. The videos were recorded on an Android 4.3 Galaxy Nexus smartphone in Fresh Market, a supermarket in Logan, UT. All videos were recorded by an operator who held a grocery product in one hand and a smartphone in the other. The videos covered four different categories of products viz. bags, boxes, bottles, and cans. Each frame was manually classified as sharp only if it was possible for a person to read the text in the nutrition label.

We implemented our algorithm in Java (JDK 1.7) and compared the performance of our algorithm with the algorithms of Postl [5] and Hull [6], frequently cited in the literature on text skew angle detection. Since we were not able to find publicly available source code of either algorithm, we also implemented both in Java (JDK 1.7) to make the comparison more objective. In the tables below, we use the terms *Algo 1*, *Algo 2*, and *Algo 3* to refer to our algorithm, Postl's algorithm, and Hull's algorithm, respectively.

We ran three algorithms on all 607 images and logged the processing time for each image. The ground truth for the text skew angle was obtained from two human volunteers who used an open source protractor program [16] to manually estimate the text skew angle, as shown in Figure 6. Table I records the average processing time in milliseconds. Table II

records the median text skew angle detection error where the error is calculated as the absolute difference between a given text skew angle and the ground truth of the human evaluators.

Table I. Processing time in milliseconds

	Algo 1	Algo 2	Algo 3
Time (ms)	341.37	6253.02	5908.18

Table II. Median error in angle estimation

	Algo 1	Algo 2	Algo 3
Median error	4.62	68.85	20.92

V. Results

Table I shows *Algo* 1 has an average processing time of 341.37 ms, which is significantly faster than *Algo* 2 and *Algo* 3. This can be attributed to the fact that *Algo* 1 has no image rotation whereas *Algo* 2 and *Algo* 3 rotate either entire images or individual pixels of the image by various angles to find a match. For the sake of objectivity, it should be noted that *Algo* 2 and *Algo* 3 were originally designed to work inside document scanners where lighting conditions are near perfect and the text skew angles are expected to be relatively small. In vision-based NL scanning, neither the lighting conditions nor the text skew angle constraints are feasible. As Table II shows, *Algo* 1 has a lower median error rate than either *Algo* 2 or *Algo* 3. Specifically, Algo 1 has a median error of 4.62 whereas *Algo* 2 and *Algo* 3 have median rates of 68.85 and 20.92, respectively.

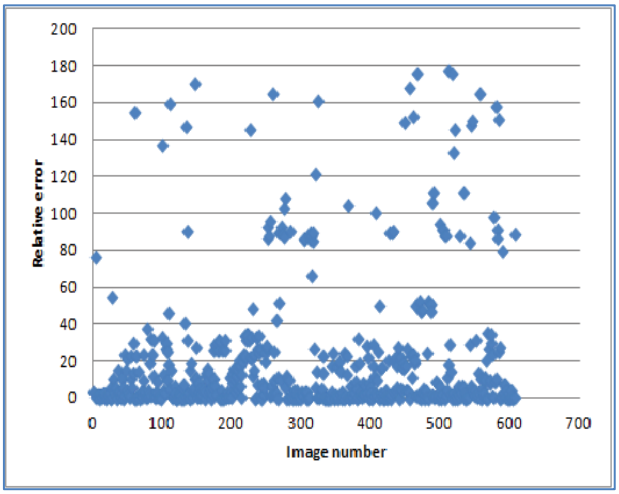


Figure 7. Error dispersion plot for Algo 1

.

Figures 7, 8 and 9 give the error dispersion plots for each algorithm. The horizontal axes in these figures record the number of images from 1 to 607 and the vertical axes record the text skew angle error. In all three figures, the zero line is the ground truth, i.e., zero deviation from the ground truth. Figure 7 shows that *Algo* 1 was less error prone on the sample of images used in the tests than either *Algo* 2 or *Algo* 3 in that most of the points are closer to the zero line and fewer points farther away than in the graphs for *Algo* 2 (Figure 8) and *Algo* 3 (Figure 9). A visual comparison of Figures 8 and 9 indicate that *Algo* 3 has a stronger clustering of points on or around the horizontal axis, which suggests that it was less error prone than *Algo* 2 on the sample of selected images, as is verified in Table II.

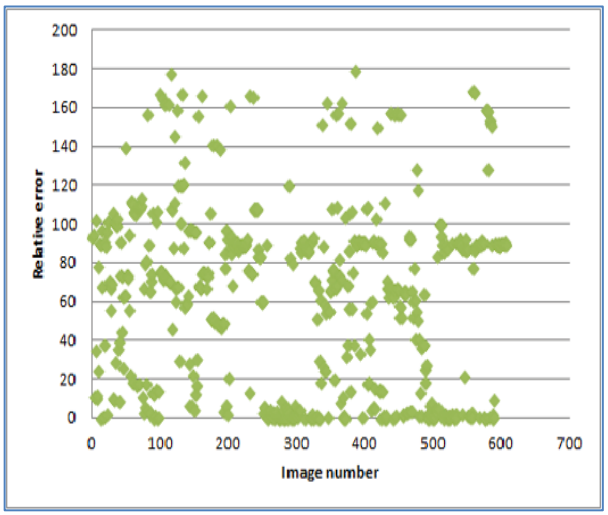


Figure 8. Error dispersion plot for Algo 2

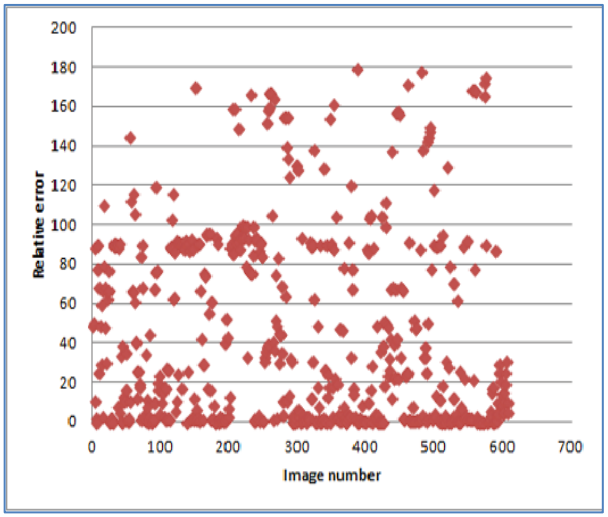


Figure 9. Error dispersion plot of Algo 3

Inadequate lighting conditions, light reflections, and irregular product shapes have caused problems for all three algorithms. For example, Figure 10 shows an image where our algorithm, *Algo* 1, has deviated from the ground truth by more than 20 degrees. The ground truth skew angle estimated by the human evaluators on the image in Figure 10 (right) was 66.29 degrees whereas the actual text skew angle returned by *Algo* 1 is 90.88 degrees. Note that the light reflections both above and inside the nutrition label resulted in point outliers and the subsequent error in the minimum area rectangle identification.

VI. Conclusions

We have proposed and implemented a text skew angle detection algorithm for vision-based NL scanning on mobile phones. The algorithm is designed to work with realistic images in which no assumptions can be made about lighting conditions, reflections, or the magnitudes of text skew angles.

The algorithm utilizes the 2D Haar Wavelet Transform (2DHWT) to effectively reduce the size of the images before processing them, thereby reducing the processing time. The algorithm takes an NL image and applies several iterations of the 2D HWT to compute horizontal, vertical, and diagonal change matrices. The values of these matrices are used to label the corresponding image pixels as text and non-text. The

convex hull algorithm is used to find a minimum area rectangle containing all text pixels [5]. The text skew angle is computed as the rotation angle of the minimum area rectangle found by the convex hull algorithm.

Figure 10. Text skew angle detection error

As a result of our experiments, we observed that the diagonal change matrix with the 2D diagonal wavelets tends to be more effective in text localization. We plan to conduct more experiments to verify this tendency of the 2D HWT in the future. A comparative study of our algorithm with the algorithms of Postl [6] and Hull [7], two text skew angle algorithms frequently cited in the literature, showed our algorithms to be faster and less error prone in vision-based scanning of NLs. This conclusion should be interpreted with caution because this comparative study reported in this paper was executed on a sample of 607 images. We plan to do further experiments on larger and more diverse image samples. We, by no means, rule out that in different domains with better lightning conditions or stricter constraints on text skew angle magnitudes our algorithm may not perform as well. We are reasonably certain, however, that our algorithm is faster than the algorithms by Postl [6] and Hull [7] because it does not rotate either images or individual pixels.

Another conclusion is that the classic text skew angle detection algorithms designed for document scanners do not work well in real world domains such as vision-based nutrition label scanning when no even illumination of documents can be assured. Such algorithms also implicitly assume that the documents have mostly printed text with little graphics. Another limitation of these algorithms is the fact that they utilize image rotation techniques to calculate either horizontal or vertical projection profiles to determine the text skew angle, which may not be suitable for real time video processing. and would be very inefficient if implemented to work on a mobile platform.

Our future work will focus on the integration of blur detection into vision-based NL scanning so that blurred images are automatically filtered out from the processing stream. Another future research objective is to couple the output of our algorithm with OCR engines to extract text from localized NLs. In our previous work, a greedy spellchecking algorithm was developed to correct OCR errors in vision-based NL scanning [17]. However, improving text skew angle detection may eliminate the need for spellchecking altogether without lowering the OCR rates.

References

[1] Kulyukin, V., Kutiyanawala, A., Zaman, T, & Clyde, S. "Vision-based localization & text chunking of nutrition fact tables on android smartphones." In *Proc. of the International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV 2013)*, pp. 314-320, ISBN 1-60132-252-6, CSREA Press, Las Vegas, NV, USA.

[2] Fog B.J. "A behavior model for persuasive design," In *Proc. 4th International Conference on Persuasive Technology*, Article 40, ACM, New York, USA, 2009.

[3] Kulyukin, V. and Zaman, T. "An Algorithm for in-place vision-based skewed 1D barcode scanning in the cloud." In *Proc. of the 18th International Conference on Image Processing and Pattern Recognition* (IPCV 2014), pp. 36-42, July 21-24, Las Vegas, NV, USA, CSREA Press, ISBN: 1-60132-280-1.

[4] Kulyukin, V. and Blay, C. "An algoritm for mobile vision-based localization of skewed nutrition labels that maximizes specificity." In *Proc. of the 18th International Conference on Image Processing and Pattern Recognition* (IPCV 2014), pp. 3-9, July 21-24, 2014, Las Vegas, NV, USA, CSREA Press, ISBN: 1-60132-280-1.

[5] Freeman, H. and Shapira, R. "Determining the minimum-area encasing rectangle for an arbitrary closed curve." *Comm. ACM*, 1975, pp.409-413.

[6] Postl, W. "Detection of linear oblique structures and skew scan in digitized documents." In *Proc*. of *International Conference on Pattern Recognition*, pp. 687-689, 1986.

[7] Hull, J.J. "Document image skew detection: survey and annotated bibliography," In J.J. Hull, S.L. Taylor (eds.), *Document Analysis Systems* II, World Scientific Publishing Co., 1997, pp. 40-64.

[8] Bloomberg, D. S., Kopec, G. E., and Dasari, L. "Measuring document image skew and orientation," *Document Recognition* II (SPIE vol. 2422), San Jose, CA, February 6-7, 1995, pp. 302-316.

[9] Kanai, J. and Bagdanov, A.D., "Projection profile based skew estimation algorithm for JBIG compressed images", *International Journal on Document Analysis and Recognition*, vol. 1, issue 1, 1998, pp.43-51.

[10] Papandreou, A. and Gatos, B. "A novel skew detection technique based on vertical projections." In *Proc. of International Conference on Document Analysis and Recognition* (ICDAR), pp. 384-388, Sept. 18-21, 2011, Beijing, China.

[11] Li, S.T., Shen, Q.H., and Sun, J. "Skew detection using wavelet decomposition and projection profile analysis." *Pattern Recognition Letters*, vol. 28, issue 5, 2007, pp. 555–562.

[12] Shivakumara, P., Hemantha Kumar, G. ., Guru, D. S., and Nagabhushan, P. "Skew estimation of binary document images using static and dynamic thresholds useful for document image mosaicing." *In Proc. of National Workshop on IT Services and Applications* (WITSA 2003), pp.51-55, , Feb 27–28, New Delhi, India, 2003,

[13] Nievergelt, Y. *Wavelets Made Easy*. Birkäuser, Boston, 2000, ISBN-10: 0817640614.

[14] Java implementation of the 2DHWT procedure. https://github.com/VKEDCO/java/tree/master/haar.

[15] Online database for sharp NL images. https://usu.box.com/s/9zk660t5h1g0dmw4pjj1x1yp6r7zovp3.

[16] Open source onscreen protractor program. http://sourceforge.net/projects/osprotractor/

[17] Kulyukin, V., Vanka, A., and Wang, W. "Skip trie matching: a greedy algorithm for real-time OCR error correction on smartphones." *International Journal of Digital Information and Wireless Communication* (IJDIWC): vol. 3, issue 3, pp. 56-65, 2013. ISSN: 2225-658X.

Encryption and Data Management Architecture to Protect Biometric Security

Obaidul Malek[a], Rabita Alamgir[a], Laila Alamgir[b], and Mohammad Matin[c]
Center for Biometrics and Biomedical Research,VA[a], Howard University, DC[b], and University of Denver, CO[c]

***Abstract*—In this paper, a novel symmetric encryption algorithm and its data management architecture to protect biometric security and privacy is proposed. Unlike current biometric encryption, the proposed method uses cryptographic keys in conjunction with extracted MultiBiometrics to create cryptographic bonds. To further enhance the security protection and to improve authentication accuracy, a data management architecture is being developed. The proposed method is being tested on images from three public databases: the "Put Face Database", the "Indian Face Database", and the "CASIA Fingerprint Image Database Version 5.1". The performance of the proposed solution has been evaluated using the Equal Error Rate (EER) and Correct Recognition Rate (CRR). The experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Biometric encryption, data management, MultiBiometrics, security, and unlinkability attack.**

I. Introduction

With the unprecedented growth of biometric systems, concerns about biometric security are the crucial issues for the 21st century. Not only does the biometric template (i.e. features) contain the unique and sensitive physiological and behavioural traits of an individual, it is also unary, and cannot be revoked or reissued if compromised. The features extracted from the biometric traits are stored in the database during enrollment in order to compare and authenticate the legitimacy of the subject of interest. This comparison also performs in the unencrypted domain, since the authentication accuracy can be largely influenced by a small variation in the feature properties if it takes place in the encrypted domain. Therefore, concerns about the security protection of biometric features are of paramount importance in the exploration of biometric systems. Ideally, the security of the template can be accomplished using mathematical algorithms that must be difficult to decrypt by the unintended recipients. In addition, a template protection algorithm should be irreversible, robust, and revokable [1-3].

In this paper, a novel symmetric biometric encryption algorithm and its Data Management Architecture (DMA) is proposed that protects the stored and dynamic biometric templates against security, privacy, and unlinkability attacks. In contrast to current biometric encryption, this method uses cryptographic keys in conjunction with extracted MultiBiometrics to create cryptographic bonds, called "$BioCryptoBond$". To further enhance security and privacy protection and to improve authentication accuracy, a multilayered DMA architecture is also proposed. The theoretical foundation of the proposed method along with the model evaluation and experimental results have also been presented in this paper.

The remainder of the paper is organized as follows: Section II presents the literature review and prerequisites; the detailed analysis and algorithmic formulation of the proposed biometric encryption and authentication systems are presented in Section III; Section IV presents the biometric Data Management Architecture (DMA); Section V studies the possible attacks; experimental results and discussions are given in Section VI; and finally, the conclusions are presented in Section VII.

II. Literature Review

The proposed method is based on the Biometric Encryption (BE). In addition, a data management architecture has been proposed to enhance the security of biometric features.

A. Cavoukian et al. [4] proposed a biometric encryption algorithm based on facial biometrics. In their method, the system is composed of two distinct stages: i) Creation of a watch list consisting of a maximum of five patrons; and ii) Implementation of a biometric encryption module and released keys for each of the top match patrons, as well as the generation of a match alert by the system that is then reviewed by administrators. This BE method can achieve an optimal FAR at the cost of FRR. In their self-exclusion model, K. Martin et al. [5] proposed a biometric encryption algorithm based on a small subset of the subject's facial biometric database. The authors here used feature vectors for their key binding process to secure the cryptographic key. It is a novel model, and they achieved low FAR at the cost of FRR. K. Nandakumar et al. [6] proposed a fuzzy vault scheme where the authors derived a multibiometrics template from multiple templates of a single user. They used fingerprint minutiae points and iriscodes templates and transformed them into a multibiometrics vault. Here, the

authors didn't properly address the challenges associated with the unlinkability attacks.

A. Ross et al. [7] proposed a visual cryptography method to protect the privacy of the biometric templates. In their method, an image is decomposed into two host images and stored in the two central databases. The original image can only be revealed when two images are available simultaneously. C. Lee et al. [8], introduced a two factor method for generating cancelable fingerprint templates using local minutia information. The transformation function is associated with the randomly generated PIN number, which is used to change the biometric template. The major drawback of this method is that it has a tradeoff between performance and changeability. D. Maio et al. [9], implemented a multihashing algorithm, where the scores of selected fingerprint matchers and those obtained by a face authenticator are combined. Furthermore, to enhance the performance of this system, a random subspace based method is further combined with the similarity matching scores. However, this method is computationally expensive. A. Teoh et al. [10] proposed a multispace random projection method. The distance-preserving property of multispace random projection is analyzed based on a normalized inner product, and an approximately zero EER is achieved; however privacy and changeability are the main concerns in this paper.

A. Prerequisites

This section introduces some fundamental concepts related to the proposed method before getting into its detailed analysis.

1) Unlinkability Attack: True anonymity requires unlinkability, which is the ability of the system to perform multiple operations anonymously. Unlinkability also implies the incapability of retrieving the information of one individual based on the information of another. Additionally, it is the measurement of the strength of a system (or object) to be unlinkable. Unlinkability is the core property for any authentication process, making it difficult for a third party recipient to be associated with the unauthorized information.

2) Data Segmentation and Foreign Key: Data segmentation is known as data grouping, and is a branch of the data mining operation. It is the process of extracting and segmenting data in such a way that the system would be able to factorize the data, reduce its volume, and classify it. It is also capable of storing data in different locations of the database system with the intention of increasing overall system performance and security. However, prior to carrying out a data segmentation analysis, appropriate care should be taken to decide which key parameters could be used for the segmentation process. This is especially important because the failure of biometric segmentation means that the system was not able to detect useful biometric features. Indexing is another technique that can be used in the data segmentation process to put segmented data in order. In addition, a foreign key is used in conjunction with the indexing process to create a link and establish a relationship amongst segmented data within database system.

III. Biometric encryption and Authentication

The biometric templates $h(t)$ created from the images received from the output of the Sequential Subspace Estimator (SSE) studied in [11],[12] are the desired templates. These biometric templates along with reference pointers will be stored (enrollment-Fig. 4) in the databases for the authentication process. The security and confidentiality of the stored and dynamic biometric features are dependent on their level of protection from security, privacy, and unlinkability attacks. Therefore, the objective of this section is to present a secure, robust, and reliable encryption and authentication algorithm for these protections.

A. Encryption

The cryptographic architecture of this method is designed to deal with two categories of people: the authorized user and the subject (target). A detailed system diagram and processing method for generating user and subject biometric encrypted bonds $BioCryptoBond$ is presented in Fig. 1.

The algorithmic architecture and formulation for creating $BioCryptoBond$ bonds have been stated below.

1) $BioCryptoBond_u$*:* The steps that are involved in creating the user cryptographic bond $BioCryptoBond_u$, are stated below:

(i) Extract and compute orientation angle (θ) (Fig. 2) from received user fingerprint features.

(ii) The tensor operation is performed on the user filtered fingerprint biometric template $h(t)$ (i.e. minutiae points) as a function of orientation angle.

(iii) Output from the tensor operation is converted into an orthogonal matrix Π.

(iv) A digital random key K^u for the user is generated (Fig. 1) and fused with an orthogonal matrix of vectors Π, creating the user cryptographic bond, $BioCryptoBond_u$. This cryptographic bond binding process can be formulated as follows:

$$\begin{aligned} \mathbb{T} &= \theta \times \digamma(s) \\ \Pi &= \mathbb{T}_{or} \\ BioCryptoBond_u &= \Pi \times K^u \end{aligned} \tag{1}$$

where $\mathbb{T}$ is the output from the tensor operation; the subscript $_{or}$ is the orthogonal operator; and $\digamma(s)$ is the fourier transformation of $h(t)$.

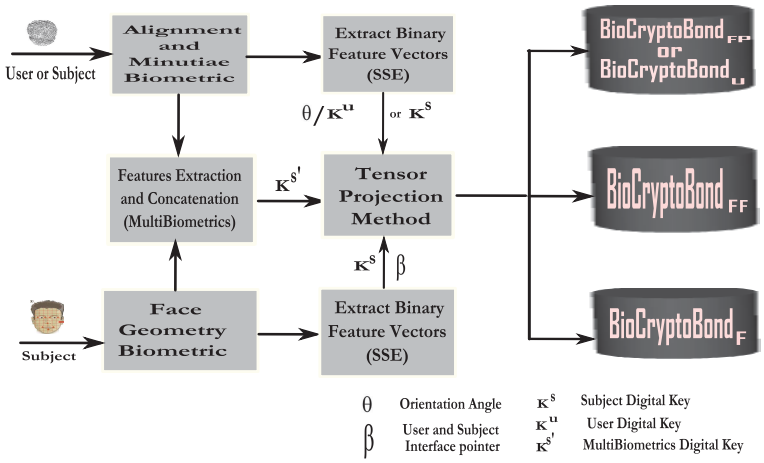


Fig. 1: System Architecture -BioCryptoBond

2) $\underline{BioCryptoBond_{FP}}$*:* The steps that are involved in creating the subject cryptographic bond $BioCryptoBond_{FP}$ using (subject) $h(t)$ fingerprint features (i.e. minutiae points (Figs. 1 and 2)) are stated below:

(i) Randomly generated key K^s is transformed into the orthogonal matrix Π_{fp}.

(ii) Matrix Π_{fp} is fused with the filtered fingerprint output $h(t)$, creating the cryptographic bond $BioCryptoBond_{FP}$.

This bond binding process can be formulated as follows:

$$\begin{aligned} \Pi_{fp} &= K^s_{or} \\ BioCryptoBond_{FP} &= \Pi_{fp} \times \digamma(s) \end{aligned} \tag{2}$$

3) $\underline{BioCryptoBond_{F}}$*:* The steps that are involved in creating the subject cryptographic bond $BioCryptoBond_F$ using subject facial biometrics (i.e. facial area; size and relative positions of eyes and lips (Figs. 1 and 3)) are stated below:

(i) An arbitrary interface pointer β is received. This interface pointer is generated by the system upon successful user authentication.

(ii) Tensor operation is performed as a function of β on received filtered facial biometric features $h(t)$.

(iii) Output of tensor operation is converted into the orthogonal matrix Π_f.

(iv) Matrix Π_f is fused with the same randomly generated digital key K^s used to create the subject's $BioCryptoBond_{FP}$ bond. This bond binding process can be formulated as follows:

$$\begin{aligned} \mathbb{T} &= \beta \times \digamma(s) \\ \Pi_f &= \mathbb{T}_{or} \\ BioCryptoBond_F &= \Pi_f \times K^s \end{aligned} \tag{3}$$

4) $\underline{BioCryptoBond_{FF}}$*:* The steps that are involved in creating the subject cryptographic bond $BioCryptoBond_{FF}$ using MultiBiometrics (the fusion of facial and fingerprint biometrics) are stated below:

(i) Subject filtered fingerprint and facial biometrics are concatenated (or fused), and a MultiBiometrics template is created.

(ii) Concatenated matrix or MultiBiometrics is converted to orthogonal matrix Π_{ff}.

(iii) Orthogonal matrix is fused with a randomly generated digital secret key $K^{s'}$ (Fig. 1) and a $BioCryptoBond_{FF}$ bond is created.

This bond binding process can be formulated as follows:

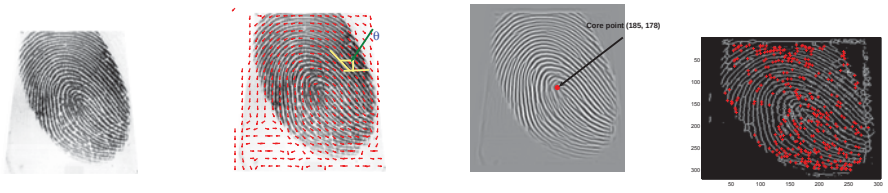
$$\begin{aligned} c(t) &= c[h_1(t) + h_2(t)] \\ \mathbb{C}(s) &= \digamma(c(t)) \\ \Pi_{ff} &= \mathbb{C}_{or}(s) \\ BioCryptoBond_{FF} &= \Pi_{ff} \times K^{s'} \end{aligned} \tag{4}$$

where $h_1(t)$ and $h_2(t)$ represent filtered outputs for fingerprint and facial biometrics, respectively; c(t) represents the concatenate operation; and $\mathbb{C}(s)$ represents the fourier transform of the concatenate operation.

B. Authentication

In the case of the user authentication process, fingerprint biometric features received from the authorized user are combined with the cryptographic bond, $BioCryptoBond_u$, and the digital secret key is released. In this stage, authentication is performed to ensure the legitimacy of the user and to release the user secret key K^u. The user authentication process is shown in Fig. 5.

During the user authentication cycle, the same algorithmic operation stated in Eq. (1) is performed on the live user fingerprint features, generating the matrix Π.



(a) Original Image (b) Ridge Orientation (c) Core Point (d) Minutiae Points

Fig. 2: Fingerprint Biometrics –Features Extraction

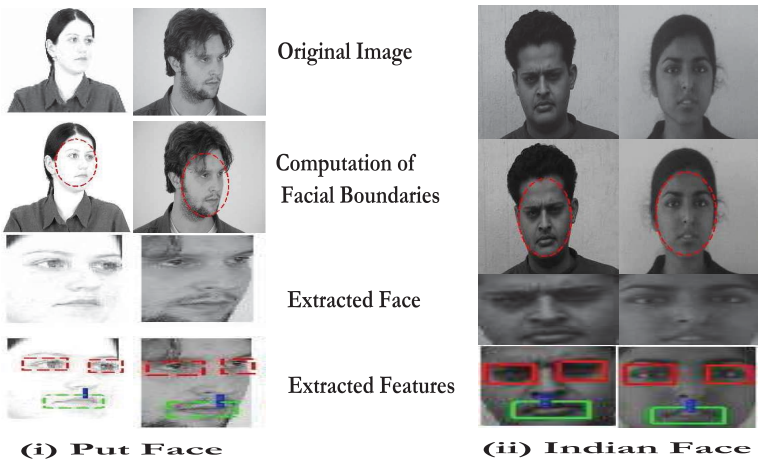


Fig. 3: Facial Biometrics –Features Extraction

Afterwards, Π is combined with the previously stored $BioCryptoBond_u$ to release the key K^u. This process can be stated as follows:

$$\begin{aligned} \mathbb{T} &= \theta \times \digamma(s) \\ \Pi &= \mathbb{T}_{or} \\ K^u &= \Pi \times BioCryptoBond_u \end{aligned} \tag{5}$$

Once the secret key is activated and released, it is hashed with the user biometric features and generates the reference pointers required to complete the final level of authenticity of the user. Afterwards, this reference pointer along with the secret key allows the user to access the system. This process can be formulated as follows:

$$\begin{aligned} \mathbb{R}^u &= \mathbb{H}[K^u \times \mathbb{I}_p] \\ \text{Required Info} &= \mathbb{R}^u\ [dB_u] \end{aligned} \tag{6}$$

where $\mathbb{R}^u$ is the user reference pointer.

Finally, a triggering signal is processed to initialize an interface between user and subject, if the user authenticity is found positive. This interface allows the user to prepare a system platform for receiving inputted subject biometric features. The system also releases an interface pointer β, which is required to ensure that the system is ready to enroll, authenticate, and release the subject information in the presence of the legitimate user and the subject of interest.

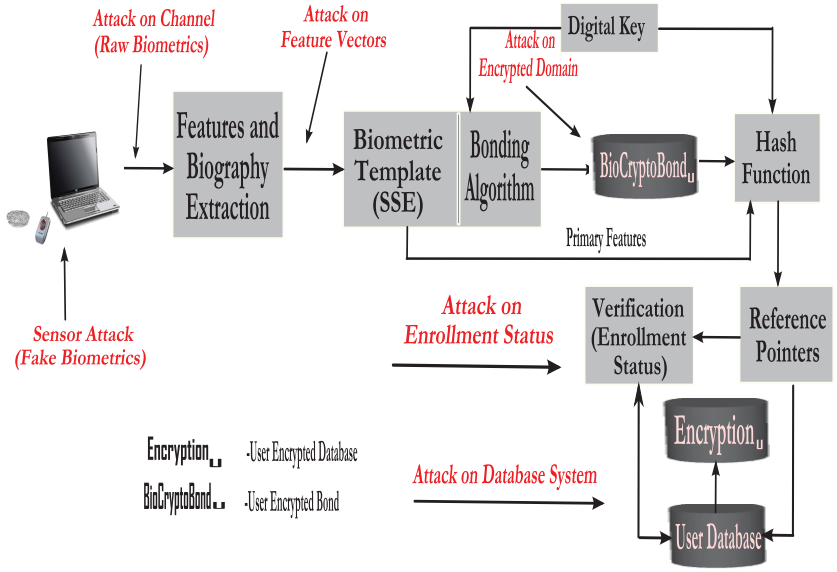
IV. Biometric Data Management Architecture

The main objective of the biometric DMA architecture is to enhance the security protection of the stored and dynamic biometric features. In this case, a multilayered and MultiBiometrics data management architecture has been proposed to protect the users' and the subjects' biometric features. The cryptographic bonding architecture and its process have already been presented in previous sections. The hash function, Hot-Key, and segmentation processes are integral parts of this management architecture, and are presented in the following subsections.

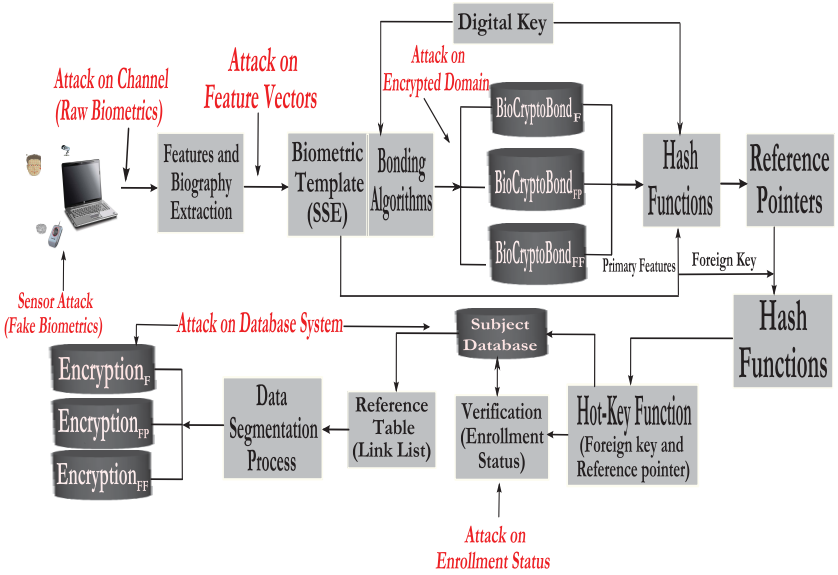
A. Hot-Key Function

The Hot-Key function is the compound function key generated from a combination of the reference pointer and foreign key. The foreign key ($\mathbb{F}$) is a 32-bit digital key generated from the primary (indexed) biometric features.

The first step of this process is to create a reference pointer for the user (or subject) from the system generated $32 - bit$ digital key hashed with the primary



(a) User Enrollment



(b) Subject Enrollment

Fig. 4: Enrollment and Possible Attacks

biometric features. This reference pointer is used to store the encrypted features (enrollment) in the user databases. In this case, the biographical information is stored in the user database dB_u and the encrypted biometric features are stored in the $Encryption_u$ database. This reference pointer is used to establish a relationship between user databases.

In the case of a subject database, the reference key is generated in the same way as the user. This reference key is hashed with the indexed foreign key generated from the subject biometric features. The output of this hash function is called the Hot-Key (Φ_{hk}) function, and its main objective is to create an extra-layer of security for the stored and dynamic biometric features of the subject. A description of the subject multilayered and MultiBiometrics authentication process is not included here, but successful user authentication in the presence of the subject is required. The experimental result of this process has been included in Section VI. The subject's biographical information and biometric features are stored in the subject databases (dB_s, $Encryption_F$, $Encryption_{FP}$, and $Encryption_{FF}$), and the generated

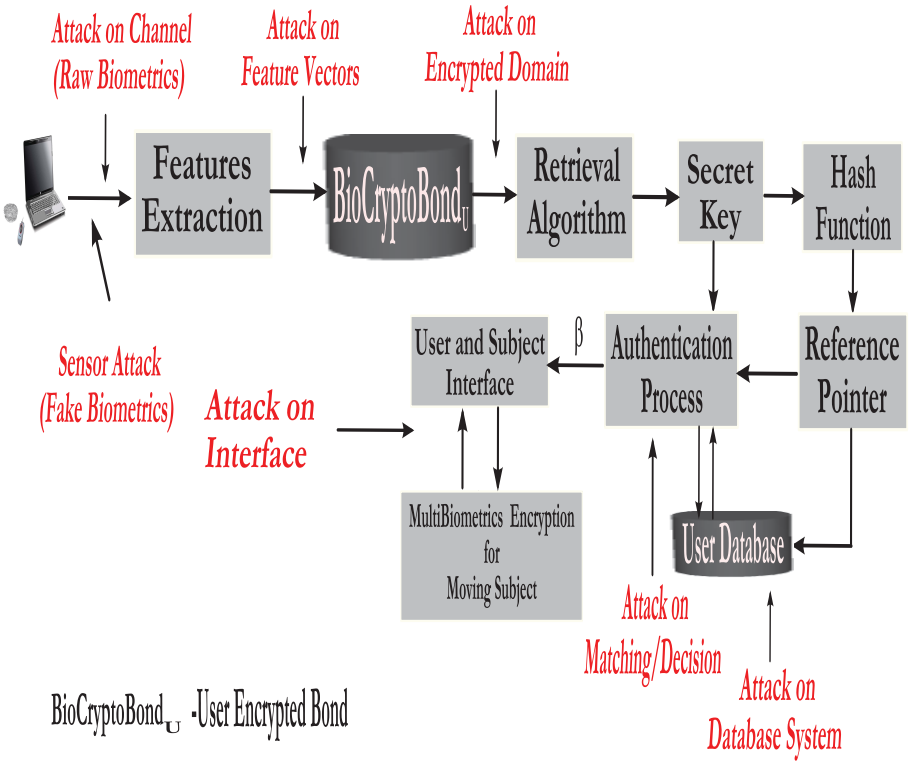


Fig. 5: User Authentication Process and Possible Attacks

reference pointers are used to create a link between the subject databases using the reference table and data segmentation process. The system architecture of this methodology is presented in Figs. $4(b)$ and 6.

B. Segmentation Process

The main purpose of the segmentation process is to cluster (or group) the subject biometric features and biographical information based on the address pointers created as shown in Figs. $4(b)$ and 6. This clustering process uses the index biometric features of face, fingerprint, and MultiBiometrics (fusion of face and fingerprint). In this process, a hash key function in conjunction with the composite foreign key and reference pointer are implemented to construct the Hot-Key algorithm. The data segmentation technique along with the Hot-Key algorithm are employed in order to develop a secure biometric DMA architecture. A reference table is created, which serves as a link list (or address pointer) for keeping reference addresses and locating records stored in the subject databases. The relationship between subject databases is also maintained by the reference table as shown in Figs. $4(b)$ and 6.

V. Possible Attacks -Authentication

The possible attacks on the user authentication process are shown in Fig. 5 (attacks on enrollment process are not included here). The subject authentication is dependent on a successful user authentication process, and the types of attacks on it are the same. The experimental results of the subject (and user) authentication process are presented in Section VI. User fingerprint biometrics is being used during the authentication process. If attackers are able to intervene at the sensor or communication channel, they still won't be able to access the system, since the biometric features need to be transformed as a function of the user fingerprint orientation angle before the authentication process occurs. Even if the attackers are able to obtain access to the system through a single point, they won't have the right to access other users' or subjects' information, since the biometric systems are unlinkable and the physical presence of the subject is required along with user in order to retrieve the biometric and biographic information. The databases are protected by multilayered encryption, hence single point access ability won't allow the attacker to retrieve unauthorized information or distinguish the identity of the subject (or user) from the received information.

Furthermore, in this DMA architecture, the biometric information is segmented, and reference pointers are used to establish a link between these segmented biometrics. In this method, it is not possible to obtain the original biometrics from these reference pointers and vice versa. As well, it is not possible to know the individual's iden-

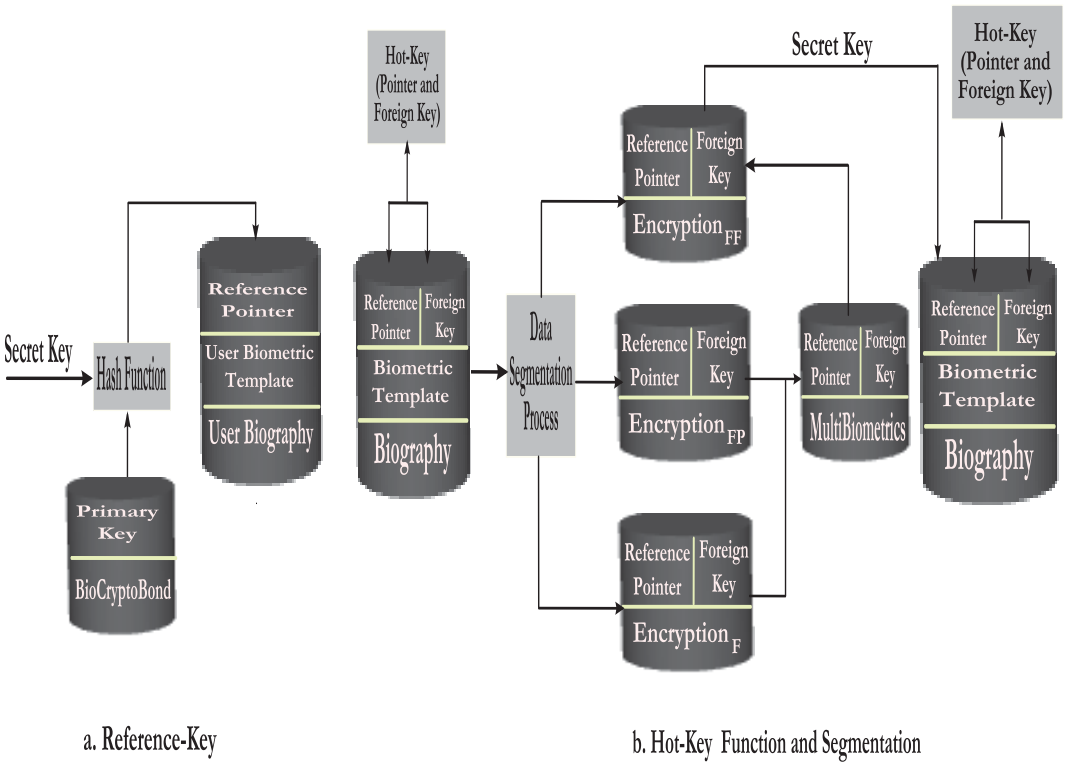


Fig. 6: Biometric Data Management Architecture

tity or construct (or guess) the original biometric features of an individual from the segmented biometrics stored in the databases. Databases (or information) are segmented and transformed, complete authorized processing is required in order to access the system. Therefore, this system is invincible to unlinkable attacks, and imposters cannot retrieve data based on information found in other parts of the system.

VI. Experimental Results and Discussions

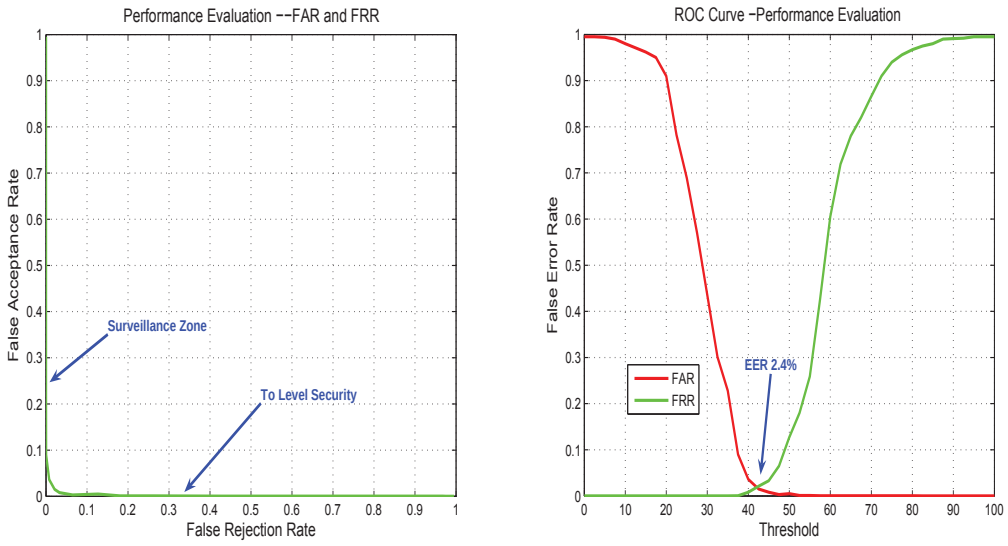
In this experiment, two types of authentication processes have been performed: i) user authentication, and ii) authentication and retrieval of the subject's information. Therefore, the experimental results and resultant analysis presented here are based on these two processes.

A. User Authentication

In this experiment, two user encrypted databases were created for 30 users, then 10 users with fingerprint biometrics from the public database "CASIA Fingerprint Image Database Version 5.1". The encrypted database set comprised of 30 users has been used for authorized user fingerprints, and the encrypted database set comprised of 10 users has been used for imposter fingerprints. The main objective of this process is to authenticate the legitimacy of a user. An evaluation of the verification performance of the encryption method is also presented in this paper. In this case, each of the 40 users have been tested against the encrypted users' biometrics stored in the databases. The performance of the verification process has been evaluated based on the False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER). The experimentation results of this verification process have been recorded in Table I, and the graphical outcome of the FAR, FRR, and ROC are presented in Fig. 7.

TABLE I: *Performance Evaluation in (%) - FAR, FRR, and EER*

Database	Person	FAR	FRR	EER
CASIA Fingerprint	40 users	1.20	3.50	2.40
Put Face	20 Subjects	1.45	8.50	4.70
Put Face	40 Subjects	1.75	9.30	5.10
Indian Face	10 Subjects	1.50	4.60	3.10
Indian Face	20 Subjects	1.86	5.40	3.45



(a) FAR and FRR (b) ROC Curve

Fig. 7: User Fingerprint Biometrics -Verification Process

B. Authentication and Retrieval of the Subject's Information

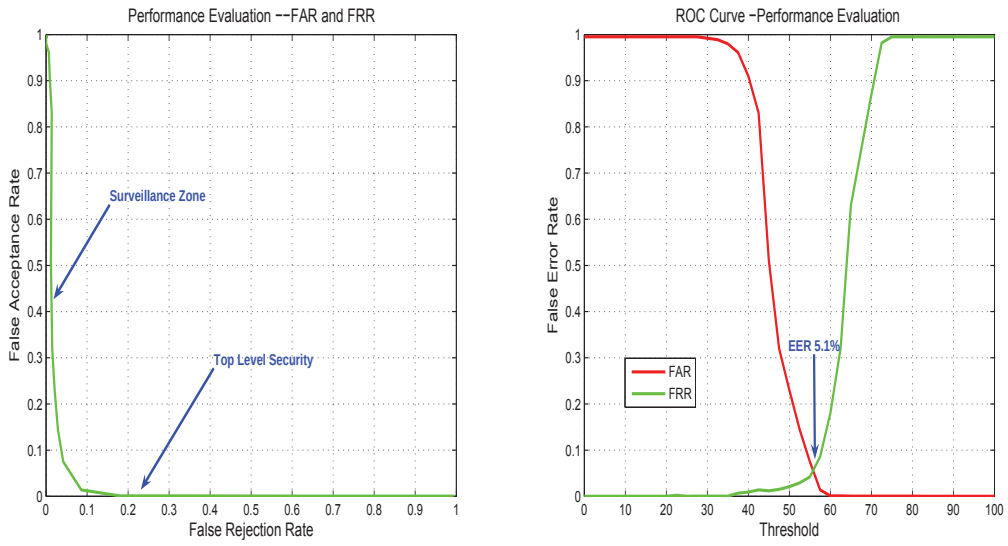
The performance of the proposed method has been evaluated based on the images of these public databases: "Put Face Database" [13], "Indian Face Database" [14], and "CASIA Fingerprint Image Database Version 5.1". The experimental results presented here are based on the authorized users' authentication processes using fingerprint biometric features in the presence of the respective subjects. In this experiment, two sets of encrypted user databases and four sets of encrypted subject databases have been created from the original image databases. This experiment tested whether the subject's information could be retrieved from their biometric database by legitimate and illegitimate users, with or without the presence of the subject. The percentages of Correct Recognition Rate (CRR), False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER) have been determined, and experimental results have been recorded. The experimental results of this authentication (verification) process have been recorded in Table I. Simulation results of the legitimate (and illegitimate) user verification process for retrieving subject biometrics in the presence (and without the presence) of the respective subjects are shown in Figs. 8. As well, the performance of the identification process (CRR) has been recorded in Table II.

TABLE II: *Performance Evaluation in (%) - CRR*

Database	10-Subject	20-Subject	40-Subject	Average
Put Face	–	91.68	88.35	90.02
Indian Face	96.20	95.55	–	95.87

VII. Conclusions

A biometric system contains attributes that exclusively represent an individual's identity. These properties don't change and are difficult to lose or fake. The main concern for the exploration of the biometric system is to protect the security and privacy of these biometric features. This cannot be neglected, otherwise it can revert the overall process in the opposite direction, since the damage to this system is irreversible and may cost more than the system it is used for. The proposed MultiBiometrics $BioCryptoBond$ is secure and efficient, since a 1.5% FAR has been achieved at the cost of 4.6% FRR. According to the experimental results, the proposed method is also found to be robust with a promising EER of 3.1%. As well, BE along with the DMA architecture provide multilayered protection against security, privacy, and unlinkability attacks for the dynamic and stored biometric features in the databases. It can be concluded that the encryption method presented in this paper is heuristic, robust, and reliable in comparison to its counterparts. This is because, unlike other key binding encryption systems, BE along with the biometric DMA architecture are implemented to enhance security protection and improve authentication accuracy. Without



(a) FAR and FRR

(b) ROC Curve

Fig. 8: MultiBiometrics Encryption –Put Face Database(40 subjects)

a successful authentication process, neither the secret key nor the biometric features can be retrieved independently from the encrypted bonds. In addition, even if the secret key or the transformed biometric features are intercepted at any point of operation by the imposter, the original biometric features are not obtainable. Finally, top level security has also been maintained for subject biometric templates, since the retrieval of the subject's biometric features would also require the physical presence of the subject along with a successful user authentication process.

REFERENCES

[1] A. Jain and A. Kumar, "Biometric of next generation: An overview", To Appear in Second Generation Biometrics Springer, Aug. 2010.

[2] A. Menezes, P. Oorschot, and S. Stone, "Handbook of applied cryptography", CRC press, Jun. 1996.

[3] A. Cavoukian and A. Stoianov, "Biometric encryption: A positive-sum technology that achieves strong authentication, security and privacy", Information and Privacy Commissioner Ontario, Mar. 2007.

[4] A. Cavoukian and T. Marinelli, "Privacy-protective facial recognition: biometric en- cryption proof of concept", Information and Privacy Commissioner, Ontario, Canada. Nov. 2010.

[5] K. Martin, H. Lu, F. Bui, K. Plataniotis, and D. Hatzinakos,"A biometric encryption system for the selfexclusion scenario of face recognition", IEEE Systems Journal: Special Issue on Biometrics Systems, vol. 3, no. 4, pp. 440-450, Mar. 2009.

[6] K. Nandakumar and A. Jain, "Multibiometric template security using fuzzy vault, Proceedings of 2nd IEEE International Conference on Biometrics: Theory, Applications, and Systems", pp. 1-6, Jun. 2008.

[7] A. Ross, K. Nandakumar, and A. Jain, "Handbook of multibiometrics", Springer, Chapter 2, Mar. 2006.

[8] C. Lee, J.Choi, K. Toh, S. Lee, and J. Kim, "Alignment-free cancelable fingerprint templates based on local minutiae information", IEEE Transactions on Systems, Man and Cybernetics, Part B, vol. 37, no. 4, pp. 980-992, Oct. 2007.

[9] D. Maio and L. Nanni, "Multihashing, human authentication featuring biometrics data and tokenised random number: a case study", Elsevier Neurocomputing, vol. 69, no. 1, pp. 242-249, Jun. 2006.

[10] A. Teoh, T. Connie, O. Ngo, and C. Ling, Remarks on biohash and its mathematical foundation, Information Processing Letter, no. 4, pp. 145-150, Sep. 2006.

[11] O. Malek, A. Venetsonoupoulous, D. Androutsos, and L. Zhao, "Sequential subspace estimator for biometric authentication", ELSEVIER-Neurocomputing, vol. 148, pp. 294-309, Jan. 2015.

[12] O. Malek, A. Venetsonoupoulous, D. Androutsos, and L. Zhao, "Subspace state estimator for facial biometric verification", IEEE Proceedings of The International Conference on Computational Science and Computational Intelligence, Las Vegas, USA, vol. 1, pp. 137-143, Mar. 2014.

[13] A. Kasiski, A. Florek, and A. Schmidt, "The PUT Face Database", Image Processing and Communications, vol. 13, no. 3-4, pp. 5964, Aug. 2008.

[14] V. Jain and A. Mukherjee, "The Indian Face Database", http://www.cs.umass.edu/ vidit/IndianFaceDatabase/, Jun. 2002.

ON INFORMED CODING AND HOST REJECTION FOR COMMUNICATION OVER INKJET PRINT-AND-SCAN CHANNELS

Joceli Mayer
Digital Signal Processing Lab - LPDS - Department of Electrical Engineering
Federal University of Santa Catarina - UFSC
Florianopolis, Santa Catarina, Brazil, CEP 88040900
email: joceli.mayer@lpds.ufsc.br

Steven J. Simske
Print and Content Delivery Lab
Hewlett-Packard Labs
USA
email: steven.simske@hp.com

ABSTRACT
This paper describes novel approaches to achieve robust communication over inkjet print-and-scan (IPS) color channels. The IPS color channel poses even greater challenges than the laser printer-and-scan channel due to the resulting mixing and spreading of the ink dots. We propose a novel informed coding and two host color rejection approaches, one based on a novel color rejection and another on a whitening filter, to deal with the aforementioned inkjet printer distortions. A substitutive spatial domain embedding is proposed to enable robustness optimization using the proposed informed coding. Analyses and examples are provided to evaluate the performance enhancement on robustness and transparency achievable by the proposed approaches.
KEY WORDS
Watermarking Methods and Protection, Hardcopy Color Watermarking, Information and Document Security.

1 Introduction

Robustly decoding side information transmitted over color printed media is very challenging [1, 2, 3, 4] due to various non-linear distortions from the color print-scan channel, particularly those originated by ink spreading and mixing existing in inkjet printers and other disturbances from the coated media properties, optical, mechanical and scanning sensor responses [5]. Several techniques have been proposed for hardcopy watermarking over print-scan channels. The technique in [6] conveys information by modulating the angle of oriented periodical sequences embedded into image spatial blocks, while dedicating one block to embed synchronisation information. However, it exploits only the luminance and considers neither informed coding nor the color channel properties. It achieves a resulting payload of 40 bits per page. The method in [2] modulates information into the luminance image phase spectrum with differential quantization index modulation. It exploits the printer halftoning to estimate the rotation, and achieves a payload of hundreds of bits for monochromatic images. The method in [3] relies on adaptive block embedding into the DFT (Discrete Fourier Transform) magnitude domain. Each block is classified into smooth or texture type and a different embedding method is applied to each block type. The Hough transform is used to detect the printed image boundaries for watermark synchronization. The approach is robust to the print-scan channel and to rotation, providing a total payload of 1024 bits with a bit error rate (BER) around 15% for monochromatic images. In [7] a circular template watermark is embedded into the Fourier transform magnitude to facilitate inversion of rotation and scaling after the print-scan process. Another template watermark is embedded in spatial domain to invert translations. The message watermark is embedded in the wavelet domain. This technique achieves a payload of about 135 bits using an error correction algorithm, resulting in a BER of 1.5%, it is also designed for monochromatic images.

The approaches just described do not deal specifically with the color channel distortions, as they embed information only in the luminance channel. In order to exploit all color channels for modulation, it is necessary to investigate new efficient strategies to address distortion from the color print channel. The work in [8] proposes to use the Discrete Fourier Transform to embed information into the red component, while the approach in [9] performs frequency domain informed embedding using halftoning modulation. Halftoning modulation may provide high capacity but requires control of the printer driver to bypass the printer processing and halftoning, which is reportedly difficult. Color hardcopy approaches based on embedding in frequency domain [8] or in halftoning [9] are more efficient for laser print-and-scan channels.

For inkjet print-and-scan channels, however, additional techniques to the aforementioned approaches are required in order to deal with the stronger channel distortions due to the IPS ink mixing and spreading. Some work in this direction is proposed in [10] and the informed coding approach is inspired on the work of Professor Max H. Costa [11] named "Writing on Dirty Paper". The informed coding approach has been exploited by [12] for single channel (monochromatic) image modulation achieving very high payload for noise, filtered and compressed channels, however it is not designed to convey information on color host images over the IPS color channel.

This paper provides detailed discussions and analyses on the robustness and detection improvements due to the novel proposed informed coding and color rejection techniques designed to deal with color IPS channels.

2 Proposed Improvements

2.1 Information Embedding

Consider an m-bit information to be conveyed by a digital color image I which is deployed as inkjet printed media. We propose to embed this information through a set of K color dot patterns (a randomly generated sparse matrix of color dots) from a set of N available patterns, $P_i, i = 1, \ldots, N, (N \geq K)$. The set of K patterns is uniquely represented by an unordered set $S = \{k_1, k_2, \ldots, k_K\}$, where $k_i \neq k_j$ for $i \neq j$. The resulting watermarked document I_w is

$$I_w = I \oplus \sum_{i \in S} P_i \tag{1}$$

where the operation $\oplus$ represents a substitution embedding, instead of the traditional additive embedding. The pixels of the image I are replaced by the pixels of the pattern whenever color pattern dots exist, as illustrated in Fig. 1(a) at left side.

This approach provides a transparent embedding only for small size square dots and using inkjet printers. This is because the inkjet printing process helps to mix and hide the embedding dots as illustrated in Fig. 1(a) at right side and in Fig. 3, provided that the dot size is smaller than 6x6 pixels in resolutions of 600 dpi/ppi for printing and scanning. Section 3 provides performance evaluations using dot size of 4x4 pixels at 600 dpi/ppi resolutions for printing and scanning.

2.2 On the Choice of Patterns for Informed Coding

We have experimentally verified that the ink spreading of the dots differs depending on the color of the embedding pattern dots, k, and on the colors of the pixels surrounding the embedded dots in the host image, as indicated in Fig. 1(b). This is due to specific ink chemical properties of the inkjet printer cartridges. As one result, the detection performance based on correlation is considerably better for certain color combinations of the embedding patterns and the host backgrounds. Hence we propose to use the correlation as the robustness metric for selecting the best pattern of one unique color from L alternative patterns depending on the pixels colors of the image host background.

For example, the robustness estimates illustrated in Fig. 2 for only 4 background colors using an HP4280 printer, indicate that higher robustness is achieved by embedding magenta pattern dots over a cyan image background rather than over a yellow background. To estimate this robustness, $R(k, b)$, of embedding a pattern of color k in a background of color b, a training is performed with the specific printing system. We need to estimate all combinations of colors of the patterns k and backgrounds b. We may use a fewer background representative colors, by clustering sets of background colors using Euclidean distance, aiming to reduce the computational complexity but also reducing the performance of the informed coding approach. Thus, for a given color k, a pattern P_{i^*k} from a set of L equivalent patterns is chosen at embedding by maximizing

$$P_{i^*k} = \max_{i=1,L} \Phi\{P_{ik}, I_w\} \tag{2}$$

where $\Phi\{P_{ik}, I_w\}$ represents the average robustness for all D pattern dots, after IPS channel, between the D color dots of pattern P_{ik} and the U pixels surrounding these dots in the host image. Thus,

$$\Phi\{P_{ik}, I_w\} = \frac{1}{UD} \sum_{d=1}^{D} \sum_{[r,s] \in \mathcal{N}_d} R(k, Color(I_w[r, s])) \tag{3}$$

where $[r, s] \in \mathcal{N}_d$ represents the set of U pixels at the neighbourhood of the embedded dot d and $Color(I_w[r, s])$ is the color of the watermarked image at location $[r, s]$.

As we propose to use L alternative patterns, for each of the N patterns, which convey the same message, both encoder and decoder must share a secret key ϕ in order to generate the same set of LN patterns. The price paid for this additional flexibility and performance optimization is the increase of the total number of required detections (time complexity) to decode the message from N to LN detections. The information payload (m bits) achievable by the proposed modulation is determined by the number K of color patterns per region, the total number N of patterns and the number N_R of embedding regions (time division modulation). Therefore, using K patterns per region from a database of N patterns and considering N_R embedding regions, we can achieve a payload of at least m bits:

$$N_R \log_2 \binom{N}{K} \geq m \tag{4}$$

where $\binom{N}{K} = N!/[K!(N-K)!]$

2.3 Color Rejection to Reduce Host Interference

Since each pattern in a region is set with a unique color k, the detection metric in (6) is computed for this pattern disregarding pixels with any other color. The rejection of pixels of other colors in the received image is based on a statistical distance as follows. Lets represent a pixel of color k as a vector with CMYK color components: $\boldsymbol{B} = [B_c\, B_m\, B_y\, B_k]^T$. Assume this pixel color is a random variable distributed as $\boldsymbol{B} \sim N(\boldsymbol{\mu_k}, \boldsymbol{C_k})$. After transmitting Z pixels of such color k over a given IPS channel, we estimate the mean vector and covariance matrix as $\boldsymbol{\mu_k} = \frac{1}{Z}\sum_i \boldsymbol{B}_i$ and $\boldsymbol{C_k} = \frac{1}{Z}\sum_i \boldsymbol{B}_i\boldsymbol{B}_i^T - \boldsymbol{\mu_k}\boldsymbol{\mu_k}^T$. Thus, when testing a pattern of color k, an unknown color pixel $\boldsymbol{X}$ of the received image is accepted only if its Mahalanobis distance to the color k, defined by

$$d_{M_k}(\boldsymbol{X}) = \sqrt{(\boldsymbol{X} - \boldsymbol{\mu_k})^T \boldsymbol{C_k}^{-1} (\boldsymbol{X} - \boldsymbol{\mu_k})}, \tag{5}$$

is smaller than to the other colors: $d_{M_k}(\boldsymbol{X}) < d_{M_i}(\boldsymbol{X}), i \neq k$. Notice that this criterion of rejection is

optimal for Normal distributed pixels as assumed here and verified in the experiments. For instance, suppose we decide to embed 4 patterns, each with one unique color from CMYK. Then, to detect the yellow pattern, we reject the other (CMK) colors generating the modified image I_{wR} before correlation.

2.4 Pattern Detection and Message Decoding

The K color patterns embedded in a region of the color image are assumed to be printed into paper and digitized using an image scanner before detection. The detection of a pattern P_{ik} is based on the correlation (performed in the frequency domain, for speed, after rejecting the other colors $\neq k$, as described above) between the observed watermarked image after the print-scan channel and the known patterns, which are locally generated with the help of a secret key ϕ.

After transforming both images to the HVS (Hue, Value and Saturation) color model, a $LoG[m,n]$ whitening filter (Laplacian of Gaussian) of dimension 3×3 is employed to decorrelate the host signal. This operation can be summarized as an average correlation over the channels of the received color image I_{wR} and known pattern P_i represented in HVS color model as

$$C_i = \frac{1}{3} \sum_{k=H,V,S} \mathcal{F}^{-1}\{\mathcal{F}(I_{wR_k}[m,n] * LoG[m,n]) \cdot \mathcal{F}(P_{i_k}[-m,-n] * LoG[-m,-n])\} \quad (6)$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote, respectively, the fast 2D direct and inverse discrete Fourier transforms. The operator $*$ represents the 2D linear convolution which performs the whitening filtering. After LN detections, the K indexes of patterns P_i corresponding to the highest correlations C_i are selected to compose the set S of indexes required to decode the message m. As an example, consider the CMYK colors ($K = 4$) and redundancy factor $L = 3$. LN correlations are performed for each of the 4 colors and the indexes of the K patterns with highest correlation peak value for each color are stored. The selected set of K indexes are associated by a look-up table to the message and the decoder will follow the same encoder assignment of message and indexes.

2.5 Expected Performance of the Proposed Detection Metric

Recalling the Central Limit theorem, where a large sum of independent small disturbances tends to follow a Normal distribution, by modeling the detection metric (correlation Ci in (6)) as $\sim N(\mu, \sigma^2)$, we find the probability of missing a pattern $P_P(\tau)$ by

$$P_P(\tau) = P_{FN}(\tau) + P_{FP}(\tau) = \quad (7)$$
$$= \frac{P_1}{\sqrt{\pi}} \int_{\frac{\mu_1 - \tau}{\sqrt{2\sigma_1^2}}}^{\infty} exp(-t^2)dt + \frac{P_0}{\sqrt{\pi}} \int_{\frac{\tau - \mu_0}{\sqrt{2\sigma_0^2}}}^{\infty} exp(-t^2)dt$$

where $P_{0,1}, \mu_{0,1}, \sigma_{0,1}^2$ parameters are estimated from data and are respectively the prior probabilities, means and variances of the detection statistics of regions with no pattern (unmarked hypothesis H_0) and regions with pattern (marked hypothesis H_1). The optimal detection threshold τ is determined and employed to decide the hypotheses based on the observed metric $C_i \underset{H_1}{\overset{H_0}{\lessgtr}} \tau$ by solving

$$(\sigma_0^2 - \sigma_1^2)\tau^2 + 2(\mu_0\sigma_1^2 - \mu_1\sigma_0^2)\tau + \quad (8)$$
$$+\sigma_0^2\mu_1^2 - \sigma_1^2\mu_0^2 + 2\sigma_0^2\sigma_1^2 \ln(\frac{\sigma_1 P_0}{\sigma_0 P_1}) = 0$$

which minimizes the probability of missing a pattern, $P_P(\tau)$. Henceforth, as LN patterns are tested in order to find the K embedded patterns in each one of the N_R regions, the estimated probability of missing the entire message, Pe, is

$$Pe(\tau) = N_R((1-(1-P_{FN}(\tau))^K)+(1-(1-P_{FP}(\tau))^{LN})) \quad (9)$$

3 Experiments

The detection performance improvements are illustrated at Fig. 4. We observe an increase of 70% in μ_1 due to the informed coding after the IPS channel for a chosen pattern P_{i^*k} in a given background. Clearly, the detection performance is superior when proper embedding patterns are defined at embedding (informed coding). Moreover, by employing the color rejection, the correlation statistics μ_1 is increased by 15% while de deviation σ_1 is decreased by 50%, providing an huge gain on detection.

For a payload of 1035 bits, we need at least 23 bits per region, $NR = 45$ regions, a color host image larger than 2100×2100 pixels, $K = 4$ patterns per region and according to Eq. (4), it would be necessary to detect $N = 140$ patterns with $L = 3$ alternatives each (informed coding). By estimating the distribution parameters from IPS experiments, the resulting probability $P_{FN}(\tau)$ is about 4×10^{-10} and $P_{FP}(\tau)$ is about 3.9×10^{-10} for a given image. For this payload the estimated probability of missing the entire message is $Pe(\tau) = 45((1 - (1 - 4 \times 10^{-10})^4) + (1 - (1 - 3.9 \times 10^{-10})^{140 \times 3})) = 7.4 \times 10^{-6}$. This estimation shows that the techniques provide a very robust embedding for the IPS channel, which can be further improved by using an error correction code.

The performance is validated by computing the correlation metrics statistics (μ, σ) from a set of 50 watermarked and scanned images (sizes of about $1.5in^2$ at 600ppi/dpi resolutions), following by the estimation of the error probabilities. Some images are illustrated at Figure 3(a). After employing the proposed color rejection and informed coding techniques, the lowest performance case for $[Pe(\tau), P_{FN}(\tau), P_{FP}(\tau)]$ is improved from $[2.4 \times 10^{-5}, 2.1 \times 10^{-7}, 5.5 \times 10^{-8}]$, respectively, to $[4.3 \times 10^{-6}, 4.4 \times 10^{-8}, 9.7 \times 10^{-9}]$. The results indicted a

consistent improvement, for different messages and printers (HP5580 and HP4280), of at least 5 times in probability of detection even for the worst of the 50 cases.

Prior to printing the image, the resulting Peak Signal-to-Watermark Ratio, $PSWR = 20log_{10}\left(\frac{255\times W\times H}{\sqrt{|I-I_w|^2}}\right)$, is very high, $PSWR = 45$ dB, in average, where W and H are respectively the width and height of the images. The structural Similarity (SSIM) index is also high prior printing, $SSIM = 0.96$ in average.

Perceptual evaluation from 15 users indicates a transparent embedding to naked eyes from a normal distance (10 inches) to the printed page using about 400 dots of 4x4 pixels size per pattern of 300 $\times$ 300 pixels at 600 ppi/dpi resolutions. IPS channel hides those dots quite well due to ink spreading and mixing, as illustrated in Fig. 1 and Fig. 3. This setup provides a payload of 100 bits/in^2 at 600 ppi/dpi for IPS color channels with a small probability of missing the message ($\sim 10^{-5}$) and good transparency, a performance very competitive to watermarking techniques discussed in Section 1.

In all the experiments, a careful placement of the printed document in the scanbed resulted in a well aligned image. Notice that the correlation method is robust to any degree of translation; however, the performance may be affected if rotation occurs. In this case we apply the following automatic method based on a coarse alignment followed by a search rotation method to achieve a finer alignment: The corners and boundaries of the image are detected and a rotation is performed on the image to achieve a coarse alignment. Next a fine search is performed aimed to improve the alignment. This is achieved by computing the correlation of some blocks and use this information to select the angle that provided the highest correlation. This approach does not require any additional visual cues or special blocks specially design to recover synchronism. The approach has the drawback of requiring extra computational time to find the best rotation angle from a range of few degrees after the coarse alignment. Fig. 5 illustrates the correlation performance dependent on the rotation angle.

Transparency, payload, decoding speed and robustness are adjustable by using a different set of parameters N, K, N_R, L, dot width and number of dots per pattern for a given number m of embedding bits. Higher payload is achievable by increasing N with some impact on computational complexity and robustness. According to (3) the embedding has complexity proportional to the product N_RKLUD. In the experiments, the embedding required about a mean of 3.5 minutes for images of size 800 $\times$ 800 pixels while the decoding using (6) required a mean of 40 seconds using a non-optimized program.

4 Conclusions

This work provide improvements on communication over IPS color channels by proposing informed coding, optimal detection and host rejection techniques. These techniques mitigate the host interference as confirmed by the results and the analyses provided. The detection performance is evaluated with the proposed optimal detection threshold and the results illustrate the significant improvement on probability of detecting a transmitted message over the IPS channel. These approaches improve communication reliability over IPS channels allowing customization for various robustness, transparencies and decoding speed tradeoffs by choosing proper embedding pattern parameters.

References

[1] P. Bulan, G. Sharma, and V. Monga, Orientation Modulation for Data Hiding in Clustered-Dot Halftone Prints, IEEE Trans. on Image Processing, Vol. 19:8, 2010.

[2] Kaushal Solanki, Upamanyu Madhow, B. S. Manjunath, Shiv Chandrasekaran, and Ibrahim El-Khalil, Print and Scan Resilient Data Hiding in Images, IEEE Trans. on Information Forensics and Security, Vol. 1, No. 4, Dec. 2006.

[3] Dajun He and Qibin Sun, A Practical Print-Scan Resilient Watermarking Scheme, IEEE International Conference on Image Processing, 2005.

[4] Q. Li, I. J. Cox, Using Perceptual Models to Improve Fidelity and Provide Resistance to Valumetric Scaling for Quantization Index Modulation Watermarking, IEEE Trans. on Information Forensics and Security, pp. 127 - 139, 2007.

[5] P.V.K. Borges, Joceli Mayer, Ebroul Izquierdo, Robust and Transparent Color Modulation for Text Data Hiding, IEEE Trans. on Multimedia, Vol. 10:8, 2008.

[6] A. Keskinarkaus, A. Pramila, T. Seppänen, Image Watermarking with a Directed Periodic Pattern to Embed Multibit Messages Resilient to Print-Scan and Compound Attacks, Journal of Systems and Software v. 83, pp. 1715-1725, 2010.

[7] Anu Pramila, Anja Keskinarkaus, and Tapio Seppänen, Multiple Domain Watermarking for Print-Scan and JPEG Resilient Data Hiding, Proceedings of the 6th International Workshop on Digital Watermarking, 2007.

[8] Guo, Chengqing Xu, Guoai Niu, Xinxin Yang, Yixian Li, Yang, A Color Image Watermarking Algorithm Resistant to Print-Scan, IEEE International Conference on Wireless Communications, Networking and Information Security, 2010.

[9] Basak Oztan and Gaurav Sharma, Multiplexed Clustered-Dot Halftone Watermarks Using Bi-Directional Phase Modulation and Detection,

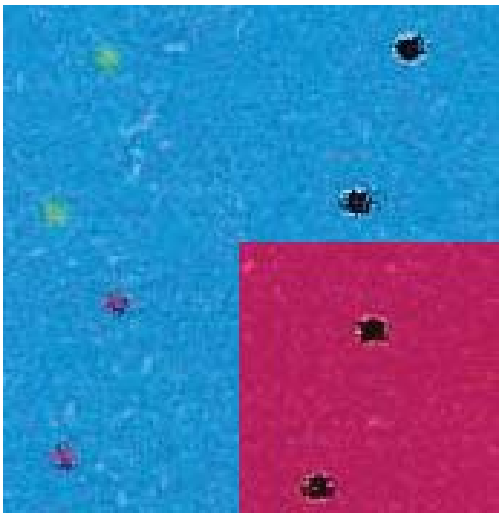
Proceedings of 2010 IEEE 17th International Conference on Image Processing, 2010.

[10] Joceli Mayer and Steven Simske, Informed Coding for Color Hardcopy Watermarking, 8th International Symposium on Image and Signal Processing and Analysis - ISPA, 2013.

[11] Max Henrique Machado Costa, Writing on Dirty Paper, IEEE Trans. on Information Theory, IT-29, 439-441, 1983.

[12] M. L. Miller, G. J. Doërr and I. J. Cox, "Applying Informed Coding and Informed Embedding to Design a Robust, High Capacity Watermark", IEEE Trans. on Image Processing, 13(6):792-807, 2004.

(a) The digital watermarked image and the image after IPS channel.



(b) Dot spreading on different backgrounds.

Figure 1. (a) The digital watermarked image detail (with 1150×850 pixels, corresponding in a real size of 1.9 in $\times 1.4$ in) with a cyan pattern. The original image, printed and scanned at 600 dpi/ppi, has 3200×2400 pixels. The digital domain is used only for embedding as the distribution media is the printed version where the embedding transparency is high. On the right is shown the printed and scanned watermarked image with a cyan pattern. (b) Detail (zoom) of the color dot pattern ink spreading and mixing for different color backgrounds.

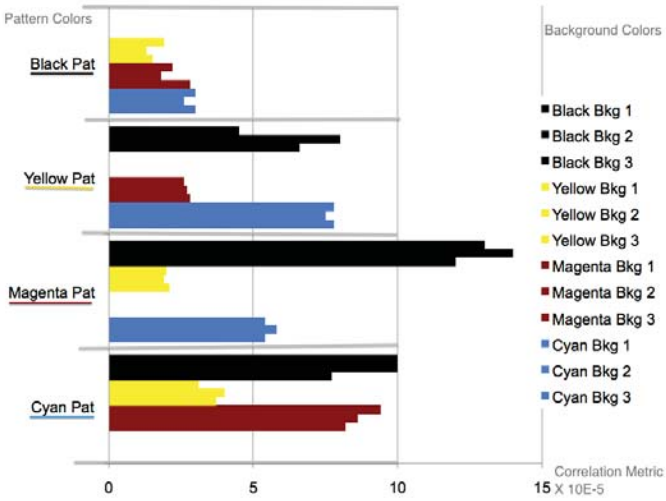
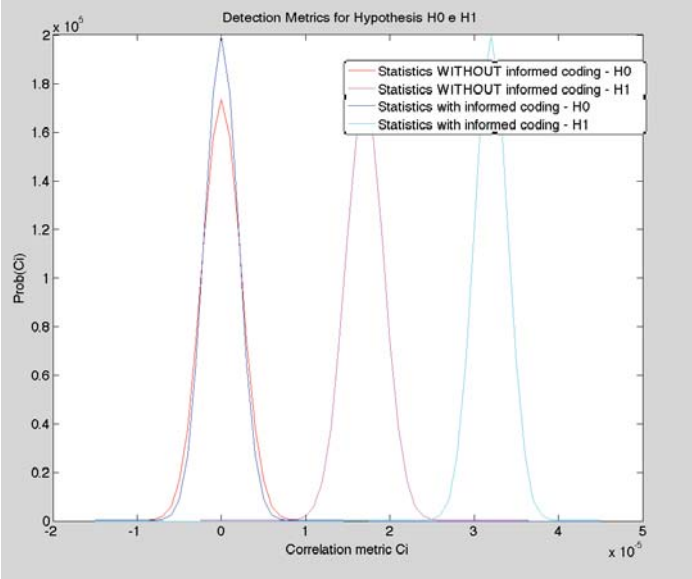


Figure 2. Expected robustness for 3 realizations of a color embedding pattern (Pat) for each (CMYK) color background (Bkg).

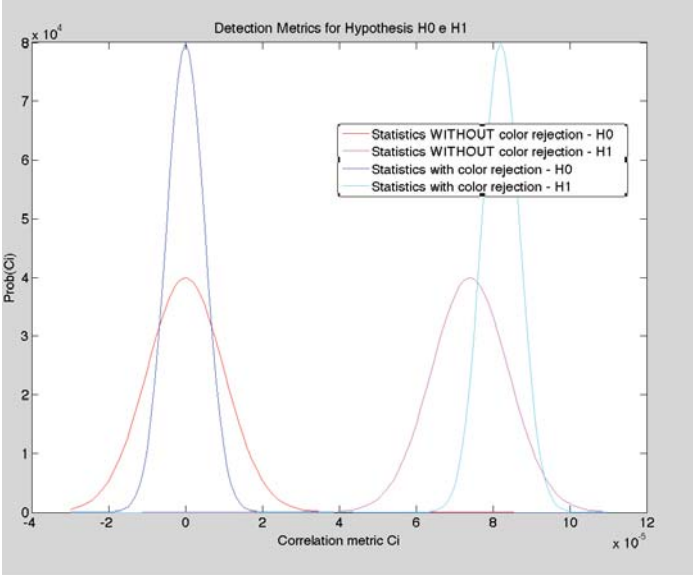
(a) Host images without watermark after IPS color channel.

(b) Watermarked images after IPS color channel.

Figure 3. (a) Original (no watermark) images of size $1.5in^2$ printed at 600 dpi and scanned at 600 ppi in HP5580. (b) Printed and watermarked images indicated high transparency when viewed from a distance of 10 inches. There are less than 1% of pixels marked as cyan dots of size of 4x4 pixels, which are barely seen by naked eye. Only with digital zoom is possible to notice the patterns.



(a) Performance of informed coding approach.



(b) Performance of color rejection approach.

Figure 4. (a) The correlation performance **with and without** informed coding: the mean μ_1 for the hypothesis H_1 is improved by 70%. (b) The correlation performance **with and without** color rejection: the μ_1 is improved by 15% and $\sigma_{0,1}$ are decreased by 50%.

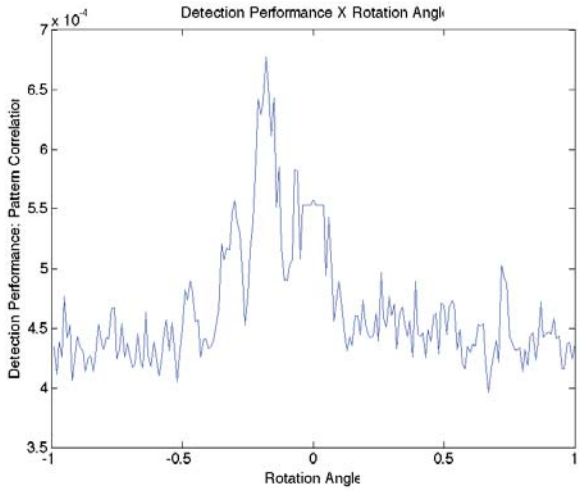


Figure 5. The correlation performance for a range of rotation angles. This approach enables to determine the best rotation angle automatically making the approach robust to angle rotation at the scanning process. The detection method based on correlation is already naturally robust to translation.

A Cryptographic, Discrete Cosine Transform and Frequency Domain Watermarking Approach for Securing Digital Images

Quist-Aphetsi Kester [1,2,3], Laurent Nana[1], Anca Christine Pascu[1], Sophie Gire[1],Jojo M. Eghan[3], Nii Narku Quaynor[3]

[1] Lab-STICC (UMR CNRS 6285), European University of Brittany, University of Brest, France
Kester.quist-aphetsi@univ-bret.fr / kquist@ieee.org
[2] Faculty of Informatics, Ghana Technology University College
[3]Department of Computer Science and Information Technology, University of Cape Coast

Abstract - *The growth in multimedia messaging, social networks, live streaming and in other applications have increased overtime. The rise in the engagement of Unmanned Ariel vehicles for agricultural, surveillance and deliver services has become a debate in its security and engagement in today's cyber physical space. Authentication and security between control units and these devices are of paramount importance. And also a fast and energy efficient algorithms are highly needed to ensure uncompromised situation of compunctions with these devices. Due to fictitious activities over communication channels by unauthorized users, security and authentication of such transmissions are needed to provide privacy, authentication and confidentiality. In this paper, we proposed a cryptographic encryption technique and a discrete cosine transform for encryption and compression of the transmitted image. We further engaged a frequency domain watermarking approach for the authentication of the digital images. These approaches were engaged to provide several security layers for transmitted image and at the, results showed to be very effective. The implementation in this work was simulated in MATLAB.*

Keywords: cryptography, discrete cosine transforms, UAVs, compression, watermarking

1 Introduction

The cyberspace today is challenged with security of transmitted and stored data. The future of computing looks forward to a full potential of sensor networks distributed across the geographical space and yielding a continuous massive data from which knowledge can be deduced from in shaping the way we perceive and adapt to our environment. This will help build and make the physical and the cyberspace be one with each other and resulting into an integrated and interdependent cyber-physical world. One can see these applications gradually evolving from the emergence of new directions in science with restless researchers aiming to get results for complex problems in today's world. Applications ranging from ubiquitous computing, internet of things, bionics etc. Brain to machine interfacing and brain to brain via device interfacing are gradually progressing in becoming a reality.

With all these advancements poses threats to mankind's security of control. Hence security approaches to securing devices in today's cyberspace has become key issue in deployment of remote control or Unmanned Ariel Vehicles. These devices first emerge with human autonomy over them and they were gradually given partial independence of autonomous behavior in time. The only interface between these devices and man is the visuals they obtain from afar. These visuals consist of transmitted valuable information such as coordinates, speed, payloads, signal strength etc. Hence a compromise situation or interception of transmitted visuals data can put the vehicle under threat and can expose it to being compromised by a third unauthorized party. And hence the safety and security of the commutations from these devices are key concern.

For most of the devices that depend on wireless networks and independent power sources, maximizing cryptographic approaches for them means putting more computational power load on them as well as demanding for more power source and memory for their processes. Hence an effective and efficient and easily implementable but a good layer of security approach is required in providing safety and security for these devices. I n contributing to the security developments and demands in these area, we proposed a cryptographic encryption technique to ensure confidentiality and a discrete cosine transform for the compression of the transmitted image. We further engaged a frequency domain watermarking approach for the authentication of the digital images. The paper has the following structure; section II Related works, section III is Methodology, section IV Results and analysis, and section V concluded the paper.

2 Literature Review

The demand for data security for multimedia image applications for use in streaming over secured and unsecured networks has risen over the years. Social network users are in high demand for both security for streaming data and still images in other to counter surveillance activities with the essence of ensuring privacy and security. Applications used for video communications by tech companies are gradually encrypting their streamed data in providing security for their clients. The discrete cosine transform (DCT) is a technique for converting a signal into elementary frequency components and widely used in image compression [1]. There is no so much concern with compression works digital images when there is some tolerance for loss, the compression factor can be greater than no loss tolerance. For this reason, graphic images can be compressed more than text files or programs [2]. PS, A. K. ih their work demonstrated a discrete wavelet technique in compressing the images using wavelet theory in VHDL, Verilog [3]. Xu, N. showed FFT approach for data compression that its histogram has a desired shape [4] and Klimesh, M et al showed the lossless image compression algorithm using FPGA technology [5]. Other works such as Glynn, E. F used Fourier analysis and Image processing technique [6] and Riet, p. shows Image compression Implementation using Fast Fourier Transform [7].]. Al-Haj, A. used imperceptible and a robust combined with DWT-DCT digital image watermarking algorithm to provide security for digital images. In their work, they approach engaged used watermarks digital image in combination of the Discrete Wavelet Transform (DWT) and the Discrete Cosine Transform (DCT). His evaluation results show that combining the two transforms improved the performance of the watermarking algorithms that were based solely on the DWT transform [8]. Tomar, Ravi et al in their work showed a robust watermarking technique to copyright an image solely based on Discrete Cosine Transform by embedding blockwise watermark against the noise, filtering and cropping attack. Their experimental results showed that the invisibility and security of the scheme was robust against signal processing [9].

3 Methodology

With our proposed approach, the plain image was encrypted based on the cryptographic approach that involved pixel displacement technique and then compressed before watermarked. Below is a block diagram that illustrates how the entire process engaged. The cryptographic approach is used to conceal the content of the image before the application of the discrete cosine function. The discrete cosine function is then used to compress the image. The watermark is then applied to the compressed image before transmission.

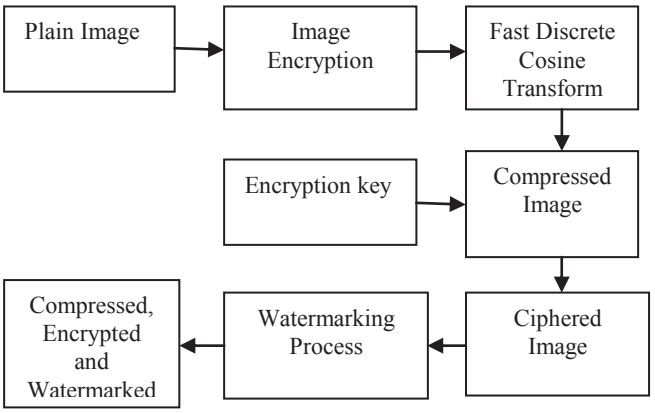


Figure 1. The image encryption, compression and watermarking process

With the proposed system the authentication of the source is verified before decompression takes place. The decompressed image is then decrypted by the approach. But there is an engagement of a symmetric key for the cryptographic process at the earlier phase that is for the image encryption process as well as the later stage which involved the decryption process.

3.1 The Encryption process

a) Import data from image and create an image graphics object by interpreting each element in a matrix.

b) Get the size of r as [c, p]

c)Get the Entropy of the plain Image

d) Get the mean of the plain Image

e)Compute the shared secret from the image

f) Engage SK for g) to q) using secret key value

g) Extract the red component as 'r'

h) Extract the green component as 'g'

i) Extract the blue component as 'b'

j) Let r =Transpose of r

k)Let g =Transpose of g

l) Let b =Transpose of b

m) Reshape r into (r, c, p)

n) Reshape g into (g, c, and p)

o) Reshape b into (b, c, and p)

p) Concatenate the arrays r, g, b into the same dimension of 'r' or 'g' or 'b' of the original image.

q) Finally the data will be converted into an image format to get the encrypted image.

The inverse of the algorithm will decrypt the encrypted image back into the plain image.

The secret key is obtained as follows:

$Sk = [(c \times p) + |(He \times 10^3)| + |(\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i)|] \bmod p$

Where c, p are dimension of the image and He is the entropy value of the image and *x* bar is the arithmetic mean for all the pixels in the image.

3.2 The Discrete Cosine Process

DCT transforms a time domain of a signal into its frequency components by only using of the real parts of the Discrete Fourier Transform (DFT) coefficients [10]. DCT turn over the image edge to make the image transformed into the form of even function with the character of discrete Fourier transform (DFT) and its inverse manner can be expressed as follows [11]:

The DCT transform of

$$F(u,v) = \frac{4C(u)C(v)}{n^2} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} f(j,k) \cos[\frac{(2j+1)u\pi}{2n}] \cos[\frac{(2k+1)v\pi}{2n}],$$

The DCT transform inverse

$$f(j,k) = \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} C(u)C(v)F(u,v) \cos[\frac{(2j+1)u\pi}{2n}] \cos[\frac{(2k+1)v\pi}{2n}],$$

Where C(w) = 1/ 2 and when w =0 C(w) = 1 also when w =1,2,3,… n -1

For digital image concept for N-by-N image matrix of real numbers we will have [12, 13]

$$F = \begin{bmatrix} f_{00} & f_{01} & \cdots & f_{0(N-1)} \\ f_{10} & f_{11} & \cdots & f_{1(N-1)} \\ \vdots & \vdots & \vdots & \vdots \\ f_{(N-1)0} & f_{(N-1)1} & \cdots & f_{(N-1)(N-1)} \end{bmatrix}$$

Where f_{ij}=pixel value and f^2_{ij} is proportional to brightness or energy. Let T be the transpose vector as follows

$$\vec{f_i} = (f_{0i}, f_{1i}, \ldots, f_{(N-1)i})^T$$

The digital image F can be expressed as formula

$$F = \begin{bmatrix} \vec{f_0} & \vec{f_1} & \ldots & \overrightarrow{f_{N-1}} \end{bmatrix}$$

The engagement of this process was used to compress the image after the image encryption technique.

3.3 The watremarking process

Let the host signal be defined by A as below. The following approach was used to embed the data into the image, A. For a given Image A, we have

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdot & \cdots & x_{2n} \\ x_{31} & \cdot & \cdot & \cdot & \cdots & x_{3n} \\ x_{41} & \cdot & \cdot & \cdot & \cdots & x_{4n} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{m1} & x_{m2} & x_{m3} & x_{m4} & \cdots & x_{mn} \end{bmatrix}$$

For a given message E to be embedded in A we have,

$$E = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdots & x_{2n} \\ x_{31} & \cdot & \cdot & \cdots & x_{3n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix}$$

Let s(R) =size of R be [row, column] = size (R) =R (c x p)
Embedding E the data into A will be
d=Eij, where d is the elements of the data to be embedded
Let the size of d be [c1, p1] =size (d)

```
for i=1:1:c1
            for j=1:1:p1
             if(c1<c)
                 Aij= Aij+Eij
               else
               end
            end
end
```

The images were encrypted and the results were analyzed below.

4 Analysis and Results

The image below was obtained from a Unmanned Ariel vehicle (UAV) was encrypted, compressed, watermarked and analyzed using the proposed approach. The recovery of the plain image from the compressed ciphered image was achieved successfully.

Figure 2. The plain image from a UAV drone.

Figure 3. The ciphered image

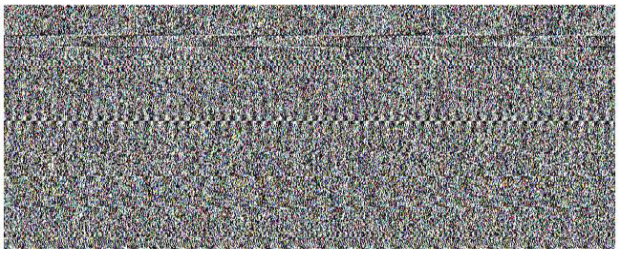


Figure 4. The DCT of the ciphered image.

Figure 5. The image of the loss pixel values

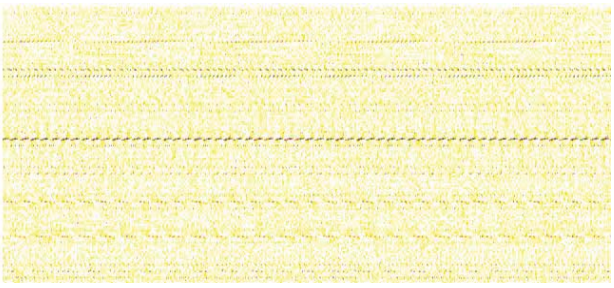


Figure 6. The watermarked and compreesed image

Figure 7. The IDCT of the dewatermarked and decompressed image

Figure 8. The recovered image from fig 3 .

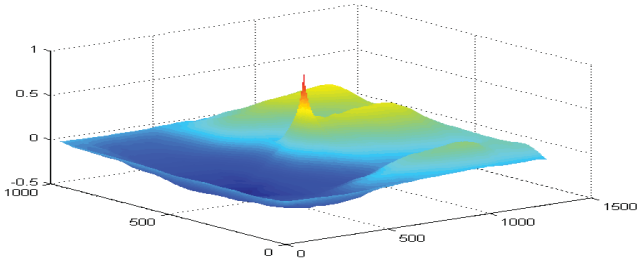


Figure 9. The graph of the normalized cross-correlation of the matrices of the plain image.

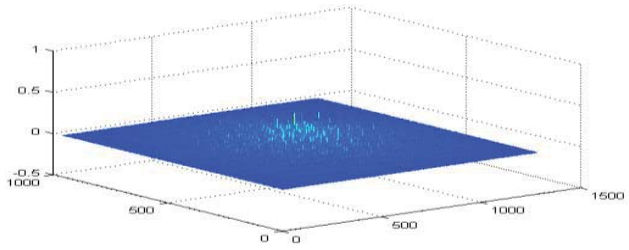


Figure 10. The graph of the normalized cross-correlation of the matrices of the ciphered image.

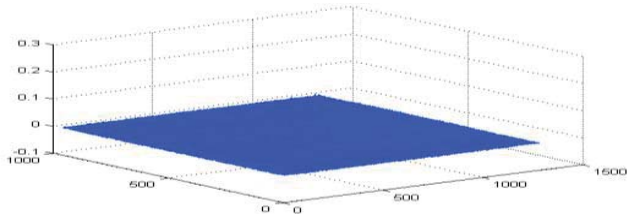


Figure 11. The graph of the normalized cross-correlation of the matrices of the loss pixel values.

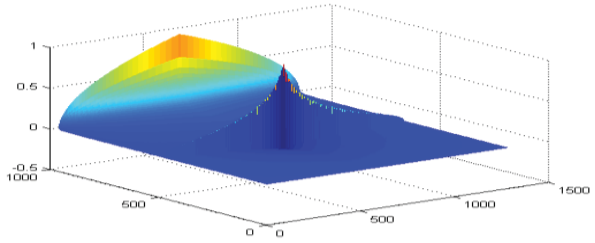


Figure 12. The graph of the normalized cross-correlation of the matrices of the watermarked and compreesed image.

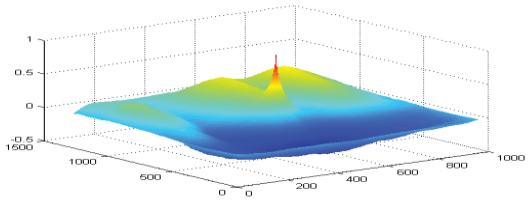


Figure 13. The graph of the normalized cross-correlation of the matrices of the IDCT of the dewatermarked and decompressed image.

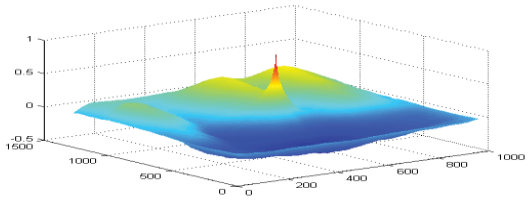


Figure 14. The graph of the normalized cross-correlation of the matrices of the recovered image.

The normalized cross-correlation of the matrices of is

$$\gamma(u,v) = \frac{\sum_{x,y}\left[f(x,y) - \bar{f}_{u,v}\right]\left[t(x-u,y-v) - \bar{t}\right]}{\left\{\sum_{x,y}\left[f(x,y) - \bar{f}_{u,v}\right]^2 \sum_{x,y}\left[t(x-u,y-v) - \bar{t}\right]^2\right\}^{0.5}}$$

f is the mean of the image templates engaged in the process, $\bar{t}$ is the mean of in the region under the image template. $\bar{f}_{u,v}$ is the mean of $f(u,v)$ in the region under the image template. At the end of the entire process, the recovered but compressed, decryped and dewatermarked image still have visual characteristics that makes it not to be visually different from its original one. And the table below is the mean and entropy values of the images in the process.

TABLE 1: ANALYSIS OF PLAIN, CIPHERD, COMPRESSED, DECOMPRESSED AND DECRIPTED IMAGE.

	Entropy(p)	*Arithmetic mean(m)*
PI	7.1726	125.8161
EI	7.1726	125.8161
CI	1.1960	5.7262
WCI	0.4492	94.7262
ICI	7.1753	125.8147
LI	2.0720	1.1645
DI	7.1753	125.8147

PI=plain image, EI= Enrypted Image, CI=Compressed Image, WCI=Watermarked and Compressed Image, ICI=Decompressed and dewatermarked Image, LI= Image Loss due to compression, DI=Decrypted Compressed Image

5 Conclusion

Our proposed approach was resistive against statistical and brute force attacks. The encryption process was effective for all the images and there was no pixel expansion at the end of the process. But there was an effective loss of pixel value during the encryption process but it was insignificant to the visuals of the image. The entropy and mean values for the images in were computed and indicated in the table above. The total entropy and the mean of the plain images never changed for all the ciphered images and that of the dewatermarked and the recovered images also remained the same.

Acknowledgments. This work was supported by Lab-STICC (UMR CNRS 6285) at UBO France, AWBC Canada, Ambassade de France-Institut Français-Ghana and the DCSIT-UCC, and also Dominique Sotteau (formerly directeur de recherche, Centre national de la recherche scientifique (CNRS) in France and head of international relations, Institut national de recherche en informatique et automatique, INRIA) and currently the Scientific counselor of AWBC.

References

[1] Watson, A. B. (1994). Image compression using the discrete cosine transform. Mathematica journal, 4(1), 81.

[2] Dubey, R. B., & Gupta, R. (2011). High quality image compression.

[3] PS, A. K. (2009). Implementation of Image Compression Algorithm using Verilog with Area, Power and Timing Constraints (Doctoral dissertation, National Institute Of Technology Rourkela).

[4] Xu, N. Implementation of data compression and FFT in TinyOS. Embedded Networks Laboratory, Computer Science Dept. USC. Los Angeles, http://enl. usc. edu/ningxu/papers/lzfft. pdf.

[5] Klimesh, M., Stanton, V., & Watola, D. (2001). Hardware implementation of a lossless image compression algorithm using a field programmable gate array. Mars (Pathfinder), 4(4.69), 5-72.

[6] Glynn, E. F. (2007, February). Fourier Analysis and Image Processing. In pdf] Lecture. Bioinformatics Weekly Seminar (Vol. 14).

[7] Riet, p. (2012). Analysis of wavelet transform and fast wavelet transform for image compression: review.

[8] Al-Haj, A. (2007). Combined DWT-DCT digital image watermarking. Journal of computer science, 3(9), 740.

[9] Tomar, Ravi, J. C. Patni, Ankur Dumka, and Abhineet Anand. "Blind Watermarking Technique for Grey Scale Image Using Block Level Discrete Cosine Transform (DCT)." In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, pp. 81-89. Springer International Publishing, 2015.

[10] Ram, B. (2013). Digital Image Watermarking Technique Using Discrete Wavelet Transform And Discrete Cosine Transform. International Journal of Advancements in Research & Technology, 2.

[11] Wen Yuan Chen and Shih Yuan Huang "Digital Watermarking Using DCT Transformation" Department of Electronic Engineering National ChinYi Institute of Technology.

[12] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing," Publishing House of Electronics Industry ; Prentice Hall (Beijin, 2002)

Wen Gao, 1994, "Technique of Multimedia Data Compression," Publishing House of Electron-ics Industry (Beijin: 1994)

An Image Encryption Algorithm with XOR and S-box

Abdelfatah A. Tamimi and Ayman M. Abdalla
Department of Computer Science, Al-Zaytoonah University of Jordan,
P.O. Box 130, Amman 11733, Jordan
E-mail: drtamimi99@gmail.com

Abstract - *This new algorithm performs lossless image encryption by combining variable-length key-dependent XOR encryption with S-box substitution. This algorithm was implemented and tested by performing different permutations of XOR encryption and S-box substitution. Empirical analysis using different types of test images of different sizes showed that this new algorithm is effective and resistant to statistical attacks. The idea presented by this algorithm may be generalized to apply to input data other than images, and may be combined with other encryption methods.*

Keywords: cryptography, block cipher, S-box, XOR.

1 Introduction

The bitwise XOR operation is normally used as a part of a more complex encryption algorithm. Numerous variations of the use of XOR in image encryption can be found in the literature. In the Advanced Encryption Standard (AES), XOR is used as a step in every iteration of the encryption procedure to effectively combine data being encrypted with the encryption key[1]. An algorithm that combines XOR encryption with a rotation operation was designed for effective image encryption[2]. Another algorithm[3] used an affine transform combined with XOR encryption to perform image encryption. Images may also be effectively encrypted using the recursive attributes of the XOR filter[4].

Examples on applying the four steps of AES including, the use of S-box substitution, are available[1]. Many encryption algorithms based on AES were also developed[5,6,7,8,9,10,11]. However, AES has limitations on some multimedia specific requirements[12,13], so other encryption algorithms need to be developed. Some algorithms[14,15] were developed for image encryption using only the S-box substitution from AES as a part of a more complex algorithm that does not use the XOR operation.

In this paper, a new algorithm is presented, which performs lossless encryption via two operations. The first operation performs XOR encryption on the image using variable length blocks. The second operation performs byte substitution using a fixed-size lookup table (S-box). The algorithm was implemented and tested with different combinations of these two operations. Analysis showed effectiveness of the cipher.

2 The New Algorithm

This algorithm takes an image and a key as input. It performs variable-length key-dependent XOR encryption and applies byte substitution using a lookup table called S-box. Different combinations of these two encryptions may be performed, where the decryption performs the inverse of the applied steps in reverse order.

In the XOR encryption operation of the algorithm, the image is regarded as a stream of bytes, and then it is divided into groups (one-dimensional blocks). Let the input image have n bytes and the key have b bytes referred to as $key[0]$ through $key[b-1]$. The image is divided into approximately $n/\sum_{i=0}^{b-1} key[i]$ groups of bytes. Group number j will consist of $key[i]$ bytes where $j = (b \times c + i)$ for some non-negative integer c. For example, for a key of 16 bytes where the value of its $key[5] = 70$, there will be groups in the image consisting of 70 bytes each, namely: the groups numbered 5, 21, 37, 53, 69, etc.

Each of the above groups is encrypted with XOR as follows. Suppose the bytes of group number j are $G[0]$ through $G[key[i] - 1]$. Then, the encrypted values will be:

$$G'[0] = (G[0] \text{ XOR } key[i]), \text{ and}$$

$$G'[p] = (G[p] \text{ XOR } G'[p-1]) \text{ for } 0 < p \le key[i] - 1. \quad (1)$$

Each group is encrypted similarly but independently of the other groups.

The two-dimensional substitution table, known as S-box, is constructed to perform two transformations: multiplicative inverse and affine transformation. This nonlinear key-dependent substitution was presented as a step in each iteration of the AES algorithm[1]. However, in the new algorithm presented here, this substitution is performed at most two times. It is either applied before, after, or both before and after the XOR encryption operation. It is applied to the entire image; block by block. The S-box substitution in this algorithm may also be skipped if needed. If S-box substitution is skipped, another substitution or shuffling operation should be applied in addition to the XOR encryption, so that no encrypted group will remain intact or in

the same location inside the image as produced by the XOR encryption operation.

The decryption algorithm is similar to the encryption algorithm, where each of the above steps can be easily inverted. The inverted steps are performed in reverse order, and the decryption restores the original image without any loss.

3 Implementation and Analysis

The security of the new algorithm comes from combining the two encryption operations; using XOR encryption and S-box byte substitution. If XOR encryption is used alone, the encryption may become vulnerable to brute-force and plaintext attacks. Using S-box substitution alone could make the encryption vulnerable to statistical attacks. A combination of these two encryption operations will provide significant resistance to all of these types of attacks.

If one or more bits in the key are changed, it causes a different grouping in the XOR step and the XOR values are changed. In addition, let an S-box of size 16×16 bytes be used in the S-box substitution step. This S-box has 2,048 different entries where each of these entries consists of 8 bits. This makes the total number of permutations for this step is 2^{11}. Consequently, for an image of ten or more kilobytes input to any combination of these two encryption operations, a brute-force attack is impossible.

The algorithm was applied to 50 images of various types and sizes. When different keys were used with the same image, they produced different encrypted images. In addition, analysis using histograms, correlation, and peak signal to noise ratio (PSNR) showed properties of the algorithm that strongly resist statistical attacks.

The histograms of the images encrypted with any combination of the operations of the new algorithm were uniform and different from the histograms of the original images. They gave no indication that may help statistical attacks.

The mean squared error for two images, stored in matrices *A* and *B*, is computed as follows:

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(A[i,j]-B[i,j])^2 \qquad (2)$$

PSNR is computed as:

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \qquad (3)$$

where *MAX* is the maximum pixel value of the image, and the PSNR measurement unit is the decibel (dB). A lower PSNR value is desired for encrypted images since it indicates more noise and, therefore, more resistance to attacks.

Figure 1 shows PSNR computed for encrypted images resulting from encrypting the original image using different combinations of XOR and S-box encryptions: without S-box, S-box first followed by XOR, S-box last (after XOR), and S-box twice (once before and once after XOR). As it appears in the figure, the PSNR values of these methods were similar. The average PSNR values for the results are shown in the first column of Table 1. The average PSNR value was the highest (i.e., best) when applying S-box exactly once, where applying S-box before XOR produced an average close to that when S-box was applied after XOR. The average was slightly lower when S-box was used twice and the lowest (worst) when it was not used at all.

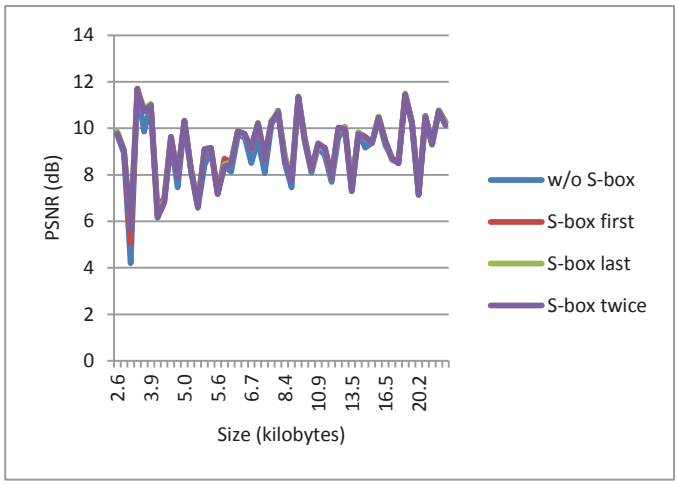


Figure 1. PSNR resulting from different combined operations

Table 1. Average values with different combinations of encryption operations

Operation Combination	PSNR	Correlation
XOR without S-box	8.999649	0.028833
S-box before XOR	9.216589	0.006321
S-box after XOR	9.208713	0.006116
S-box before & after XOR	9.174472	0.008030

The correlation, *r*, between two images, stored in matrices *A* and *B*, is computed as follows, where $\bar{A}$ and $\bar{B}$ are mean values for matrices *A* and *B*, respectively:

$$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(A[i,j]-\bar{A})(B[i,j]-\bar{B})}{\sqrt{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(A[i,j]-\bar{A})^2\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(B[i,j]-\bar{B})^2\right)}} \qquad (4)$$

A lower correlation value between an image and its encryption indicates less resemblance between them, which provides more resistance to attacks.

The correlation value computed for encrypted images resulting from encrypting the original image using different combinations of XOR and S-box encryptions is shown in Figure 2. As seen in the figure, using XOR without S-box generally gave the highest (worst) value of all four encryption combinations, while other encryption combinations gave values similar to each other. This observation is supported by the average correlation value computed for each operation combination, shown in the second column of Table 1, where the average was taken for the absolute values of correlation for the sample images. The average correlation computed when applying S-box once was lower than the average computed when S-box was used twice, and did not make much difference whether S-box was applied before or after XOR.

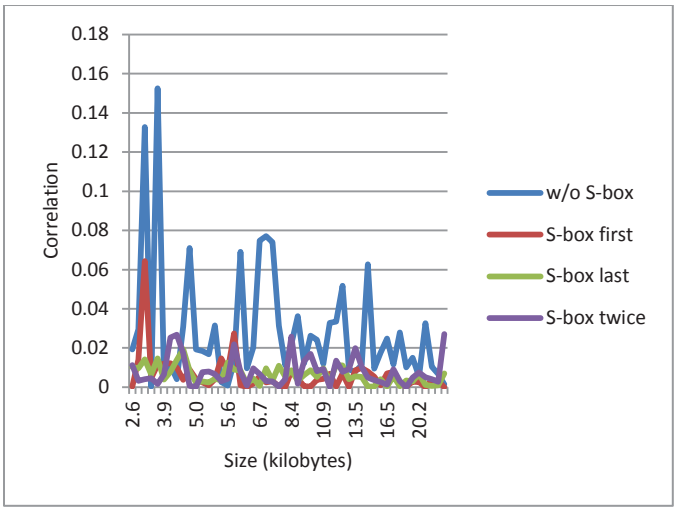


Figure 2. Correlation resulting from different combined operations

Overall, all PSNR values were high and all correlation values were low. This indicates resistance to statistical attacks. The correlation results agreed with the PSNR results in showing the best and worst combinations of XOR and S-box encryptions.

4 Conclusions

A new encryption algorithm was presented. The new algorithm performs encryption using an XOR encryption and S-box substitution. Analysis of different combinations of these two encryptions showed that applying S-box substitution once with XOR encryption produced the best results compared to using it twice or not using it at all.

Statistical analysis using histograms, PSNR, and correlation showed the algorithm is not vulnerable to statistical attacks. In addition, the huge number of possible keys combined with a huge number of possible substitutions makes a brute-force attack on the algorithm impossible.

For future work, the XOR encryption method presented in this paper may be combined with other encryption methods. It is recommended that it should be combined with methods that change the order of bytes through substitution or shuffling. In that case, S-box may not be necessarily used.

5 References

[1] Federal Information Processing Standards (FIPS 197). The Advanced Encryption Standard, 2001. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

[2] M.A.F. Al-Husainy, "A novel encryption method for image security," Int. J. Security & Its Applications, vol. 6(1), pp. 1-8, 2012.

[3] A. Nag, J.P. Singh, S. Khan, S. Biswas, D. Sarkar and P.P. Sarkar, "Image encryption using affine transform and XOR operation," Proc. 2011 Int. Conf. Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), Thuckafay, India. 21-22 July 2011. DOI: 10.1109/ICSCCN.2011.6024565

[4] S.A. Chatzichristofis, L. Bampis, O. Marques, M. Lux and Y. Boutalis, "Image encryption using the recursive attributes of the exclusive-or filter," J. Celluar Automata, vol. 9, pp. 125-137, 2014.

[5] M. Benabdellah, M.M. Himmi, N. Zahid, F. Regragui and E.H. Bouyakhf. "Encryption-compression of images based on FMT and AES algorithm"; Appl. Math. Sci. (Hikari Ltd.), vol. 1 (45), pp. 2203–2219, 2007.

[6] D.A. Duc, T.M. Triet and L.H. Co. "The extended Rijndael-like block ciphers"; Proc. Int. Conf. Info. Tech.: Coding and Computing, pp. 183-188, 2002. DOI: 10.1109/ITCC.2002.1000384

[7] N. El-Fishawy and O.M. Abu Zaid. "Quality of encryption measurement of bitmap images with RC6, MRC6, and Rijndael block cipher algorithms"; Int. J. Net. Sec. (Femto Technique Co.), vol. 5 (3), pp. 241-251, 2007.

[8] A. Yahya and A. Abdalla. "An AES-based encryption algorithm with shuffling"; Proc. 2009 Int. Conf. Security & Management (SAM '09), pp. 113-116, 2009.

[9] M. Zeghid, M. Machhout, L. Khriji, A. Baganne and R. Tourki. "A modified AES based algorithm for image encryption"; Int. J. Comp. Sci. & Eng. (World Academy of Science, Engineering and Technology), vol. 1 (1), pp. 70-75, 2007.

[10] M. Zeghid, M. Machhout, L. Khriji, A. Baganne and R. Tourki. "A modified AES based algorithm for image encryption"; Enformatika (World Enformatika Society), vol. 21, pp. 206-211, 2007.

[11] J.-M. Do and Y.-J. Song, "Secure streaming media data management protocol," Int. J. Security & Its Applications, vol. 8(2), pp. 193-202, 2014. DOI: 10.14257/ijsia.2014.8.2.20

[12] D. Socek, S. Magliveras, D. C'ulibrk, O. Marques, H. Kalva and B. Furht. "Digital video encryption algorithms based on correlation-preserving permutations"; EURASIP J. Inform. Security, 2007.

[13] J.W. Yoon and H. Kim. "An image encryption scheme with a pseudorandom permutation based on chaotic maps"; Commun. Nonlinear Sci. Numer. Simulat., 2010. DOI: 10.1016/j.cnsns.2010.01.041

[14] A. Abdalla and A. Tamimi, "Algorithm for image mixing and encryption," Int. J. Multimedia & Its Applications, vol. 5(2), pp. 15-21, 2013.

[15] A. Tamimi and A. Abdalla, "A Double-Shuffle Image-Encryption Algorithm," Proc. 2012 Int. Conf. Image Processing, Computer Vision, and Pattern Recognition (IPCV '12), pp. 496-499, 2012.

Application of Chaotic Maps for B-Spline functions to Robust Image Cryptography

H.B.Kekre[1], Tanuja Sarode[2], Pallavi N Halarnkar[3]
[2]Computer Engineering Department, TSEC, Mumbai University, Mumbai India
[3]Computer Engineering Department, MPSTME, NMIMS University, Mumbai India

Abstract - *Image encryption is one of most useful way of securing image data. Traditional data security algorithms are also used over images but due to the bulkiness of image data they are not useful. Chaotic sequences play a very good role in image cryptographic system, due to their high sensitivity to initial condition and other properties. In this paper, B-spline functions of first three orders are studied, explored for their chaotic nature and used for image encryption. The experimental results show the proposed approach is useful for encrypting images using the B-Spline functions as chaotic maps. A subjective analysis shows B-spline Map of first order gives good unintelligible encrypted images.*

Keywords: *Chaotic B-spline, Bifurcation, Image Encryption*

1 Introduction

Many cryptographic systems are developed using chaotic functions. The chaotic system are very well known for their properties of sensitivity to initial condition, ergodicity and control parameters which makes them highly suitable for image encryption. Traditional image encryption algorithms are not robust against attacks like noise and shear. Jiu-Lun FAN et al. presented an image encryption method [1] which is robust against noise and shear attacks. The method is based on location transformation. An extension to magic square matrix generation algorithm is also provided. Shiguo Lian et al. made use of skew tent map for image encryption[2]. The proposed method was tested against statistical attack and was found to be robust. The standard map was also modified so as to (0,0) pixel is also shuffled by using a shift operation. Skew Tent map along with m-dimensional cat map is used for image encryption[8] by H.S Kwok. Skew tent map generates the chaotic sequence which is then randomized using m-dimensional cat map. An image encryption algorithm using chaotic system having large key space and high level security was proposed in [3] Haojiang Gao et al. and Amir Akhavan et al.[4], the former uses Non Linear chaotic map and the later makes use of Polynomial for the same. The traditional algorithm only using chaotic map can't meet the demands of encryption. To overcome this, XYYu et al. proposed an image encryption scheme based on Logistic map and SDES which has anti statistic and exhaustion attack[5]. A. N. Pisarchik et al. made use of chaotic map lattices for image encryption [6]. The map parameters, number of iterations and number of cycles are used as a key. A scrambled image in the spatial domain may not resist the statistical attacks. GUO-SHENG GU et al. proposed an image encryption in DWT domain using chaotic function [7]. Experimental results show that the method can resist the attacks offered. WANG Juan [17] proposed a novel image position scrambling method and used it in DWT domain to scramble the image. A fast chaotic image encryption scheme based on dynamic twice interval division is proposed by Liu Xiangdong et al.[9]. The method proves better when tested statistically. An image scrambling technique based on Poker shuffle method controlled dynamically by chaotic system is proposed by Xiaomin Wang [10]. When compared to traditional methods the proposed method has a large key space, non linearity and non analytic formula. Sai Charan Koduru et al. proposed an image encryption scheme to overcome the drawback of traditional approaches in which the confusion and diffusion processes are separated out[11]. In the proposed approach both the processes are carried out in a single step thus improving the efficiency of the encryption algorithm. Meng Jian-liang et al. proposed an image encryption method based on chaotic sequence ranking[12]. The Lorenz system is used to generate the chaotic sequence. Yong Feng et al. used a Line map which consists of two sub maps, left line map and right line map which are used for image encryption and decryption [13]. F. Belkhouche et al. proposed an image encryption method based on Hyperchaos [14]. The method makes use of 2D and 3D maps which permutes the pixel value as well as pixel positions simultaneously. The drawback of limited accuracy using traditional image encryption methods based on chaos systems are overcome by ZHANG Yun-peng et al. [15]. The Logistic map along with DES are used in the process. A multi-chaotic system based on Pixel Chaotic Shuffle and Bit chaotic Rearrangement is proposed by H. H. Nien et al. [16]. The proposed method has good encryption performance. Chen Wei-bin et al. proposed an image encryption method [18] based on scrambling the positions of pixels using Arnold Cat Map and changing the gray values of the image using Henon chaotic map, thus making the image more robust to attackers. The image encryption scheme based on coupled map lattices prove to be efficient and secure when tested statistically for correlation and key sensitivity. The method was proposed by Lin Jinqiu et al. [19].

A new compound two dimensional chaotic function is built using two one dimensional chaotic functions which are switched randomly and used for chaotic sequence generator, the said method is proposed by Xiaojun Tong et al. [20]. Using Logistic map as a chaotic sequence generator and bitXOR operator for encrypting the image values an image encryption scheme is been proposed by Wang Yanling [21], the logistic map with a MOD operator is used for image encryption in a scheme proposed by Guodong Ye [25]. An enhanced version of Logistic Map is proposed by Shatheesh Sam et al. [27] , experimental results prove the method's speed and its suitability in real time image encryption. An extension to one dimensional logistic map to two dimensional is proposed by Rashidah Kadir et al. [22]. A three dimensional Lorenz chaotic map is used to encrypt the image by shuffling the positions of the pixels in the R,G and B plane based on

chaotic sequence ranking. The method is proposed by MU Xiu-chun et al. [23] The 2D Ikeda map is used as a chaotic sequence generator for image encryption. the technique is proposed by Xiaogang Jia [24]. Rui LIU et al. proposed an image scrambling algorithm based on space bit plane operation along with chaotic sequence.[26].

2 B-Spline Functions

The Normalised B-spline blending functions are defined recursively by

$$N_{i,1}(\mathrm{t}) = \begin{cases} 1 & if\ u\ \epsilon[t_i, t_{i+1}) \\ 0 & otherwise \end{cases} \quad (1)$$

and if K> 1,

$$N_{i,k}(t) = \left(\frac{t-t_i}{t_{i+k-1}-t_i}\right) N_{i,k-1}(t) + \left(\frac{t_{i+k}-t}{t_{i+k}-t_{i+1}}\right) N_{i+1,k-1}(t) \quad (2)$$

Using the above equations (1) and (2) B-spline functions of any order can be calculated. In this paper, B-Spline functions of first order, second order and third order are derived and their chaotic nature is studied and used for chaos based image encryption. Following are the equations obtained.

For k=2, the first order B-spline function is given below

$$N_{0,2}(t) = \begin{cases} t & if\ 0 \le t < 1 \\ 2-t & if\ 1 \le t < 2 \\ 0 & otherwise \end{cases} \quad (3)$$

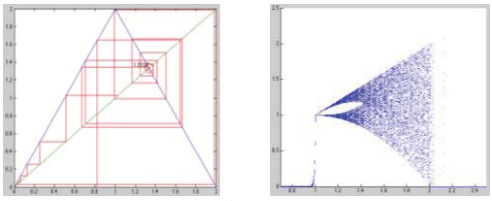


Fig 1.
(a) B-Spline First Order Map (b) Bifurcation Diagram

For k=3, the second order B-spline function is given below

$$N_{0,3}(t) = \begin{cases} \frac{t^2}{2} & 0 \le t < 1 \\ \frac{-3t+6t-2t^2}{2} & 1 \le t < 2 \\ \frac{(3-t)^2}{2} & 2 \le t < 3 \\ 0 & otherwise \end{cases} \quad (4)$$

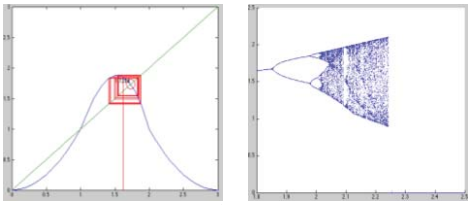


Fig. 2.
(a) B-Spline Second Order Map (b) Bifurcation Diagram

For k=4, the third order B-spline function is given below.

$$N_{0,4}(t) = \begin{cases} \frac{t^3}{6} & 0 \le t < 1 \\ \frac{-3t^3+12t^2-12t+4}{6} & 1 \le t < 2 \\ \frac{3t^3-24t^2+60t-44}{6} & 2 \le t < 3 \\ \frac{(4-t)^3}{6} & 3 \le t < 4 \\ 0 & otherwise \end{cases} \quad (5)$$

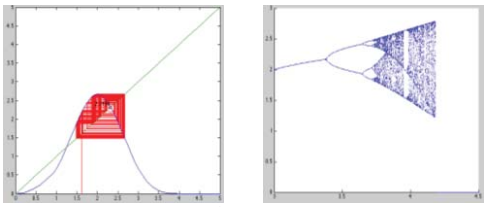


Figure 3.
(a) B-Spline third Order Map (b) Bifurcation Diagram

3 Proposed Approach

In this paper, B-spline functions of first order, second order and third order are been derived (the derivation is not included due to space constraint) and used as chaotic functions and chaotic sequence is generated for some appropriate initial condition. All the different order functions used are given above. Unlike traditional BitXOR operator, this paper makes use of MOD 2-bit, MOD 4-bit and MOD 8-bit operator [28] for the encryption and decryption process.

The encryption process is as follows

1) Read the 24 bit color image, separate the R,G and B planes of the image.
2) For encrypting each plane in the image, Generate three different sequences using three different initial condition with first order/ second order/ third order B-spline function.
3) Use the chaotic sequence generated above, the 3 digits after the decimal point is considered for encryption using the appropriate MOD 2 bit/4 bit/8 bit operator.
4) The above steps results in image encryption.

The steps for image decryption are as follows

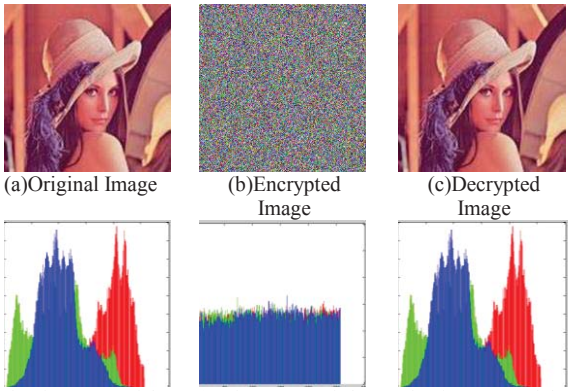
1) Read the encrypted image, separate the R, G and B planes
2) Using the same initial conditions used at the time of encryption (key) are used to generate the chaotic sequence using the appropriate order B-spline function.
3) Use the appropriate MOD operator and apply the inverse procedure to decrypt the encrypted values.
4) The above step will result in a decrypted image.

4 Experimental Results

The proposed approach was applied over 256X256, 24-bit color images. The results shown below are for six different color images. Experimental parameters like Average Row correlation, Average column correlation, PAFCPV [29], Entropy, NPCR[30], Entropy and Histogram were used for analysis purpose.

4.1 First Order B-Spline Map Results

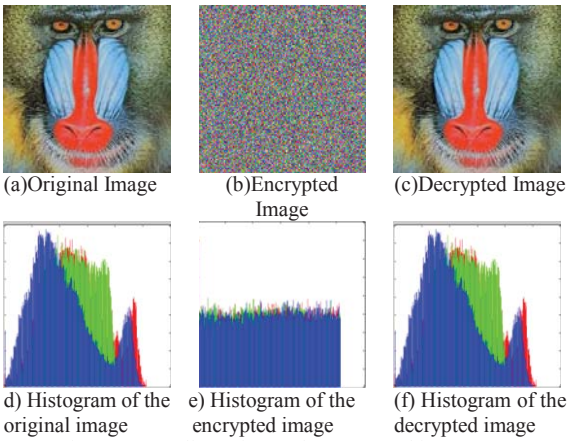
Following are the results obtained for first order B-Spline Map.



(a)Original Image (b)Encrypted Image (c)Decrypted Image

d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

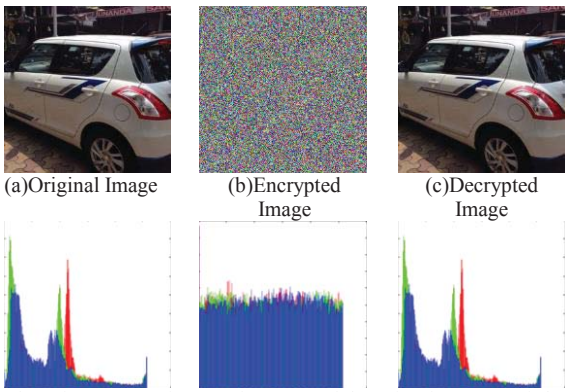
Fig. 4. B-Spline (First Order) MOD 2-bit Operator



(a)Original Image (b)Encrypted Image (c)Decrypted Image

d) Histogram of the original image e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 5. B-Spline (First Order) MOD 4-bit Operator



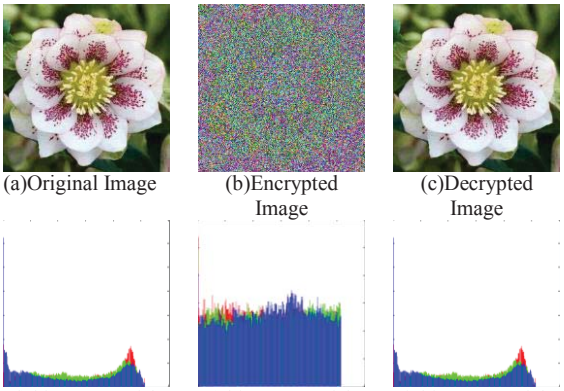
(a)Original Image (b)Encrypted Image (c)Decrypted Image

(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 6. B-Spline (First Order) MOD 8-bit Operator

4.2 Second Order B-Spline Map results

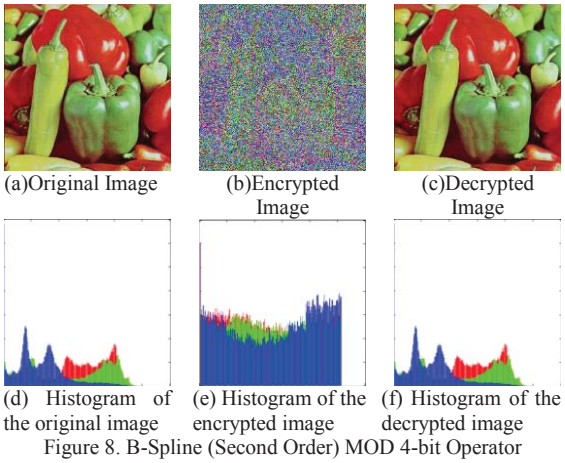
Following are the results obtained for second order B-Spline Map.



(a)Original Image (b)Encrypted Image (c)Decrypted Image

(d) histogram of the original image (e) histogram of the encrypted image (f) histogram of the decrypted image

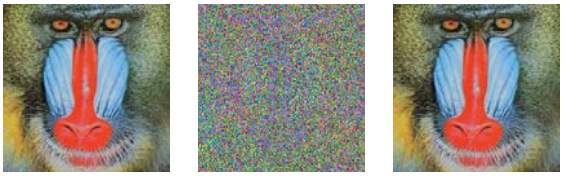
Figure 7. B-Spline (Second Order) MOD 2-bit Operator



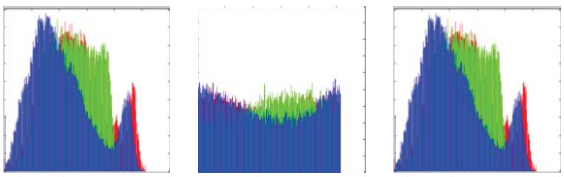
(a)Original Image (b)Encrypted Image (c)Decrypted Image

(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 8. B-Spline (Second Order) MOD 4-bit Operator



(a)Original Image (b)Encrypted Image (c)Decrypted Image

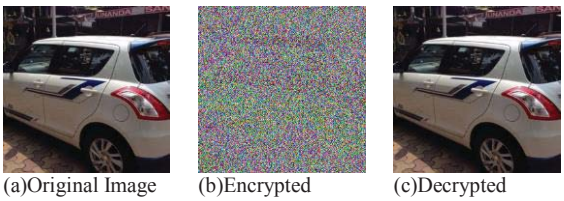


(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

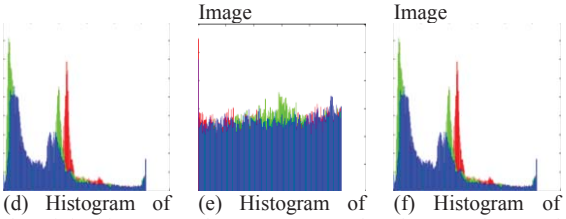
Figure 9. B-Spline (Second Order) MOD 8-bit Operator

4.3 Third Order B-Spline Map results

Following are the results obtained for third order B-Spline Map.

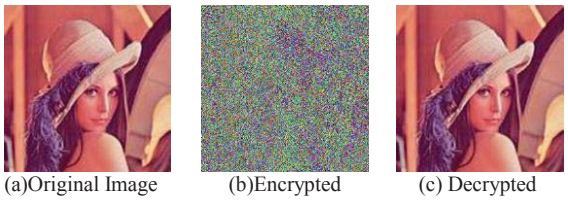


(a)Original Image (b)Encrypted Image (c)Decrypted Image

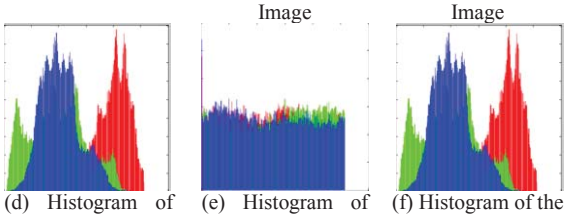


(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 10. B-Spline (third Order) MOD 2-bit Operator

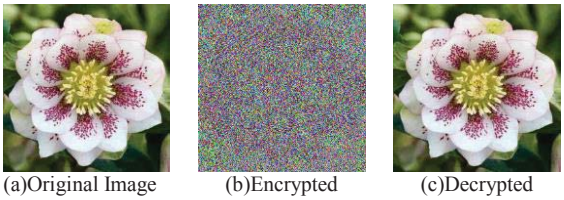


(a)Original Image (b)Encrypted Image (c) Decrypted Image

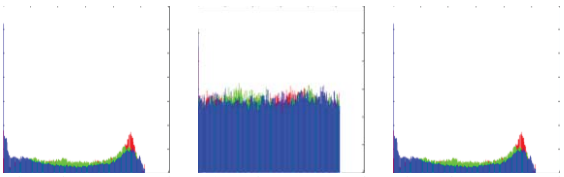


(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 11. B-Spline (third Order) MOD 4-bit Operator



(a)Original Image (b)Encrypted Image (c)Decrypted Image



(d) Histogram of the original image (e) Histogram of the encrypted image (f) Histogram of the decrypted image

Figure 12. B-Spline (third Order) MOD 8-bit Operator

TABLE NO I. VALUES OF INTERSECTION POINT, MIN AND MAX VALUE FOR PERIOD DOUBLING AND RANGE FOR B-SPLINE ORDER 1, B-SPLINE ORDER 2 AND B-SPLINE ORDER 3

Type	Intersection point	Min r	Max r
B-Spline Order 1	1.333	1.1	2.1
B-Spline Order 2	1.728	2	2.25
B-Spline Order 3	2.319	3.5	4.2

TABLE NO II. VALUES USED FOR ENCRYPTION AND DECRYPTION PROCESS IN B-SPLINE ORDER 1, B-SPLINE ORDER 2 AND B-SPLINE ORDER 3 IN R, G AND B PLANES

Type	Plane	Multiplying Factor (R)	Initial Value (Z)
B-Spline Order 1	R-Plane	1.823	0.525
	G-Plane	1.799	0.626
	B-Plane	1.710	0.545
B-Spline Order 2	R-Plane	2.123	1.525
	G-Plane	2.199	1.626
	B-Plane	2.110	1.745
B-Spline Order 3	R-Plane	4.123	2.002
	G-Plane	4.003	2.110
	B-Plane	4.110	2.123

Table No I gives the values of the intersection point of the 45 degree line with the B-spline map, it also shows the chaotic range for first , second and third order B-spline map. It can be observed from the table, B-spline of the first order has the longest range.

Table No II gives the values used for generating the chaotic sequence for B-spline first , second and third order map for encrypting the R, G and B plane of the 24 bit color images used for experimental analysis. Three different initial values (Z) and three different multiplying factors are made use of, to generate the required chaotic sequence.

TABLE NO III. VALUE OF ENTROPY IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 1 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Entropy	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	7.3411	7.9961	7.9965	7.9964
Baboon	7.5814	7.9962	7.9959	7.9962
Pepper	7.3540	7.9967	7.9967	7.9967
Micky	4.7265	7.9955	7.9949	7.9948
Flower	7.8102	7.9970	7.9970	7.9970
Car	7.3024	7.9964	7.9966	7.9966

TABLE NO IV. VALUE OF ROW CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 1 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Row Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.8446	0.1807	0.1817	0.1808
Baboon	0.3609	0.1897	0.1915	0.1902
Pepper	0.6748	0.1819	0.1817	0.1817
Micky	0.5164	0.1807	0.1812	0.1816
Flower	0.5680	0.1792	0.1806	0.1806
Car	0.5572	0.1798	0.1805	0.1805

TABLE NO V. VALUE OF COLUMN CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 1 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Column Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.7001	0.1809	0.1798	0.1816
Baboon	0.4469	0.1902	0.1910	0.1899
Pepper	0.6276	0.1810	0.1814	0.1814
Micky	0.5468	0.1804	0.1811	0.1792
Flower	0.5485	0.1800	0.1816	0.1816
Car	0.7408	0.1815	0.1820	0.1820

TABLE NO VI. VALUE OF PAFCPV IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 1 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.2984	0.3024	0.3040
Baboon	0.2894	0.2937	0.2949
Pepper	0.3143	0.3196	0.3196
Micky	0.4490	0.4582	0.4629
Flower	0.3403	0.3472	0.3472
Car	0.3410	0.3483	0.3483

TABLE NO VII. VALUE OF NPCR IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 1 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	99.5682	99.5687	99.5687
Baboon	99.5753	99.5753	99.5753
Pepper	99.5702	99.5702	99.5702
Micky	99.5702	99.5702	99.5702
Flower	99.5702	99.5702	99.5702
Car	99.5702	99.5702	99.5702

Table No III to VII shows the experimental results obtained for different parameters like Entropy, row correlation, column correlation, PAFCPV and NPCR by the proposed approach using B-spline first order map's chaotic sequence. Parameters entropy, row correlation and column correlation are been compared with the original image values. The highest entropy is obtained for micky image for mod 2 bit, mod 4 bit and mod 8 bit operators. Minimum row correlation is obtained in Lena image and minimum column correlation is obtained in car image for all the three operators. Highest value of PAFCPV is obtained in micky image for mod 8 bit operator. NPCR parameter is giving similar results across all the encrypted images.

TABLE NO VIII. VALUE OF ENTROPY IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 2 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Entropy	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	7.3411	7.9817	7.9747	7.9736
Baboon	7.5814	7.9891	7.9853	7.9846
Pepper	7.3540	7.9842	7.9795	7.9786
Micky	4.7265	7.9397	7.9288	7.9230
Flower	7.8102	7.9936	7.9927	7.9925
Car	7.3024	7.9835	7.9830	7.9811

TABLE NO IX. VALUE OF ROW CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 2 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Row Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.8446	0.1819	0.1805	0.1817
Baboon	0.3609	0.1907	0.1878	0.1901
Pepper	0.6748	0.1804	0.1818	0.1821
Micky	0.5164	0.1828	0.1814	0.1825
Flower	0.5680	0.1808	0.1811	0.1806
Car	0.5572	0.1804	0.1826	0.1803

TABLE NO X. VALUE OF COLUMN CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 2 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Column Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.7001	0.1822	0.1835	0.1844
Baboon	0.4469	0.1916	0.1918	0.1919
Pepper	0.6276	0.1823	0.1849	0.1853
Micky	0.5468	0.1807	0.1795	0.1825
Flower	0.5485	0.1818	0.1820	0.1829
Car	0.7408	0.1811	0.1822	0.1804

TABLE NO XI. VALUE OF PAFCPV IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 2 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.3167	0.3175	0.3155
Baboon	0.3082	0.3118	0.3106
Pepper	0.3349	0.3439	0.3446
Micky	0.4612	0.4611	0.4570
Flower	0.3590	0.3650	0.3649
Car	0.3604	0.3691	0.3721

TABLE NO XII. VALUE OF NPCR IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 2 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	99.7187	99.7187	99.7187
Baboon	99.7123	99.7123	99.7123
Pepper	99.7187	99.7187	99.7187
Micky	99.7187	99.7187	99.7187
Flower	99.7187	99.7187	99.7187
Car	99.7187	99.7187	99.7187

Table No VIII to XII shows the experimental results obtained by the proposed approach using B-spline second order map's chaotic sequence. The highest entropy is obtained for micky image for mod 2 bit, mod 4 bit and mod 8 bit operators. Minimum row correlation is obtained in lena image and minimum column correlation is obtained in car image for all the three operators. Highest value of PAFCPV is obtained in micky image for mod 2 and 4 bit operators. NPCR parameter is giving similar results across all the encrypted images.

TABLE NO XIII. VALUE OF ENTROPY IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 3 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Entropy	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	7.3411	7.9941	7.9931	7.9929
Baboon	7.5814	7.9934	7.9926	7.9925
Pepper	7.3540	7.9947	7.9939	7.9938
Micky	4.7265	7.9607	7.9570	7.9528
Flower	7.8102	7.9961	7.9960	7.9958
Car	7.3024	7.9926	7.9926	7.9918

TABLE NO XIV.VALUE OF ROW CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 3 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Row Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.8446	0.1811	0.1802	0.1806
Baboon	0.3609	0.1907	0.1908	0.1930
Pepper	0.6748	0.1806	0.1807	0.1811
Micky	0.5164	0.1822	0.1826	0.1823
Flower	0.5680	0.1815	0.1826	0.1817
Car	0.5572	0.1816	0.1811	0.1799

TABLE NO XV.VALUE OF COLUMN CORRELATION IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 3 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Original Column Corr	Encrypted Images		
		MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.7001	0.1806	0.1797	0.1807
Baboon	0.4469	0.1904	0.1904	0.1902
Pepper	0.6276	0.1807	0.1808	0.1784
Micky	0.5468	0.1811	0.1794	0.1787
Flower	0.5485	0.1813	0.1810	0.1804
Car	0.7408	0.1819	0.1795	0.1802

Table No XVI. Value of PAFCPV in Original and Encrypted Images for B-spline Order 3 using MOD 2 bit, MOD 4 bit and MOD 8 bit operator

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	0.3120	0.3136	0.3132
Baboon	0.3062	0.3095	0.3078
Pepper	0.3332	0.3358	0.3339
Micky	0.4786	0.4585	0.4513
Flower	0.3629	0.3632	0.3609
Car	0.3648	0.3668	0.3668

TABLE NO XVII. VALUE OF NPCR IN ORIGINAL AND ENCRYPTED IMAGES FOR B-SPLINE ORDER 3 USING MOD 2 BIT, MOD 4 BIT AND MOD 8 BIT OPERATOR

Image Name	Encrypted Images		
	MOD 2-bit	MOD 4-bit	MOD 8-bit
Lena	99.7055	99.7050	99.7050
Baboon	99.7169	99.7169	99.7169
Pepper	99.7050	99.7050	99.7050
Micky	99.7050	99.7050	99.7050
Flower	99.7050	99.7050	99.7050
Car	99.7050	99.7050	99.7050

Table No XIII to XVII shows the experimental results obtained by the proposed approach using B-spline third order map's chaotic sequence. The highest entropy is obtained for micky image for mod 2 bit, mod 4 bit and mod 8 bit operators. Minimum row correlation is obtained in lena image and minimum column correlation is obtained in car image for all the three operators. Highest value of PAFCPV is obtained in micky image for mod 2 bit operator. NPCR parameter is giving similar results across all the encrypted images.

5 Conclusion

In this paper we have given image encryption method for B-spline first, second and third order. We have studied the chaotic nature of B-Spline maps of different order. The chaotic sequences generated using different initial conditions were used in the encryption process. Based on experimental results obtained it is proved that the method has good robustness due to its chaotic nature of high sensitivity to initial condition. Various parameters like entropy obtained is very close to maximum entropy in the encrypted images across all the B-spline functions. Row correlation is minimum in Lena image across all the B-spline functions and column correlation is minimum in Car image across all the B-spline functions. PAFCPV is maximum in Micky image across all the B-spline functions. NPCR parameter values obtained are similar across all the three B-spline functions for all images. From the Flat Histogram of the encrypted images it can be concluded that the proposed technique is ideal for image encryption.

6 References

[1] Fan, Jiu-Lun, and Xue-Feng Zhang, "Image encryption algorithm based on chaotic system", 7th IEEE International Conference on Computer-Aided Industrial Design and Conceptual Design, CAIDCD'06. pp.1-6, 2006.

[2] Lian, Shiguo, Jinsheng Sun, and Zhiquan Wang "A block cipher based on a suitable use of the chaotic standard map", Chaos, Solitons & Fractal. 26(1): pp.117-129, 2005.

[3] Gao, Haojiang, Yisheng Zhang, Shuyun Liang, and Dequn Li, "A new chaotic algorithm for image encryption", Chaos, Solitons & Fractals. ; 29(2):pp.393-399, 2006.

[4] Akhavan, Amir, Hadi Mahmodi, and Afshin Akhshani, "A new image encryption algorithm based on one-dimensional polynomial chaotic maps", In Computer and Information Sciences–ISCIS 2006. Springer Berlin Heidelberg, pp.963-971, 2006.

[5] Yu, X. Y., J. Zhang, H. E. Ren, G. S. Xu, and X. Y. Luo, "Chaotic image scrambling algorithm based on S-DES", In Journal of Physics: Conference Series. IOP Publishing, ;48(1): p. 349, 2006

[6] Pisarchik, A. N., N. J. Flores-Carmona, and M. Carpio-Valadez. "Encryption and decryption of images with chaotic map lattices", Chaos: An Interdisciplinary Journal of Nonlinear Science 16(3)., 2006.

[7] Gu, Guo-Sheng, and Guo-Qiang Han, "The application of chaos and DWT in image scrambling", In International Conference on Machine Learning and Cybernetics.pp.3729-3733, 2006.

[8] Kwok, H. S., and Wallace KS Tang, "A fast image encryption system based on chaotic maps with finite precision representation", Chaos, solitons & fractals. ; 32(4): pp.1518-1529., 2007.

[9] Xiangdong, Liu, Zhang Junxing, Zhang Jinhai, and He Xiqin, "A new chaotic image scrambling algorithm based on dynamic twice interval-division", In 2008 International Conference on Computer Science and Software Engineering.;(3) pp.818-821, 2008.

[10] Wang, Xiaomin, and Jiashu Zhang, "An image scrambling encryption using chaos-controlled Poker shuffle operation", IEEE International Symposium on Biometrics and Security Technologies, ISBAST.pp. 1-6, 2008.

[11] Koduru, Sai Charan, and V. Chandrasekaran, "Integrated confusion-diffusion mechanisms for chaos based image encryption", IEEE 8th International Conference on Computer and Information Technology Workshops, 2008. (CIT) : pp. 260-263, 2008.

[12] Jian-liang, Meng, Pang Hui-jing, and Gao Wan-qing, "New color image encryption algorithm based on chaotic sequences ranking", International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08. Pp.1348-1351,2008.

[13] Feng, Yong, and Xinghuo Yu, "A novel symmetric image encryption approach based on an invertible two-dimensional map", 35th Annual Conference of IEEE on Industrial Electronics. IECON'09. pp. 1973-1978, 2009:

[14] Belkhouche, Fethi, and I. Gokcen, "Digital image encoding using hyperchaos", IEEE International Conference on Systems, Man and Cybernetics, SMC.pp. : 1349-1352, 2009.

[15] Yun-Peng, Zhang, Liu Wei, Cao Shui-ping, Zhai Zheng-jun, Nie Xuan, and Dai Wei-di, "Digital image encryption algorithm based on chaos and improved DES", IEEE International Conference on Systems, Man and Cybernetics.pp.474-479, 2009.

[16] Nien, H. H., Wei-Tzer Huang, C. M. Hung, S. C. Chen, S. Y. Wu, C. K. Huang, and Y. H. Hsu, "Hybrid image encryption using multi-chaos-system", 7th International Conference on Information, Communications and Signal Processing, ICICS. pp.1-5, 2009.

[17] Juan, Wang, "Image encryption algorithm based on 2-D wavelet transform and chaos sequences", IEEE International Conference on Computational Intelligence and Software Engineering CiSE pp. 1-3, 2009.

[18] Wei-bin, Chen, and Zhang Xin, "Image encryption algorithm based on Henon chaotic system", International Conference on Image Analysis and Signal Processing, 2009. IASP. pp.94-97, 2009.

[19] Jinqiu, Lin, and Si Xica, "Image encryption algorithm based on hyperchaotic system", International Workshop on Chaos-Fractals Theories and Applications. IWCFTA'09. pp.153-156,2009.

[20] Tong, Xiaojun, and Minggen Cui, "Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator", Signal processing. ; 89(4): pp.480-491, 2009.

[21] Yanling, Wang, "Image scrambling method based on chaotic sequences and mapping", First International Workshop on Education Technology and Computer Science. ETCS'09. ;3: pp. 453-457.IEEE, 2009.

[22] Kadir, Rashidah, Rosdiana Shahril, and Mohd Aizaini Maarof, "A modified image encryption scheme based on 2D chaotic map", International Conference on Computer and Communication Engineering (ICCCE). pp: 1-5, 2010.

[23] Xiu-chun, Mu, and E. Song, "A New Color Image Encryption Algorithm Based on 3D Lorenz Chaos Sequences", First International Conference on Pervasive Computing Signal Processing and Applications (PCSPA). Pp. : 269-272, 2010.

[24] Jia, Xiaogang, "Image Encryption using the Ikeda map", International Conference on Intelligent Computing and Cognitive Informatics (ICICCI). pp. 455-458, 2010.

[25] Ye, Guodong, "Image scrambling encryption algorithm of pixel bit based on chaos map", Pattern Recognition Letters ; 31(5): 347-354, 2010.

[26] Liu, Rui, and Xiao-ping Tian "A space-bit-plane scrambling algorithm for image based on chaos", Journal of Multimedia. ;6(5): 458-466, 2011.

[27] Sam, I. Shatheesh, P. Devaraj, and Raghuvel S. Bhuvaneswaran, "Chaos based image encryption scheme based on enhanced logistic map", In Distributed Computing and Internet Technology Springer Berlin Heidelberg,pp.290-300, 2011.

[28] H.B.Kekre, Tanuja Sarode, Pallavi N Halarnkar, "Performance evaluation of digital image encryption using Discrete Random distrbutions and MOD operator", IOSR Journal of Computer Engineering (IOSR-JCE) 16(2) Ver. V pp.54-68, March -April 2014.

[29] H.B.Kekre, Tanuja Sarode, Pallavi N Halarnkar, "Symmetric Key image encryption using continuos distributions with MOD operator", International Journal of Engineering Science and Technology (IJEST), 6(6) pp.316-330, June 2014.

[30] Mohammed A. Shreef, Haider K. Hoomod "Image Encryption Using Lagrange-Least Squares Interpolation", International Journal of Advanced Computer Science and Information Technology (IJACSIT) 2(4), 35-55, 2013.

SESSION

MOTION SEGMENTATION, TRACKING ALGORITHMS, AND APPLICATIONS + VIDEO PROCESSING, ANALYSIS AND APPLICATIONS

Chair(s)

TBA

Articulated Structure from Motion through Ellipsoid Fitting

Peter Boyi Zhang, and Yeung Sam Hung
Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
{byzhang, yshung}@eee.hku.hk

Abstract - *We present a new method to reconstruct non-rigid objects from orthographic projections based on the assumption of articulated model. We introduce an ellipsoid property to identify points belonging to a rigid subset. This enables us to formulate the problem of motion segmentation as an ellipsoid fitting problem. The obtained rigid subsets are then linked as kinematic chains to constitute 3D articulated structure. This method is practical of computational complexity O(N) mainly based on linear least squares. We demonstrate the effectiveness of this method through experiments on real tracking data, motion capture data, and challenging human dataset with missing data, in comparison with existing methods.*

Keywords: non-rigid structure from motion, articulated structure, ellipsoid fitting, motion segmentation, orthographic camera

1 Introduction

The Non-Rigid Structure From Motion (NRSFM) problem aiming at recovering dynamic 3D structure from a sequence of 2D image measurements has attracted much attention over the years. As demonstrated in Figure 1, under affine projection, in each frame of measurement the depth information for all feature points needs to be estimated. This is an ill-posed problem unless additional constraints (e.g. based on the property of the object) are introduced to regulate the estimation to achieve meaningful results.

Bregler *et al.* [1] introduce low rank condition on the object, constraining the object's movement to be a mean shape plus some degrees of freedom. Akhter *et al.* [2] also propose trajectory basis method based on this low rank assumption. Generally, low rank methods are only effective when the object is relatively rigid, and the movement is confined to be linear with few degrees of freedom. Rabaud and Belongie [3] tackle the problem assuming the shape of the object repeats itself at different times, and is thus captured in a few different frames. They identify these frames and recover the object in the same way as a rigid object.

In a real scenario, many objects can fit to an articulated model. They can be modeled as a few rigid subsets connected to each other. With this assumption, as shown in Figure 1, feature points can be grouped into rigid subsets. Points in a same subset are fixed relative to each other. Furthermore, the relative depth between subsets can be recovered based on the knowledge of linkage between subsets.

A major challenge of this problem is the grouping of feature points into rigid subsets, which amounts to motion segmentation. Tresadern and Reid [4] perform motion segmentation based on rank condition. Yan and Pollefeys [5] project points to low-rank manifolds and group points sharing the same subspace. Ross *et al.* [6] employ a probabilistic model to divide and merge the subsets iteratively. Russell *et al.* [7] propose an energy-based approach to label points as belonging to different subsets.

The rank condition is not strong enough to accurately classify points into different subsets, and the number of subsets has to be provided a priori as an input. In this paper we show that beyond the low rank property, a group of points on a rigid structure also has the property of constituting an ellipsoid. The motion segmentation problem can thus be transformed to an ellipsoid model fitting problem. This helps achieve a more precise segmentation of feature points with low computational cost, and leads to better performance in 3D reconstruction.

We describe the ellipsoid fitting method in section 2, and how to perform full 3D reconstruction in section 3. We demonstrate the effectiveness of the method through experiments on challenging datasets in section 4. Some conclusions are drawn in section 5.

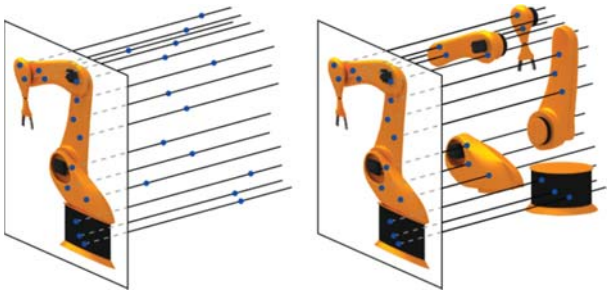


Figure 1. Without constraint, the feature points can have arbitrary depth. With the articulated assumption, the relative depths between feature points of the same rigid subset are fixed, and the relative depth between rigid subsets can be resolved by means of joint linkages. [8]

2 Motion segmentation

Motion segmentation is a critical step for recovering articulated objects. In this section, we describe the problem, reveal the ellipsoid property of rigid subsets, and propose an efficient method to perform motion segmentation through ellipsoid fitting.

2.1 Articulated object

For an articulated object consisting of N feature points measured over F frames by an orthographic camera, let K denote the number of rigid subsets (K is not known); let $\mathrm{W} \in \mathbb{R}^{2F\times N}$ denote the measurements, and let W_f denote the measurement of the fth frame; let $\mathrm{S}^i \in \mathbb{R}^{3\times N^i}$ denote 3D structure of the ith subset, where N^i is the number of points in the ith subset; let $\mathrm{R}_f^i \in \mathbb{R}^{3\times 3}$ and $\mathrm{T}_f^i \in \mathbb{R}^{3\times 1}$ denote the rotation and translation of the ith subset in the fth frame relative to a world coordinate. Let

$$\mathrm{I}_{2\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \tag{1}$$

we have

$$\mathrm{W}_f = \mathrm{I}_{2\times 3}\begin{bmatrix} \mathrm{R}_f^1 \mathrm{T}_f^1 & \cdots & \mathrm{R}_f^K \mathrm{T}_f^K \end{bmatrix} \begin{bmatrix} \mathrm{S}^1 & & \\ & 1 & \\ & & \ddots \\ & & & \mathrm{S}^K \\ & & & & 1 \end{bmatrix}. \tag{2}$$

We denote

$$\mathrm{R}^i = \begin{bmatrix} \mathrm{I}_{2\times 3}\mathrm{R}_1^i \\ \vdots \\ \mathrm{I}_{2\times 3}\mathrm{R}_F^i \end{bmatrix}, \mathrm{T}^i = \begin{bmatrix} \mathrm{I}_{2\times 3}\mathrm{T}_1^i \\ \vdots \\ \mathrm{I}_{2\times 3}\mathrm{T}_F^i \end{bmatrix}. \tag{3}$$

Through motion segmentation, we aim to group the feature points into subsets, so that each subset is rigid throughout all frames. To achieve that, firstly a property needs to be utilized to effectively classify the feature points into rigid subset; secondly an efficient method needs to be proposed to apply the property to different groups of points. We describe the ellipsoid property in section 2.2, and describe the grouping method in section 2.3.

2.2 Ellipsoid fitting

The problem of grouping points into rigid subsets can be transformed to a problem of ellipsoid model fitting. In this section we reveal the ellipsoid property as a fundamental property of rigid structure under orthographic projection, and explain how ellipsoid fitting can help identify rigid subsets.

An ellipsoid in $\mathbb{R}^N$ centered at the origin is defined as a set of points $\mathrm{P} \in \mathbb{R}^N$ satisfying

$$\mathrm{P} = \begin{bmatrix} \mathrm{r}_1\sigma_1 & \mathrm{r}_2\sigma_2 & \mathrm{r}_3\sigma_3 \end{bmatrix} \begin{bmatrix} \mathrm{V}_1 \\ \mathrm{V}_2 \\ \mathrm{V}_3 \end{bmatrix}, \tag{4}$$

subject to

$$\mathrm{r}_1{}^2 + \mathrm{r}_2{}^2 + \mathrm{r}_3{}^2 = 1, \tag{5}$$

where $\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3 \in \mathbb{R}^N$ is a set of orthonormal vectors specifying the axis-direction of the ellipsoid; $\sigma_1, \sigma_2, \sigma_3 > 0$ are constants representing the lengths of the semi-axes; and $\begin{bmatrix} \mathrm{r}_1 & \mathrm{r}_2 & \mathrm{r}_3 \end{bmatrix}$ is a unit vector.

Let W^i denote the measurements of the ith group of points. Subtract the mean of each row of W^i from entries of the same row, we obtain a registered measurement matrix $\widetilde{\mathrm{W}}^i$. Suppose this group is a rigid subset, we have

$$\widetilde{\mathrm{W}}_f^i = \mathrm{W}_f^i - \mathrm{I}_{2\times 3}\mathrm{T}_f^i = \mathrm{I}_{2\times 3}\mathrm{R}_f^i\mathrm{S}^i. \tag{6}$$

We show rows of $\widetilde{\mathrm{W}}^i$ lie on an ellipsoid through the following analysis, and after that we will describe how to perform ellipsoid fitting. Express S^i by SVD as

$$\mathrm{S}^i = \mathrm{U}\Sigma\mathrm{V}^T, \tag{7}$$

then

$$\widetilde{\mathrm{W}}_f^i = \mathrm{I}_{2\times 3}\mathrm{R}_f^i\mathrm{U}\Sigma\mathrm{V}^T. \tag{8}$$

We absorb the orthogonal matrix U into R_f^i. Let $\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3$ be the first three rows of V^T; $\sigma_1, \sigma_2, \sigma_3$ be the diagonal elements of Σ. Denote the entries of $\mathrm{I}_{2\times 3}\mathrm{R}_f^i$ as

$$\mathrm{I}_{2\times 3}\mathrm{R}_f^i = \begin{bmatrix} \mathrm{r}_{f1} & \mathrm{r}_{f2} & \mathrm{r}_{f3} \\ \mathrm{r}_{f4} & \mathrm{r}_{f5} & \mathrm{r}_{f6} \end{bmatrix}, \tag{9}$$

then

$$\widetilde{\mathrm{W}}_f^i = \begin{bmatrix} \mathrm{r}_{f1}\sigma_1 & \mathrm{r}_{f2}\sigma_2 & \mathrm{r}_{f3}\sigma_3 \\ \mathrm{r}_{f4}\sigma_1 & \mathrm{r}_{f5}\sigma_2 & \mathrm{r}_{f6}\sigma_3 \end{bmatrix} \begin{bmatrix} \mathrm{V}_1 \\ \mathrm{V}_2 \\ \mathrm{V}_3 \end{bmatrix}. \tag{10}$$

According to (4) and (5), rows of $\widetilde{\mathrm{W}}^i$ lie on a same ellipsoid in $\mathbb{R}^{N^i}$. Moreover, the two rows of $\widetilde{\mathrm{W}}_f^i$ satisfy

$$\begin{bmatrix} \mathrm{r}_{f1} & \mathrm{r}_{f2} & \mathrm{r}_{f3} \\ \mathrm{r}_{f4} & \mathrm{r}_{f5} & \mathrm{r}_{f6} \end{bmatrix} \begin{bmatrix} \mathrm{r}_{f1} & \mathrm{r}_{f2} & \mathrm{r}_{f3} \\ \mathrm{r}_{f4} & \mathrm{r}_{f5} & \mathrm{r}_{f6} \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{11}$$

To fit ellipsoid to W^i, we first project its row vectors to a 3D subspace. We register rows of W^i to their centroid to get $\widetilde{\mathrm{W}}^i$. As illustrated in Figure 2, each row of $\widetilde{\mathrm{W}}^i$ can be represented as a point in N^i-dimensional space. We then perform a rank-3 approximation on $\widetilde{\mathrm{W}}^i$, so that points in N^i-dimensional space are projected to points in a 3D subspace, given by

$$\widetilde{\mathrm{W}}_f^i = \begin{bmatrix} \mathrm{x}_{f1} & \mathrm{y}_{f1} & \mathrm{z}_{f1} \\ \mathrm{x}_{f2} & \mathrm{y}_{f2} & \mathrm{z}_{f2} \end{bmatrix} \begin{bmatrix} \hat{\mathrm{V}}_1 \\ \hat{\mathrm{V}}_2 \\ \hat{\mathrm{V}}_3 \end{bmatrix}, \tag{12}$$

where $\hat{\mathrm{V}}_1, \hat{\mathrm{V}}_2, \hat{\mathrm{V}}_3 \in \mathbb{R}^{N^i}$ is an orthonormal basis of the 3D subspace. (12) can be expressed in the form of (10) subject to (11) if and only if there exists $\mathrm{A} \in \mathbb{R}^{3\times 3} > 0$, such that

$$\begin{bmatrix} \mathrm{x}_{f1} & \mathrm{y}_{f1} & \mathrm{z}_{f1} \\ \mathrm{x}_{f2} & \mathrm{y}_{f2} & \mathrm{z}_{f2} \end{bmatrix} \mathrm{A} \begin{bmatrix} \mathrm{x}_{f1} & \mathrm{y}_{f1} & \mathrm{z}_{f1} \\ \mathrm{x}_{f2} & \mathrm{y}_{f2} & \mathrm{z}_{f2} \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{13}$$

This is equivalent to fitting an ellipsoid characterized by A to the projected points. Since $\mathrm{A} > 0$, we may write

$$\mathrm{A} = \hat{\mathrm{U}}\Sigma^{-2}\hat{\mathrm{U}}^T, \tag{14}$$

where $\widehat{\mathrm{U}} \in \mathbb{R}^{3\times3}$ is orthogonal. Then

$$\mathrm{I}_{2\times3}\mathrm{R}_f^i = \begin{bmatrix} \mathrm{x}_{f1} & \mathrm{y}_{f1} & \mathrm{z}_{f1} \\ \mathrm{x}_{f2} & \mathrm{y}_{f2} & \mathrm{z}_{f2} \end{bmatrix} \mathrm{A}^{1/2}, \tag{15}$$

$$\begin{bmatrix} \mathrm{V}_1 \\ \mathrm{V}_2 \\ \mathrm{V}_3 \end{bmatrix} = \widehat{\mathrm{U}}^T \begin{bmatrix} \widehat{\mathrm{V}}_1 \\ \widehat{\mathrm{V}}_2 \\ \widehat{\mathrm{V}}_3 \end{bmatrix}. \tag{16}$$

In our method, we fit the ellipsoid A by least squares as in (13), and then estimate R_f^i using (15). Although this method of first fitting a subspace then an ellipsoid gives a sub-optimal solution compared to directly fitting an ellipsoid in $\mathbb{R}^{N^i}$, it greatly reduces the computational cost. The 3D shape of this subset can be recovered as

$$\mathrm{S}^i = \mathrm{R}^{i\dagger}\widetilde{\mathrm{W}}^i. \tag{17}$$

We define 2D reprojection error

$$\varepsilon = \left\|\widetilde{\mathrm{W}}^i - \mathrm{R}^i\mathrm{S}^i\right\| / (2F \times N^i)^{1/2} \tag{18}$$

as a measurement of how much this group of points deviates from rigidity.

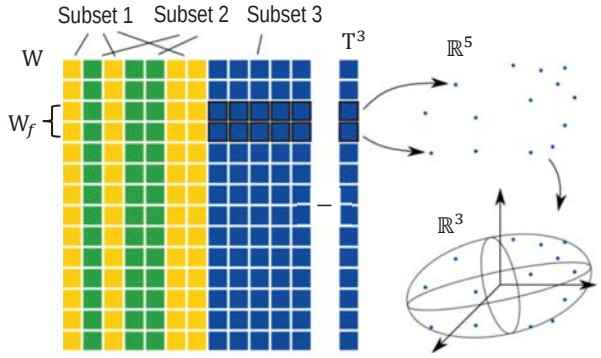


Figure 2. Ellipsoid fitting. Each row in subset 3 is represented as a point in 5D space. The 5D space is projected to its most significant 3D subspace, and an ellipsoid is fitted to the points.

2.3 Grouping of points as rigid subsets

Through the previous analysis, the problem of finding rigid subsets among feature points can be transformed to a problem of finding appropriate ellipsoids parameterized by R^i and T^i, so that each group of columns in W can be fitted to an ellipsoid as

$$\mathrm{W}^i = \begin{bmatrix} \mathrm{R}^i & \mathrm{T}^i \end{bmatrix} \begin{bmatrix} \mathrm{S}^i \\ 1 \end{bmatrix}. \tag{19}$$

To fit an ellipsoid to any possible group of points, there would be as many as 2^N groups to test. We propose a method to perform ellipsoid model fitting with only $O(N)$ complexity. We first estimate the model by fitting ellipsoid to a minimal seed, and then add inliers to the estimated model. A minimal seed contains 4 feature points, because registering each row by its mean reduces the rank of the measurement matrix by 1, and we need at least 3 points to fit an ellipsoid centered at the origin in a 3D subspace. In a real scenario points of a rigid subset are usually close to each other, so in our case, instead of randomly selecting groups of 4 points as seeds, we only consider points close to each other. This would only require $O(N)$ ellipsoid fitting computations. The groups with ellipsoid fitting error ε below a threshold θ are chosen as seeds for rigid subsets.

For a seed and the fitted model R^i, T^i, we evaluate ε of each column in W through reprojection as in (18), and include the points with $\varepsilon < \theta$ to expand the rigid subset. As the rigid subset expands, R^i and T^i may be updated to minimize ε. In total, this process requires $O(N)$ ellipsoid fitting and $O(N^2)$ projections.

After we expand each rigid seed, we might obtain a number of rigid subsets overlapping with each other. This is because in a real scenario the object might not be a perfectly articulated object, and there may be ambiguity as to which subset the points at the boundary should belong to. We resolve this issue by first picking subsets with minimal overlap, and then assign the points belonging to more than one rigid subsets to the subset of minimal reprojection error ε.

If some points are not grouped to any rigid subset, they tend to introduce large error. Either we can choose to discard them; or we can include each point to the rigid subset for which it has minimal ε at the final stage after the recovery of R_f^i and S^i, to reduce their influence. The ability to distinguish outliers is an advantage of this algorithm.

As we shall show in our experiments, the range of choice of θ to provide optimal result is wide. As θ increases from 0, the number of ungrouped points decreases, and the number of duplicate points (points occurring in more than one subsets) increases. As θ reaches the value that well distinguishes the rigid subsets, the number of ungrouped and duplicate points are kept small and stable. The threshold θ should be small if the object can be well modeled as an articulated object. θ should be large if the object is not strictly articulated and/or there is much noise. We shall also show that the 2D reprojection error can be taken as an indicator of the 3D error, for there is a correlation between them.

3 Reconstruction of 3D structure

After segmenting the feature points into rigid subsets, we describe how to reconstruct the 3D structure in this section. We connect the subsets through kinematic chain, eliminate mirror ambiguity based on the linkage between subsets, resolve the relative depth between subsets, and handle missing data.

3.1 Kinematic chain

Similar to Kirk et al. [9] and Yan and Pollefeys [5], we construct a graph recording the cost of connecting every two rigid subsets through joints (called joint cost), and build kinematic chains through performing minimum spanning tree

search in the graph; however, we calculate the joint cost through a different method that only requires least squares computation.

If two rigid subsets S^i and S^j can be connected through a link, we should be able to find virtual points p^i and p^j attached to the coordinate frame of S^i and S^j, respectively, such that the distance between trajectory of p^i and p^j is small (ideally zero). Place the coordinate frame of S^i at the centroid of S^i, $I_{2\times3}T_f^i$ can be recovered as the row mean of W_f^i. Let c^i denote the coefficients combining trajectories of S^i to p^i, the trajectory of p^i as observed in the measurement is

$$p_f^i = I_{2\times3}R_f^i S^i c^i + I_{2\times3}T_f^i. \quad (20)$$

Substitute (6) into (20),

$$p_f^i = \left(W_f^i - I_{2\times3}T_f^i\right)c^i + I_{2\times3}T_f^i. \quad (21)$$

By (21), p_f^i can be approximated directly from W_f^i, without recovery of R_f^i and S^i. The joint cost can be found through least squares minimization as:

$$d^{ij} = \min_{c^i,c^j} \sum_f \left\| p_f^i - p_f^j \right\|^2. \quad (22)$$

We construct an undirected graph with the rigid subsets as nodes, and joint costs d^{ij} as weights of edges. In this graph we perform minimum spanning tree search to recover the connectivity among rigid subsets and then kinematic chains from connected rigid subsets.

3.2 Mirror ambiguity

R_f^i and S^i can be recovered through (15) and (17). Yet a mirror ambiguity remains unresolved, since $I_{2\times3}\bar{R}_f^i = I_{2\times3}R_f^i diag(1,1,-1)$ and $\bar{S}^i = diag(1,1,-1)S^i$ is an equally acceptable solution. We resolve this ambiguity using the physical condition that two subsets linked by a joint generally have similar rotational transformation. For two subsets labeled i and j linked by joints, we align R^i and $\bar{R}^i$ with R^j, and pick the camera motion with smaller error. We arbitrarily pick one subset as reference and propagate alignment to its neighbors and so on. As a result, we will only have one overall mirror ambiguity left to be handled manually.

3.3 3D structure

As we have recovered R_f^i, S^i, the exact joint location p^i in the coordinate frame of S^i can be reevaluated by applying (20) and (22) to subsets linked by joint. Now the only parameter remains to be estimated is the depth z_f^i, the third element of T_f^i. For a set of linked subsets i and j, we set the relative depth between p^i and p^j to be 0 in each frame to minimize the joint distance, assigning one subset as the reference, the depth of its neighboring subsets can be solved through vector addition as shown in Figure 3.

In our model, the joint points p^i and p^j are close to each other, but may not strictly coincide with each other. This relaxed condition allows us to model objects in the physical world that are not perfectly articulated, e.g. the human body.

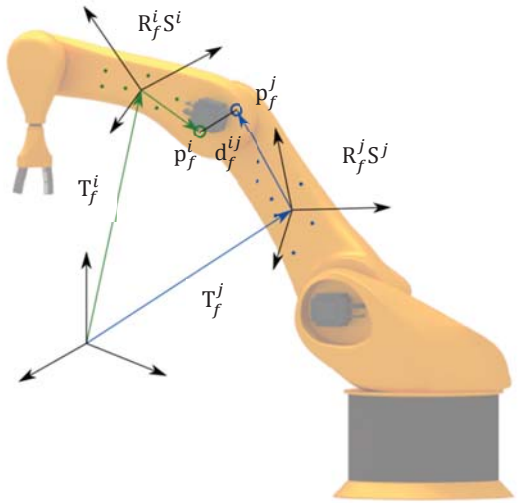


Figure 3. Two subsets linked by joints. The circles represent recovered joint locations. (Best viewed in color.)

3.4 Missing data

Missing data does not present any problem and can be handled in a natural way in our method. For points that are visible in incomplete set of frames, we only use information in the observable frames to perform motion segmentation for these points. Once the points are grouped into rigid subsets, we can estimate the rigid substructure and its transformation in all frames.

4 Experimental Results

We evaluate our method on both motion capture data and real tracking data. The motion capture datasets include 3D ground truth, providing a foundation for quantitative measurement of the accuracy of the algorithm. We evaluate the 2D reprojection error against the measurement matrix, and the 3D error against the 3D ground truth. The 2D reprojection error is defined as the Frobenius norm of the difference between 2D measurement and 2D reprojection, divided by $\|W\|_F$. The 3D error is defined as the Frobenius norm of the difference between 3D ground truth and reconstructed 3D points, divided by the Frobenius norm of 3D ground truth.

We evaluate our method qualitatively on real scenes 'two cranes' and 'toy truck' from [5]. We also perform 3D reconstruction on 'toy truck'. The robustness of our method in the presence of missing data is demonstrated through motion capture data 'skin'.

Depending on its size, each of these datasets takes two seconds to a few minutes for an intel-i5 3.10GHz PC to process.

4.1 Motion capture data

The 'stepstool' dataset taken from subject 40 trial 6 of Carnegie Mellon University Motion Capture Database [10] records a person that climbs, steps over, sits on, and jumps over a stepstool. The camera is fixed while the person performs substantial movements, making it a very challenging NRSFM problem. Before testing our algorithm, we down sampled the number of frames by a factor of 6, and removed the duplicate point tracks, resulting in 1097 frames and 149 points. From the 3D trajectory of all the tracks, we retain the information in y and z axis as 2D measurements. Dataset 'dance' from [2] is a standard NRSFM dataset evaluated by many methods, it has 264 frames and 75 points. The camera rotates around the person horizontally for about 720 degrees throughout the frames. θ is set to 0.3% and 2.2% times $\|W\|_F/(2F \times N)^{1/2}$ for 'stepstool' and 'dance' respectively.

Table 1. 3D error.

	Trajectory basis	Kernel	Ellipsoid Fitting
Stepstool	22.70%	20.73%	1.60%
Dance	18.99%	16.90%	8.85%

We list the results in comparison with the trajectory basis method [2] and the kernel method by Gotardo and Martines [11] in Table 1. We calculate the 3D error after shifting each frame of the 3D ground truth to the centroid. Our method reduces the 3D error by more than an order of magnitude in 'stepstool', and by a factor of two in 'dance'. This shows the limitation of methods based on low-rank assumptions for reconstructing highly flexible objects.

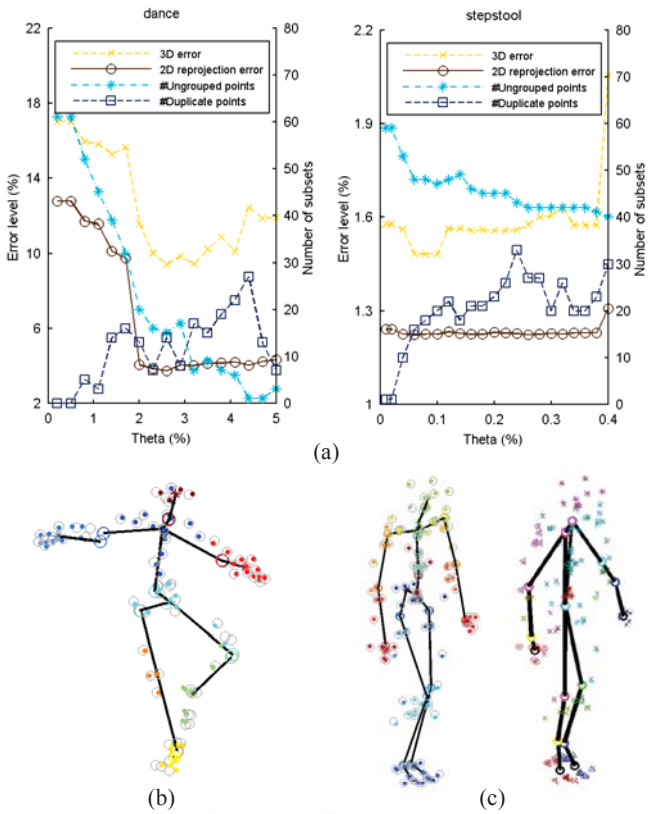


Figure 4. (a) Effect of varying θ. Recovered kinematic structure of (b) 'dance', and (c) 'stepstool' in comparison with result by Ross *et al.* [6] (Best viewed in color. See Figure 7 for meaning of symbols and coloring.)

Figure 4 (a) shows the effect of varying θ. The θ leading to optimal result is chosen where the number of ungrouped and duplicate points are small, the error is robust to change of θ, and the 3D error correlates with 2D error. Figure 4 (b) and (c) show the recovered kinematic chains. We observe that as few as 4 points is enough to determine a rigid subset. For comparison, we include result reported by Ross *et al.* [6] on 'stepstool' to the right. Our method results in 20 subsets, distinguishing the chest, the waist and the hip, and also the shoulder from the upper arm. Although [6] use 3D data instead of 2D measurements as input, they can only identify 15 subsets. Moreover, we recover the joints at more reasonable positons. Figure 7 shows 3D 'stepstool' results in different frames.

4.2 Real tracking data

Data 'two cranes' has 30 frames and 94 points. One crane rotates relative to the camera, while the other crane has two moving parts, but little camera motion. As shown in Figure 5, we segment the data into 3 parts, while Fayad *et al.* [12] further segment one more part on each crane, at least one of which appears to be redundant. Compared with Yan and Pollefeys' result [5] where two points are misclassified as belonging to the other crane, our result is also better. Because there is almost no camera motion to provide depth information, it is not possible to perform 3D reconstruction. Yet this experiment shows that our method is capable of separating the rigid subsets as long as there is relative motion

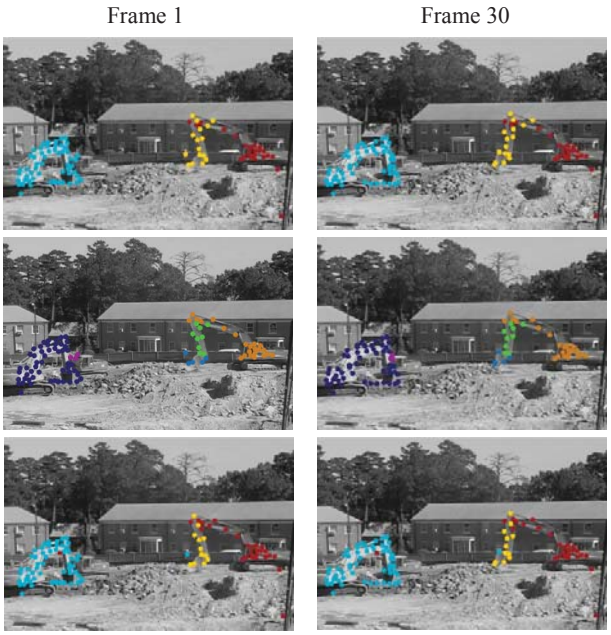


Figure 5. Experiments on 'two crane'. From top to bottom are results by our method, Fayad *et al.* [12], and Yan and Pollefeys [5].

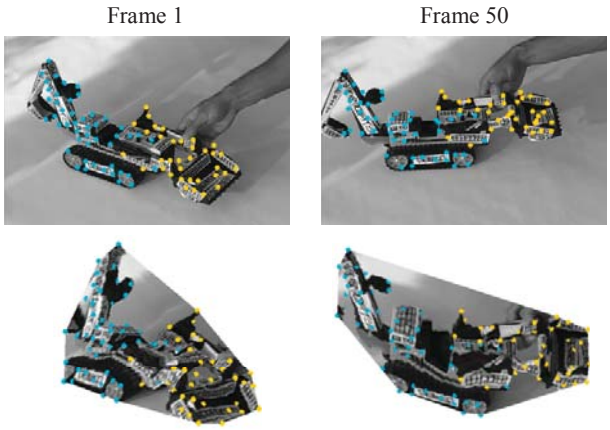


Figure 6. Experiments on 'toy truck'. The first row shows the image and segmentation result, the second shows the 3D reconstruction result.

between them.

Data 'toy truck' has 60 frames and 83 points. It rotates slightly with respect to the camera, and its shovel moves up and down. Figure 6 shows that our method accurately differentiates the shovel from the rest of the truck. We also perform 3D reconstruction. As there is no ground truth for this type of real data, the reconstruction can only be assessed qualitatively. Figure 6 shows textured 3D views of frame 1 with shovel down and frame 50 with shovel up from a different angle, which look reasonable.

4.3 Data with missing observations

The 'skin' dataset from [13] depicts a person flexing his muscles. Approximately 350 markers are placed on the subject to capture the subtle movement of human skin. But since the motion capture system sometimes lose track of the markers because of occlusion, as a result 467 tracks are

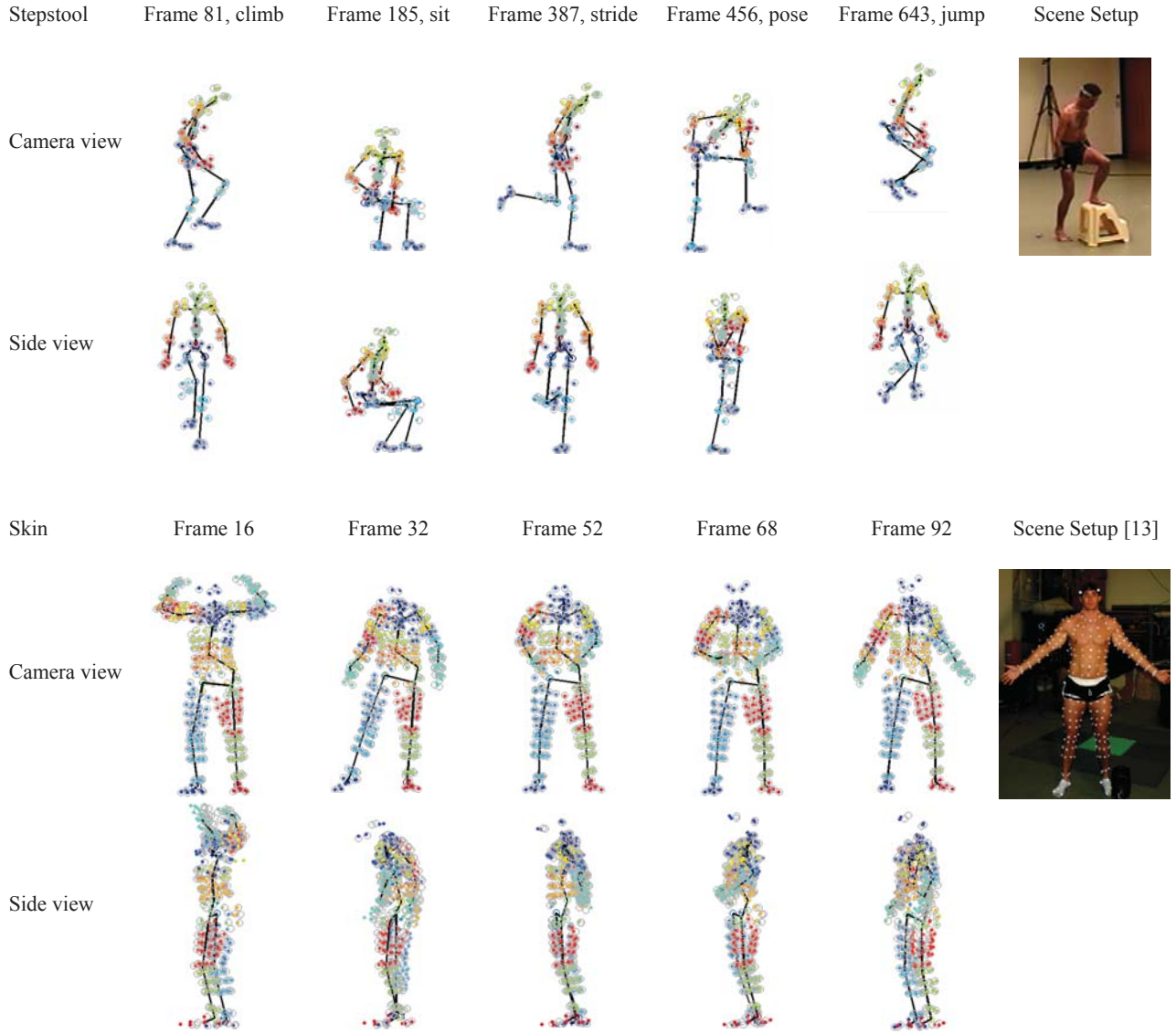


Figure 7. Results on datasets 'stepstool' and 'skin'. The reconstructed points are colored according to different subsets, the gray circles represent ground truth, the colored circles are joint locations, and the black lines are recovered kinematic chains. (Best viewed in color.)

captured, some full and some partial. Thus we have 97 frames and 467 points. In total 31.75% of data are missing in the measurements. The camera motion is also relatively small (about 45 degrees), making this very challenging data for NRSFM problem.

We process this data with θ set from 0.5% to 0.9% times $\|W\|_F/(2F \times N)^{1/2}$, resulting in 2D reprojection error ranging from 0.60% to 1.02%, and 3D error ranging from 2.85% to 3.94%. Figure 7 shows the segmentation and reconstruction with θ set to 0.8% times $\|W\|_F/(2F \times N)^{1/2}$. The left foot exhibits larger 3D error than other rigid subsets because its motion relative to camera is very small.

Comparatively, Fayad *et al.* [13]'s reconstruction from the same dataset based on the same input and same measure of error results in a 3D error of 7.13%. Our 3D error is about half of theirs. Our segmentation further divides the upper body and the arm, capturing the subtle movement of the skin.

5 Conclusion

We have presented a method to systematically reconstruct dynamic 3D structure from orthographic 2D measurements based on the articulated model assumption. We introduce the ellipsoid property as a property to effectively distinguish feature points belonging to different rigid subsets. This enables us to convert the motion segmentation problem to ellipsoid model fitting, and propose a practical method with computational complexity $O(N)$. Then, joint constraints are used to build kinematic chains and 3D articulated structures are recovered. The method is proven effective on both standard real tracking data and challenging human motion capture datasets with missing data, where the proposed method yields better motion segmentation and significant improvements in both 2D and 3D errors compared to existing methods.

6 Acknowledgements

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 712911E) and CRCG of the University of Hong Kong. We thank João Fayad for his kindness in sending us the datasets necessary to carry out the experiments.

7 References

[1] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2000.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In Neural Information Processing Systems, 2008.

[3] V. Rabaud, and S. Belongie. Re-thinking non-rigid structure from motion. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[4] P. Tresadern, and I. Reid. Articulated structure from motion by factorization. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[5] J. Yan, and M. Pollefeys. A factorization-based approach for articulated non-rigid shape, motion and kinematic chain recovery from video. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.

[6] D. Ross, D. Tarlow, and R. Zemel. Learning articulated structure and motion. In International Journal of Computer Vision, 88(2):214-237, 2010.

[7] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[8] Robot arm model, http://www.blendswap.com/blends/view/67329/

[9] A. Kirk, J. O'Brien, and D. Forsyth. Skeletal parameter estimation from optical motion capture data. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[10] Carnegie Mellon University motion capture database, http://mocap.cs.cmu.edu/

[11] P. F. U. Gotardo, and A. M. Martinez. Kernal non-rigid structure from motion. In Proceedings of IEEE International Conference on Computer Vision, 2011.

[12] J. Fayad, C. Russell, and L. Agapito. Automated articulated structure and 3D shape recovery from point correspondences. In Proceedings of IEEE International Conference on Computer Vision, 2011.

[13] S. I. Park, and J. K. Hodgins. Capturing and animating skin deformation in human motion. In Proceedings of the ACM SIGGRAPH Conference on Computer Graphics, 2006.

Motion Vector based Abnormal Moving Vehicle Detection in Nighttime

Cuong Nguyen Khac[1], Ju H. Park[2], Ho-Youl Jung[3]
[1,3]Department of Information and Communication Engineering, Yeungnam University, South Korea
[2]Department of Electrical Engineering, Yeungnam University, South Korea
[3]Correspondence: Prof. Ho-Youl Jung (hoyoul@yu.ac.kr)

Abstract – *Collision prediction during nighttime driving is helpful to avoid accidents because the driver has time to prepare upcoming situations. This research intends to find a real-time solution for frontal and lateral collision warning in an intelligent vehicle. Motion vectors of scene objects are estimated from time-varying frames. Small motion vectors are eliminated by using empirical threshold values pre-determined based on distribution of motion magnitudes. The remaining motion vectors are segmented to rectangular regions by using the unsupervised clustering K-means algorithm. After ROI setting, all segment candidates are classified into vehicle or non-vehicles by using SVM algorithm. From our experiments on real driving situations, the collision risk from lateral and preceding abnormal moving vehicles can be predicted with 91.96% accuracy. The detected abnormal moving vehicles include on-coming, lane change, abrupt speed change, roadside-parking, and overtaking.*

Keywords: Collision prediction, motion information, empirical threshold, K-means clustering, SVM classification.

1 Introduction

Over the past decade, a large number of researches have proposed to support drivers in various driving situations.

For intelligent monitoring system (surveillance), an unsupervised model of activity perception by vehicle trajectories has been proposed for vehicle behavior detection [1]. Similarly, abnormal moving vehicles have been identified by other ideas. For example, a novel kernel density estimation approach to vehicle trajectory learning and motion [2], a graph based approach [3], movement string based approach [4], background subtraction and information chain of tracked vehicles analysis [5], local features based approach [6], short-term continuous velocity and trajectory analysis [7], and so on. In surveillance systems, the aforementioned approaches aim to detect automatically the abnormal behaviors of moving vehicles that can lead to collisions or abnormal traffic events.

In case of dynamic scenes, a camera is usually placed behind the windshield of a vehicle. Common types of accidents are rear-end and lateral collisions that take place on roads and freeways. Therefore, it is crucial to recognize the moving vehicle behaviors for collision avoiding. There have been many existing methods that proposed for vehicle behavior detection in various conditions.

Under daytime condition, future behavior of an ego-vehicle in an inner-city environment is predicted using a sequence of elementary states termed behavior primitives [8]. Perspective Transformation and Template Matching are used to detect vehicle behaviors based on road-markings and stop signs [9]. Grey System Theory and a method of car-following behavior in multilane are introduced to detect vehicle behaviors [10]. Under nighttime condition, lots of factors cause collision such as low illumination, moving vehicles with high speed, drowsy and neglecting of drivers, dangerous glare from on-coming vehicles. In [11], HSV color information and magnitude of vector are used to find the warning threshold of dangerous headlight glare. Tail-lights of preceding vehicles are identified by using multi-level image processing algorithms and clustering processing. Then, related distance between the host vehicle and the preceding vehicle are estimated for collision warning [12]. Head-lights of rear moving vehicles are extracted by using the principle of Blob Analysis, DOF (Depth of Field) theory and identify the distance with perspective technique. The distance information is used to obtain the collision warning of overtaking vehicles on two-lane highway while driving during nighttime [13]. Besides, vehicle detection, tracking and behavior analysis are obtained from monocular vision, stereo vision, and active sensor–vision fusion [14].

Especially, driving in nighttime on freeway with non-lane discipline barriers is a high risky task. Our previous work was abnormal driving vehicle detection using motion information and a fixed value of threshold [20]. This paper contributes an improvement of the previous work. The previous practical threshold is extended to low and high threshold values. These thresholds eliminate more effectively the interferences. Moreover, we apply a new idea of ROI (Region of Interest) setting. This ROI is useful to reduce significantly the number of segment candidates. Thus, the number of training samples in training dataset is reduced, also. This idea helps to optimize the training time faster. This paper also performs the evaluation of proposal approach that we did not do it in the previous work. The proposal method is

useful for collision warning and increase the concentration of drivers while driving in nighttime.

The rest of the paper is organized as follows. In the next section, we describe the related theories that are Lucas-Kanade sparse optical flow and K-means clustering algorithm. In Section III, we explain the proposal method. In Section IV, we describe the experimental results and evaluation of proposal method. Finally, the paper is concluded in Section V.

2 Related theories

In this paper, the motion vectors of moving objects are estimated by using Pyramidal Lucas-Kanade optical flow algorithm. K-means clustering algorithm is used to group the similar motion vectors together. This section aims to summary briefly the theories of aforementioned algorithms. From that, we can see how these techniques are able to segment moving objects from frame by frame.

2.1 Lucas-Kanade optical flow algorithm

Optical flow is the motion of image points over successive frames. Popular gradient based optical flow techniques were introduced by Horn-Schunck [15] and Lucas-Kanade [16]. The performance evaluation of these optical flow techniques can be found in [17]. The basic idea of Lucas-Kanade is described here. Let $I(x,y,t)$ is an image or a frame of video, $m=[x,y]^T$ is a pixel coordinate.

The first assumption of Lucas-Kanade algorithm is that the intensity of pixel m will not be changed during temporal domain dt. This assumption is represented by below equation

$$I(x+v_x dt, y+v_y dt, t+dt) = I(x,y,t) \tag{1}$$

The second assumption of Lucas-Kanade algorithm said that an object does not move very far from frame to frame. In another words, object's motion is small change follow time. The representing of this assumption is described as below

$$I(x,y,t)+\frac{\partial I}{\partial x}v_x dt+\frac{\partial I}{\partial y}v_y dt+\frac{\partial I}{\partial t}dt+O(dt)^2 = I(x,y,t) \tag{2}$$

Because the higher order terms $O(dt)^2$ is very small, then the equation (3) can be simplified as follow

$$\frac{\partial I}{\partial x}v_x+\frac{\partial I}{\partial y}v_y+\frac{\partial I}{\partial t}=0 \text{ or equivalent form } \nabla I.v_m+\frac{\partial I}{\partial t}=0 \tag{3}$$

where $\frac{\partial I}{\partial x}v_x$, $\frac{\partial I}{\partial y}v_y$, $\frac{\partial I}{\partial t}$ are difference between successive frames of spatial derivative follow x direction, y direction and temporal derivative of image I, respectively.

And $\nabla I=\left[\frac{\partial I}{\partial x},\frac{\partial I}{\partial y}\right]^T$ is image gradient at pixel m.

The above derivative operations are applied to corner points only. This is also the reason why Lucas-Kanade algorithm is called the sparse optical flow estimation.

The equation (3) is called Lucas-Kanade optical flow constraint. If equation (3) can be solved, then the motion vector of each corner point will be obtained as follow

$$v_m=[v_x,v_y]^T=\left[\frac{dx}{dt},\frac{dy}{dt}\right]^T \tag{4}$$

However, equation (3) is single equation with two unknowns. We need more equations to find the motion vector v_m. The third assumption of Lucas-Kanade is used in this case. It said that neighbor points belong to the same object should have similar motions. Thus, this assumption helps to obtain more equations by writing a Lucas-Kanade optical flow constraint for each of point within a small window Ω surrounding the considering corner point. After writing more equations, the system becomes over-determined. To solve this system, Least-Squares minimization is used to find the best vector v_m which is closest to the real solution of the system. Let $W(m)$ is a window function where $(m\in\Omega)$. The idea of Least Squares minimization is solved by finding the min of residual function as follow

$$\min_v E=\sum_{m\in\Omega}W^2(m)\left(\Delta I.v+\frac{\partial I}{\partial t}\right)^2 \tag{5}$$

Writing out the derivatives of residual function follow x and y direction, we have

$$\frac{\partial E}{\partial v_x}=\sum W^2(m)\left(\frac{\partial I}{\partial x}v_x+\frac{\partial I}{\partial y}v_y\frac{\partial I}{\partial t}\right)\frac{\partial I}{\partial x}=0 \tag{6}$$

$$\frac{\partial E}{\partial v_y}=\sum W^2(m)\left(\frac{\partial I}{\partial x}v_x+\frac{\partial I}{\partial y}v_y\frac{\partial I}{\partial t}\right)\frac{\partial I}{\partial y}=0 \tag{7}$$

Finally, the equation system need to solve will be

$$A^TW^2Av=A^TW^2b \tag{8}$$

Using algebra, the motion vector for point m can be obtained as

$$v=(A^TW^2A)^{-1}A^TW^2b \tag{9}$$

where

$$W=diag\left(W(m_1),\ldots,W(m_N)\right)_{N\times N}$$

$$A=\begin{bmatrix}\frac{\partial I_1}{\partial x_1} & \frac{\partial I_1}{\partial y_1}\\ \vdots & \vdots\\ \frac{\partial I_N}{\partial x_N} & \frac{\partial I_N}{\partial y_N}\end{bmatrix}_{N\times 2}$$

A disadvantage of Lucas-Kanade method is that the flow is described within a small size local window, so the fast moving points cannot be detected. However, this weak point is fixed by using Lucas-Kanade optical flow algorithm and pyramidal technique [18].

2.2 K-means clustering

K-means clustering is an algorithm to group the N objects based on features into K groups (K<N). The grouping operation is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Mathematically, K-means algorithm clusters N data points into K disjoint subsets S_j containing data points so as to minimize the sum of squares criterion

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left| x_n - \mu_j \right|^2 \quad (10)$$

where x_n is a vector representing the n^{th} data point and μ_j is the geometric centroid of the data points in S_j. K-means is the unsupervised clustering algorithm. The data vectors are automatically assigned to clusters without prior knowledge of data. K-means clustering is simple and fast algorithm. Details of K-means clustering algorithm and evaluations of K-means clustering with respect to others clustering algorithms can be found in [19].

3 Proposal method

The proposal flowchart is shown in Fig.1.

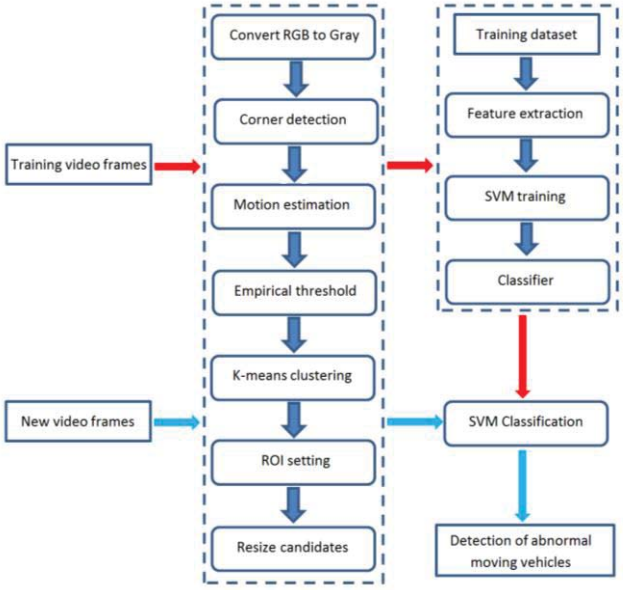


Figure 1. The proposal flowchart

Various videos of real driving situations are recorded for training purpose. Each RGB video frame is converted to gray-scale. The gray-scale frame is passed through Shi-Tomasi corner detector to obtain a set of key points per each frame. Each of corner point in frame f_i is considered to determine its new position in which it moves to within frame f_{i+1}. If two corner points in frames f_i and f_{i+1} satisfy the Lucas-Kanade optical flow constraint, then the motion vector will be obtained. A sample of raw motion vectors is shown in Fig. 2.

The vehicle regions in one thousand frames are selected manually to build a database of vehicle ground truths. The lengths of motion vectors belong to vehicle regions are considered to find the practical threshold values. All of motion vectors in one thousand frames whose positions locate inside the ground truths are used to compute the average μ and standard deviation σ of length distribution. Then, low and high threshold are obtained as follow

$$low\ threshold = \mu - \sigma \quad (11)$$
$$high\ threshold = \mu + \sigma$$

Any motion vector has the length shorter than the low threshold or longer than the high threshold is considered as interference. In practice, most of motion vectors locate outside the vehicle regions are eliminated by applying the above proposal threshold. The result is shown in Fig. 3. However, the practical threshold is not strong enough to eliminate all of non-vehicle motion vectors. Therefore, the vehicle regions need to learn by using machine learning.

After applying threshold, the remaining motion vectors are clustered by using unsupervised clustering algorithm K-means. The similar (position and direction) motion vectors are grouped into the same group. Using spatial coordinates of the top-most, bottom-most, left-most, right-most corner point in current frame f_{i+1}, each group is represented by a bounding box as shown in Fig. 3. Each bounding region is considered as a segment candidate. The candidates are collected manually into positive samples (vehicle bounding boxes) and negative samples (non-vehicle bounding boxes). The number of samples can be reduced by using a practical Region-of-interest (ROI) setting. All samples are normalization by resizing in the same size. Each sample is represented as a feature vector by using its original RGB color information. The aforementioned dataset is trained by using Support Vector Machine (SVM) to obtain the classifier. This classifier will be used to detect moving vehicles in new videos.

4 Experimental results and Evaluations

The motion vectors of all moving objects are extracted by using Lucas-Kanade optical flow algorithm. The sample result is as follow

Figure 2. Motion vectors of all moving object are estimated by using Lucas-Kanade optical flow algorithm

After applying the proposal practical threshold, the remaining motion vectors are clustered to obtain the segment candidates. The sample result is as follow

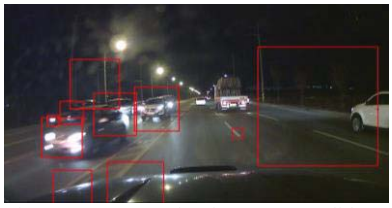


Figure 3. An example of segment candidates

The ROI setting is considered carefully that is suitable to reject the obvious interferences from street lamp regions and the preceding part of the host vehicle. The remaining candidates that locate inside the ROI are resized to size of 50×50. These normalization candidates are collected manually from one thousand video frames to build the dataset for SVM training stage. The SVM classifier is used to detect the abnormal moving vehicles in the new video that are shown as follow

Figure 4. Detection result of change lane vehicle

Figure 5. Detection result of change speed vehicle

Figure 6. Detection result of on-coming vehicles

Figure 7. Detection result of overtaking and on-coming vehicles

Figure 8. Detection result of roadside parking vehicle

As shown in the above Figures, the detected vehicles are satisfied the criteria of proposal method. The non-detected vehicles are moving vehicles that have very small motion. It means that those vehicles are not dangerous with respect to the host vehicle.

EVALUATIONS OF PROPOSAL METHOD

Rate of testing video	Time to view one frame	Time to process two successive frames	Total time to view + process	Rate of result video
30 fps	2 ms	0.0127 ms	2.0127 ms	29.8 fps

Table 1. The processing time is fast for real-time system

Total testing vehicle ground truths	True Positive	True Positive Rate
4,815	4,428	91.96%

Table 2. The detection rate of moving vehicles

Total non-vehicle samples	True Negative	True Negative Rate
5,000	4,236	84.72%

Table 3. The True Negative Rate for Non-Vehicles samples

From Table 1, the processing time for each of two successive frames is 0.0127ms. The rate of the test video is 30fps. Thus, time to view one frame is 60ms / 30fps = 2ms. Total time for viewing and processing per one frame is 2ms + 0.0127ms = 2.0127ms. Finally, the rate of result video including viewing and processing is 60ms / 2.0127ms = 29.8fps. It is fast for real-time system. Besides, the detection rate is 91.96% (Table 2). The False Positive Rate is 8.04%. In this case, the small motion vehicles also contribute to this 8.04%.

5 Conclusions

This paper proposes a new approach for real-time detection of abnormal moving vehicles in dynamic scene and nighttime driving. The main idea is using motion information of moving objects, practical thresholds and machine learning to classify the abnormal moving vehicles. If a vehicle has strong motion then it can be dangerous with respect to the host vehicle. While other vehicle has no motion or small motion then it is not necessary to detect. The proposal approach in this paper has just detected the strong motion vehicles in ROI regions, only. Based on the experimental results, we can see that most of dangerous moving vehicles that can cause collision to the host vehicles are detected. Whereas the vehicles that have the same motion with the host vehicle are not detected, because they are safe with respect to the host vehicle. The proposal system is useful for lateral and frontal collision warning with respect to the abnormal moving vehicles. Especially, it is more important for drivers who are driving in nighttime and on non-lane discipline barriers freeways.

6 Future works

In Fig. 8, Table 2 and Table 3, we can see the wrong detection results. The main reason is that we used a simple feature extraction method for learning vehicle and non-vehicle regions. In this paper, we used original RGB color information for feature extraction. In the next works, we need to investigate what feature extraction methods are good to fix the error. For example, local feature or gradient based feature and so on. We also need to enhance the evaluation method, because most of non-detected vehicles are normal vehicles. Non-detected vehicles have no motion or small motion. They are not abnormal moving vehicles. They should be eliminated out of the set of ground truths for evaluation.

7 Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011- 0011096)

8 References

[1] W. Desheng, W. Jia. "A new method of vehicle activity perception from live video". International Symposium on Computer Network and Multimedia Technology, pages 1-4, Jan 2009.

[2] J. Zhou, K. Wang, S. Tang. "Trajectory learning and analysis based on kernel density estimation". International Conference on Intelligent Transportation Systems, pages 1-6, 2009.

[3] L. Brun, B. Cappellania, A. Saggese, M. Vento. "Detection of anomalous driving behaviors by unsupervised learning of graphs". International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 405-410, 2014.

[4] J. Hao, S. Hao, C. Li , Z. Xiong, E. Hussain. "Vehicle behavior understanding based on movement string". International Conference on Intelligent Transportation Systems, pages 1-6, 2009.

[5] S. Hai-feng, W. Hui, W. Dan-yang. "Vehicle abnormal behavior detection system based on video". International Symposium on Computational Intelligence and Design (ISCID), pages 132-135, 2012.

[6] L. Cui, K. Li, J. Chen, Z. Li, "Abnormal event detection in traffic video surveillance based on local features". International Congress on Image and Signal Processing (CISP), pages 362-366, 2011.

[7] H. Li, Q. Wu, A. Dou. "Abnormal traffic events detection based on short-time constant velocity model and spatio-temporal trajectory analysis". Journal of Information & Computational Science, pages 5233-5241, 2013.

[8] M.G. Ortiz, J. Schmudderich, F. Kummert, A. Gepperth. "Situation-specific learning for ego-vehicle behavior prediction systems". International Conference on Intelligent Transportation Systems (ITSC), pages 1237-1242, 2011.

[9] R. Ishizaki, M. Morimoto, K. Fujii. "An evaluation method of driving behavior by in-vehicle data camera". International Conference on Emerging Trends in Engineering and Technology (ICETET), pages 293-297, 2012.

[10] C. Yu, J. Wang. "Drivers' car-following correlative behavior with preceding vehicles in multilane driving". IEEE Intelligent Vehicles Symposium - Dearborn, Michigan, USA, pages 64-69, 2014.

[11] S. Pharadornpanitchakul, A. Duangchit, R. Chaisricharoen. "Enhanced danger detection of headlight through vision estimation and vector magnitude". International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE), pages 1-4, 2014.

[12] Ying-Che Kuo, Hsuan-Wen Chen. "Vision-based vehicle detection in the nighttime". International Symposium on Computer Communication Control and Automation (3CA), pages 361-364, 2010.

[13] P. Saengpredeekorn, J. Srinonchat. "A new technique to define the overtake distance using image processing". International Conference on Electrical Engineering, Electronics, Computer, Telecommunications and Information Technology, pages 1142-1145, 2009.

[14] S. Sivaraman, M.M. Trivedi. "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis". IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 4, December 2013.

[15] B.K.P. Horn and B.G. Schunck. "Determining optical flow". Artificial Intelligence, vol. 17, pages 185–203, 1981.

[16] B.D. Lucas, T. Kanade. "An iterative image registration technique with an application to stereo vision". Proceeding 7th Conference on Artificial Intelligence (IJCAI), Vancouver, British Columbia, pages 674-679, 1981.

[17] J.L. Barron, D.J. Fleet, S.S. Beauchemin, T.A. Burkitt. "Performance of Optical Flow Techniques". IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 236-242, 1992.

[18] J.Y. Bouguet. "Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm". Intel Corporation Microprocessor Research Labs, 2001.

[19] T. Warren Liao. "Clustering of time series data a survey". Journal of The Pattern Recognition Society, pages 1857-1874, 2005.

[20] Cuong Nguyen Khac, Ju H. Park , Ho-Youl Jung. "An effective detection method of abnormal driving vehicles in nighttime". International Conference on Green and Human Information Technology (ICGHIT), Feb 2015.

Low-Complexity and Low-Delay Structure from Motion Approach for Advanced Driver Assist Systems

J. Andrade[1], C. Prakash[1], F. Akhbari[2], and L. Karam[1]
[1]Arizona State University, Tempe, Arizona, USA.
[2]Intel Corporation, Chandler, Arizona, USA.

Abstract - *Structure from Motion (SfM) is a fundamental capability in Advanced Driver Assist Systems (ADAS) which require accuracy and performance. This paper presents an improved solution to depth estimation based on a low-complexity and low-delay (LOCAD) 3D SfM approach that requires as few as two images from a calibrated camera and that makes use of a new Multi-scale Fast Feature Point Detector (MFFPD). Results are presented to illustrate the improved performance of the proposed method as compared to existing depth estimation methods.*

Keywords: depth estimation, SfM, ADAS, feature detector

1 Introduction

ADAS have enjoyed considerable growth during the last ten years providing commercial applications such as collision awareness and intervention systems. These systems utilize onboard sensors such as tachometers, cameras and RADAR systems to measure depth and relative velocity between vehicles. Video cameras and vision systems provide cost-effective solutions and are employed in a wide array of ADAS use cases [1], including object detection, tracking and depth estimation.

There are a variety of methods devoted to depth estimation using cameras, most of which use stereo cameras [2] allowing a straightforward utilization of the epipolar geometry constraints [3]. However a monocular vision system can also be used to recover 3D information about the scene in a more cost effective way. The accuracy of a depth estimation algorithm directly correlates with the quality of the features it is working with. An efficient feature detector must be able to produce a rich set of features covering all critical corners in a given frame. Invariance to factors such as motion, scaling, illumination and geometric transformation is key to the quality of the edges and corners detected by a quality feature detector. Not only the feature detector must be able to detect distinct features that can be tracked across multiple frames with minimum effort, but it also must do so in real time, matching the input video stream's frame rate.

In an earlier work [4], the authors proposed a robust sparse depth estimation system for ADAS using a real-time key-point detector and a multi-frame based SfM method. However, this latter method exhibits a relatively high computational complexity as it makes use of iterative optimization and a high-complexity metric upgrade stage in order to make the depth estimation robust to drifts in camera calibration parameters. In addition, the multi-frame (also known as multi-view SfM) method [4] introduces excessive pipeline delay as it requires 4 to 8 video frames for performing depth estimation. Moreover, the feature detector used in [4] can be shown to produce falsely detected key-points along diagonal edges.

In this paper, we address the above issues by providing a fast LOw-Complexity And low-Delay (LOCAD) depth estimation algorithm that requires as few as two images or two video frames from a calibrated camera. The proposed LOCAD 3D algorithm also consists of an improved Multi-scale Fast Feature Point Detector (MFFPD) that is robust to edge orientation and that results in a significantly reduced false key-point detection. The proposed LOCAD 3D SfM method can perform depth estimation from a single video stream that is captured using a calibrated camera. The new solution is low-delay as it does not require more than two frames from a video stream. Moreover, the proposed LOCAD 3D algorithm exhibits a significantly reduced complexity since it eliminates the need for the computationally intensive metric upgrade.

This paper is organized as follows. Section 2 presents a background on feature detectors and SfM. Section 3 presents the proposed MFFPD that is used as part of the proposed SfM approach. Section 4 describes the proposed LOCAD 3D SfM method. Results and comparison with existing depth estimation methods are presented in Section 5. A conclusion is given in Section 6.

2 Background

2.1 Key-point Detection

The Harris corner detector [5] is one of the most popular detectors for corner detection. The algorithm uses the structure tensor matrix to approximate the second-order derivative of sum of squared differences over a local neighborhood. Many improved versions of the Harris corner detector were proposed such as the one presented by Shi and Tomasi [6] which employs non maximal suppression to generate a ranking for a list of good corners.

There exists many algorithms [7] [8] that provide high accuracy in terms of repeatability and invariance towards scale and rotation. While these algorithms provide good accuracy, they are computationally expensive and are time consuming on a general CPU platform with limited or most

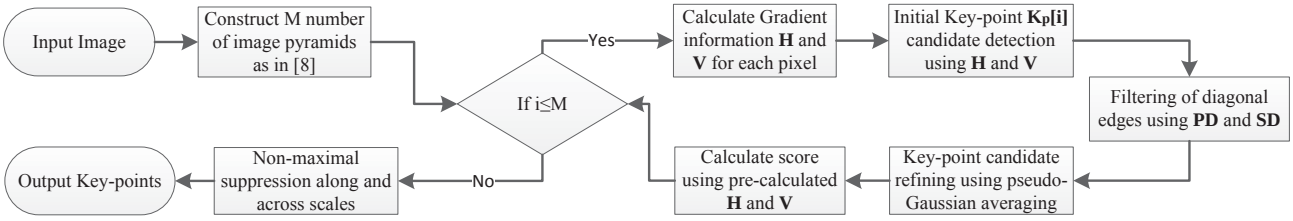


Figure 1. Block diagram of improved MFFPD.

likely no parallel processing capabilities. Therefore, such detectors might not be appropriate for an application with hard real-time constraints such as ADAS use cases.

Numerous other key-point detectors were proposed in the recent past with emphasis on real-time performance and low-computational complexity. These algorithms exploit the pixel intensities around the neighborhood of the pixel of interest [9] [10] [11] [12]. One such algorithm proposed by is called the Features from Accelerated Segment Test, abbreviated as FAST [10]. The algorithm calculates the absolute differences between the center pixel and 16 pixels that form a circle around the reference pixel. Machine learning techniques are used to pick the order in which the circular neighborhood pixels are selected for processing in order to accelerate the computation.

While FAST is known to provide high repeatability, it has a few drawbacks, including false detection at the corners [11]. In order to overcome this drawback, Rublee *et al.* proposed the oFAST detector [11] which uses the Harris measure [5] and intensity centroid [13] to filter falsely detected corners and hence make the detection more robust. However, the addition of the Harris measure and intensity centroid increases the computational intensity. Another major drawback of FAST is its dependency on training dataset since it utilizes machine learning techniques for corner detection. AGAST [12] was proposed in order to overcome this drawback which frees the algorithm from depending on any dataset. AGAST also provides improvements in terms of execution speed.

A key-point descriptor is often used for robust key-point matching across different views of the scene with consideration for scale and rotation. SIFT [7] and SURF [8] provide highly robust key-point matching but are computationally intensive. Recent state-of-the-art algorithms such as BRISK [14] and FREAK [15] serve as real-time key-point descriptor while producing reasonable accuracy and repeatability. BRISK also implements a multi-scale version of the FAST that can be used to obtain scale invariance [14].

Recently, Nain *et al.* proposed a key-point detector, referred to as Fast Feature Point Detector or FFPD for short, that uses only a 4-pixel neighborhood, instead of the 16-pixel neighborhood used in FAST and AGAST, to detect key-points [16]. Additionally, in FFPD, non-maximal suppression of key-points is performed by calculating an information content (IC) based score in order to determine the significance of each detected key-point [16]. The authors of [16] have shown that the algorithm is resilient to noise and uses only 28 operations per pixel. However, FFPD was developed for single-scale applications and, hence, it cannot provide scale invariance when coupled with a key-point descriptor.

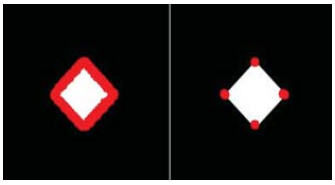
A multi-scale fast feature point detector, known as MFFPD was proposed recently by the authors to address the above mentioned issues [4]. Similar to FFPD, MFFPD uses a 4 pixel mask to detect key-points. As one of the improvements over FFPD, MFFPD eliminates the need for calculation of IC which, in turn, improves the speed of execution. It was shown [4] that MFFPD has a robust scoring pattern along the same scale as well as across multiple scales and that it exhibits better speed and accuracy as compared to the state-of-the-art multi-scale FAST when using the BRISK descriptor [4]. In this paper, we propose an improvement to the MFFPD. The improved detector provides improved performance with respect to reduced false key-point detection over diagonal edges. The improved MFFPD detector does not require many branching operations, is highly parallelizable, and makes use of simple masking operations. These characteristics make the implementation highly favorable for computing resources that support SIMD architecture, for example, GPUs. More details about the improved MFFPD key-point detector are presented in Section III.

2.2 Structure from Motion (SfM)

In [17] a real-time SfM of a static scene from a relatively narrow region is presented. The method requires a collection of several hundreds of images under short baseline displacement. The approach does not use feature tracking but depends on stable photometric information between neighboring frames. This makes this method not resilient to local illumination changes.

In [18] a real-time SfM system specifically designed to work with a calibrated handheld camera and in a small AR (Augmented Reality) workspace is presented. The tracking and mapping tasks are processed in parallel threads using a multi-core computer. A restrictive condition over the scene is that it should be mostly static.

As in the aforementioned approaches, we also assume that internal camera parameters are kept fixed and therefore a pre-calibrated camera is used; however, in contrast to the previous approaches, depth estimation in ADAS is expected to work reliably under wide baseline displacements, continuously changing scenes, and using a very small number of video frames (in our case, only 2) for acceptable real-time performance.



(a) (b)

Figure 2. Comparison of key-point detection accuracy against diagonal edges between (a) Original MFFPD and (b) improved MFFPD on a synthetic image.

Table 1. Comparison of execution time between the original MFFPD and improved MFFPD using a 4-core Intel i7-3770 with 3.48 GHz processor and 8GB RAM

	Original MFFPD	**Improved MFFPD**
Synthetic Image	0.0011	0.0008
Natural Image	0.025	0.018

In order to estimate the depth information that is lost during the image acquisition ($\mathbb{R}^3 \rightarrow \mathbb{R}^2$), correspondences among multiple images are used [19], [20], [3].

Most of the SfM approaches require some kind of feature detection to be performed on every image or every video frame that is used. The SIFT key-point detector presented by Lowe [7] has been dominantly used for this purpose.

For best 3D reconstruction, the detected features have to be matched across all images. This task was initially accomplished using methods like patch correlation, which were only valid under restrictive conditions e.g., short baseline rigid motion. However feature detectors that provide a feature descriptor [21]; a.k.a. feature signature, allow matching under more general conditions such as wide baseline matching which is nowadays a requirement for depth estimation for ADAS. Feature correspondences across multiple images are used to estimate the camera pose for every image, i.e., rotation and translation, which allows performing the 3D projective reconstruction using methods like triangulation. In order to upgrade the projective reconstruction to a metric reconstruction, the intrinsic information about the imaging sensor should be used. When this information is not available it can be estimated using auto calibration methods [22]. Finally, in batch or non-real-time applications, a global optimization process known as bundle adjustment [23] can be used to refine results and produce a jointly optimal scene reconstruction; however, this approach is slow and may not be feasible in applications such as the ADAS space, where actions have to be triggered in real time.

3 Multi-Scale Fast Feature Point Detector

A block diagram illustrating the main stages of the proposed MFFPD key-point detector, is shown in Figure 1. A scale-space pyramid of *n* octaves and *n* intra-octaves is first built as described in the BRISK framework [14] using the input image in order to obtain high accuracy. All comparisons and experimental results in this paper are generated using

(a) (b)

Figure 3. Comparison of key-point detection performance against diagonal edges between (a) Original MFFPD and (b) Improved MFFPD on a natural image.

n=3. Typically, *n* is in the range of 2 to 5 and is selected based on the image size and required accuracy.

The method for detecting key-points is as in [4] but the detection quality is improved by additional checks for diagonal edges. The key-point detection is performed in multiple stages to make the algorithm highly robust to noise.

The first stage of MFFPD uses a 2×2 neighborhood around each pixel to detect corners. The four gradients H1, H2, V1 and V2 are calculated using the 2×2 neighborhood as in [4]. A pixel propagates through the further stages of the algorithm only if *((H1 OR H2) AND (V1 OR V2) > P1)*, where *P1* is a user-defined tunable parameter that is empirically determined based on the number of key-points required and the quality of the image per camera and lens parameters. The idea behind the masking operations is that most of the pixels in the image having small or no gradient intensity provide low response in both directions. Horizontal and vertical edges provide high response in a single direction while a corner key-point provides a high response in both directions. However, the initial MFFPD detector [4] tends to provide high response to diagonal edges as all masking operations namely *H1*, *H2*, *V1* and *V2* produce high responses. Therefore, the algorithm tends to falsely classify diagonal edges as corner key-points.

In this paper, we propose an improvement to the existing MFFPD by adding an additional check to detect and filter diagonal edges. Thus, the second stage of the key-point detector consists of 4 masking operations to capture the gradient intensities along the primary diagonal and the secondary diagonals as follows:

$$PD1 = \left|I_{(x,y)} - I_{(x-1,y-1)}\right| \quad (1)$$

$$PD2 = \left|I_{(x,y)} - I_{(x+1,y+1)}\right| \quad (2)$$

$$SD1 = \left|I_{(x,y)} - I_{(x-1,y+1)}\right| \quad (3)$$

$$SD2 = \left|I_{(x,y)} - I_{(x+1,y-1)}\right| \quad (4)$$

A pixel is considered an initial key-point candidate if *((PD1 OR PD2) AND (SD1 OR SD2) > P1).* These additional checks ensure that the initial key-point candidates selected do not contain pixels belonging to diagonal edges.

It is also important to note that, while the addition of diagonal checks adds to computation, it is performed only on the pixels that are corners or diagonal edges, and not on all the pixels in the image. As most of the pixels in the image are already filtered using the first stage of the detector, the diagonal check operations are performed only on a small fraction of pixels and hence the computational latency is

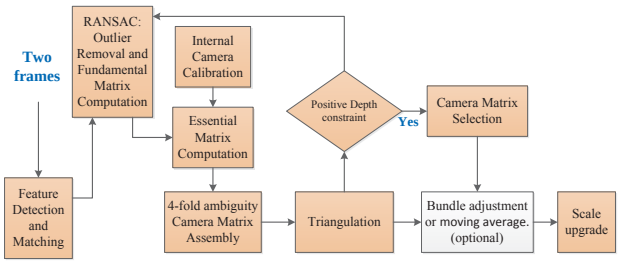


Figure 4. Flowchart of the proposed LOCAD 3D depth estimation approach.

minimal. Moreover, by performing the diagonal check and filtering pixels belonging to diagonal edges, we eliminate all further processing on these pixels. Therefore, the improved MFFPD detector produces key-points that are robust to diagonal edges while maintaining the speed of the original MFFPD.

The flow of the remainder of the algorithm is similar to that of [4]. The initial key-points are refined using Pseudo-Gaussian masks in order to remove pixels that are falsely classified as key-points due to noise. The product of H1, H2, V1 and V2 is allocated as the score for the refined key-points in each scale space. It is shown in [4] that this scoring pattern is more robust in terms of non-maximal suppression than the IC-based score of FFPD. The proposed scoring pattern also eliminates the need for additional computations that are required for calculation of IC.

The improved MFFPD detector is highly robust to false detections along diagonal edges while maintaining the speed of the original MFFPD. In order to check the robustness of the improved MFFPD, key-point detection was performed on a synthetically generated image of size 200×200 containing a white square rotated at 45° from the horizontal. Figure 2 shows the comparisons between the original and the improved MFFPD using the same threshold parameters. It can be seen from Figure 2 that, in contrast to the original MFFPD, the improved MFFPD does not result in false detections along diagonal edges. The detectors were also compared using natural images and the improved MFFPD produced better accuracy with respect to false detection of corners. An example is shown in Figure 3. A comparison of the execution speed between the original MFFPD and the improved MFFPD is shown in Table 1. The improved MFFPD results in a lower execution speed as compared to the original MFFPD as the computation required in the further stages of the key-point detection are eliminated for a large number of falsely detected key-points along the edges. All experiments were conducted on the same computer platform with a 4-core Intel i7-3770 3.48GHz processor and 8GB RAM.

4 Low-Complexity and Low-Delay Structure from motion Using MFFPD Key-points

We propose a low-complexity and low-delay (LOCAD 3D) SfM algorithm that is able to provide depth estimations

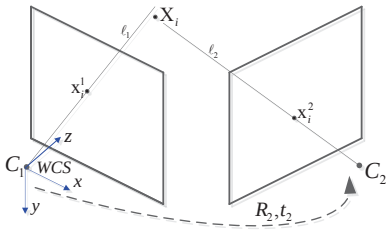


Figure 5. Stereo vision framework where the WCS is attached to camera 1. Back-projected lines $\boldsymbol{\ell_1}, \boldsymbol{\ell_2}$ passing by the camera centers $\boldsymbol{C_1}, \boldsymbol{C_2}$ and the corresponding 2D matched features $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$ in each image, do not intersect due to the presence of noise.

using as few as two frames from a single calibrated camera. Figure 4 presents the flowchart of the proposed LOCAD 3D SfM algorithm. The pipeline utilizes two images I_1, I_2 of the same scene with time stamps t_1, t_2. I_1 and I_2 do not need to be consecutive images as long as there is minimum partial commonality in the scenes. In other words, both images have information of the same scene but from different poses of the camera. The camera pose change between I_1 and I_2 can be described in terms of the rotation matrix $R \in \mathbb{R}^{[3\times3]}$ and the translation vector $t \in \mathbb{R}^{[3\times1]}$.The two frames I_1 and I_2 are analyzed in terms of their features $\mathbf{x}_i^1$ and $\mathbf{x}_j^2$, respectively. The features, i.e., $\mathbf{x}_i^1$, $\mathbf{x}_j^2$, are detected using the proposed improved MFFPD detector and are matched using the BRISK feature descriptor, providing the correspondences $\mathbf{x}_k^1 \leftrightarrow \mathbf{x}_k^2$ $k = 1,2,\ldots,L$ which are utilized to compute the Fundamental matrix $F \in \mathbb{R}^{[3\times3]}$. The F matrix together with the internal camera parameters, which are computed off-line, are used to define the Essential matrix, which provides the external parameters R, t within a four-fold ambiguity framework [3]. The aforementioned ambiguity can be solved experimentally by testing positiveness on 3D reconstructed points' depths. A triangulation stage is used to obtain 3D points from the 2D correspondences once the camera matrices for each pair of images I_1, I_2 have been defined. Since the E matrix provides pose parameters up to an unknown scale factor, it has to be corrected using a ground truth depth. More details about the different components of the proposed approach are further given in the subsequent sections.

4.1 Internal camera calibration

The toolbox provided by [22] was used to determine the K matrix. The K matrix is assumed to remain constant during the image or video acquisition. In other words the assumption is that the camera's zoom and focus are fixed.

4.2 Computation of Fundamental and Essential matrices

The F matrix can be computed with as few as eight noiseless feature correspondences between I_1 and I_2 [3]. However, in order to compensate for noise, MFFPD is used jointly with RANSAC [24] for a more robust estimation of F. In RANSAC, a subset of correspondences, greater than eight, is used to compute F as follows:

$$\mathbf{x}_k^1 \cdot F \cdot \mathbf{x}_k^2 = 0, k = 1, \dots, L \quad (5)$$

The computed F is then scored by the number of inliers, i.e., the number of correspondence pairs that satisfy (5) within a predefined tolerance, the F matrix that gets the higher number of inliers is used to compute the E matrix. Since a unique camera is used The Essential matrix $E \in \mathbb{R}^{[3\times3]}$ is computed as:

$$E = K^T F K \quad (6)$$

4.3 Camera projection matrices

Camera projection matrices P_1, P_2 extracted from F are subject to the well-known projective ambiguity whereas the Essential matrix provides camera projection matrices up to a known scale factor and a four-fold ambiguity.

The unknown scale factor ambiguity is considerably easier and faster to correct compared to the projective ambiguity which is also known as metric upgrade [25]. The four-fold ambiguity on the other hand, means that given E and considering that the world coordinate system is attached to one of the images e.g., I_1, the camera projection matrices are given by:

$$P_1 = K[R_1|t_1] \quad (7)$$

$$\begin{aligned} P_2^1 &= K[R_{2a}|t_2] \\ P_2^2 &= K[R_{2a}|-t_2] \\ P_2^3 &= K[R_{2b}|t_2] \\ P_2^4 &= K[R_{2b}|-t_2] \end{aligned} \quad (8)$$

where $R_1 = I$ is a 3×3 identity matrix, $t_1 = \mathbf{0}$ is a 3×1 zero vector, and R_{2a}, R_{2b} and t_2 are obtained from the Singular Value Decomposition (SVD) of E [3] .

4.4 Triangulation

In order to solve the four-fold ambiguity over the camera matrices, the 3D reconstruction of the matched features $\mathbf{x}_k^1 \leftrightarrow \mathbf{x}_k^2$ is performed for all four possible pairs of $P_1 \leftrightarrow P_2^{1\sim4}$. Figure 5 illustrates the 3D reconstruction from 2D correspondences, where C_1 and C_2 are the camera centers of the equivalent cameras corresponding, respectively, to frames I_1 and I_2 (note that only one camera is acquiring the frames but each frame can be thought of as being acquired by a separate camera). C_1 corresponds to the [0,0,0] 3D coordinates i.e., the origin of the World Coordinates System (WCS), and C_2 corresponds to the right null zero of P_2 [3].

As shown in Figure 5, for every pair of 2D correspondences on the set $\mathbf{x}_k^1 \leftrightarrow \mathbf{x}_k^2$ two 3D lines ℓ_1, ℓ_2 can be generated passing by C_1, and C_2, respectively. In the noiseless case the intersection of ℓ_1 and ℓ_2 defines the 3D position of the point. However when noise is present the intersection of ℓ_1 and ℓ_2 is not guaranteed. In this case, a triangulation algorithm that minimizes the geometric error is utilized [3]. The originally detected feature point correspondences are refined by finding neighboring feature points that satisfy the epipolar constraint (5) and that would result in intersecting back-projected lines ℓ_1, ℓ_2.

4.5 Camera projection matrix selection

According to Figure 5, the reconstructed 3D points $\mathbf{X}_i$ should be in front of both cameras; i.e., their Z coordinate with respect to the WCS is always positive**.** This observation is used to select the correct pair of camera projection matrices out of the four possible $P_1 \leftrightarrow P_2^{1\sim4}$. Given a 3D point $\mathbf{X}$ with homogeneous coordinates (X, Y, Z, T) and a camera projection matrix $P \in \mathbb{R}^{[3\times4]}$, the 2D projected point $\mathbf{x}$ has homogeneous coordinates given by:

$$\mathbf{x}(x, y, 1) = P \cdot \mathbf{X}(X, Y, Z, T) \quad (9)$$

The depth of $\mathbf{X}$ in front of the principal plane of the camera can be determined as follows [3]:

$$Depth(X, P) = \frac{sign\big(Det(M)\big) \cdot \mathbf{x}}{T \cdot \|\mathbf{r}^3\|} \quad (10)$$

where $P = [M|\mathbf{c_4}]$ and $\mathbf{r}^3$ is the third row of $M \in \mathbb{R}^{[3\times3]}$. In order to address the noise effects, we look for the pair of camera projection matrices that provide positive depths in all matched features $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$. If that condition is not met the process starting with the computation of the fundamental matrix F is repeated. The random behavior of RANSAC used to compute F provides a slightly different solution for F in every run. At the end, if the set max number of iterations does not result in the total positiveness condition, the requirement is relaxed by selecting the pair of camera projection matrices that provides positive depth in at least 99% of the features.

4.6 Moving average and scale upgrade

In order to produce a jointly optimized scene reconstruction, bundle adjustment [23] can be used. However, bundle adjustment is computationally expensive and a lighter algorithm is warranted. For a faster execution time with minimum impact to the quality of the result, a moving average is applied across sequential in place of bundle adjustment. The scale ambiguity cannot be avoided but it is easily resolved with ground truth depth information of a single 3D point as [26].

5 Results

In order to test the proposed approach two experiments were carried out, one in a controlled indoor environment and one under real outdoor conditions.

Images used in both experiments were captured using a calibrated camera Sony IMX135 Exmor RS CMOS sensor with pixel size 1.12μm, and the resolution of the images was 1280×960. The ground-truth depths of objects in the scene of both experiments were measured manually. The proposed LOCAD 3D algorithm was tested on a sequence of eight frames for the indoor experiment and nine frames for the outdoor one. The key-points were detected using the improved MFFPD on the initial image or frame. In both experiments these features were tracked in subsequent images using the KLT with an error threshold of 10 pixels

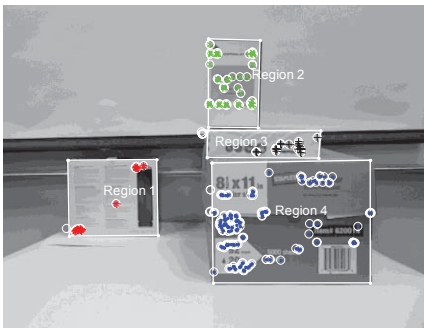


Figure 6. Indoor experiment setup with 4 regions at different depths. Detected features (colored markers according to every region) and re-projected features (white circles) are also shown.

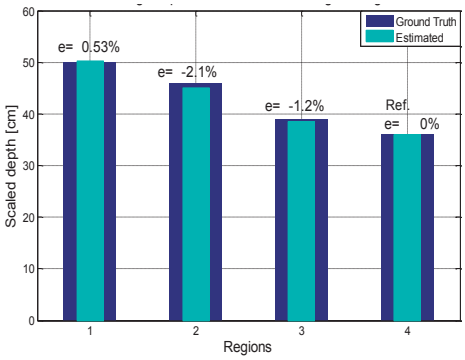


Figure 7. Average depth estimation error for the indoor experiment.

and the average re-projection error threshold for estimating the projective matrix was 0.25 pixels.

The indoor experimental setup, which is shown in Figure 6, has four regions at four different depths. Detected features within every region are also shown in Figure 6. The average depth estimation error for every region is shown in Figure 7 where the ground truth depth of region four was used to scale the computed depth results.

Samples from the outdoor image data set are shown in Figure 8. In order to examine the accuracy of our depth estimation results we have defined three regions in the scene, where points in each region have nearly the same depth. These regions are shown in Figure 9 together with their detected feature sets. For clarity, only 6% of the detected features have been randomly selected and plotted. In Figure 10 the ground-truth of the three defined regions are plotted together with the estimated depth measurements for every pair of correspondences $\mathbf{x}_i^1, \mathbf{x}_i^2$. Due to the fact that features over region 2 (features over the motorcycle) are used as ground-truth references for scaling purposes, that region presents zero mean variation. However, regions 1 and 3 show 3.1% and 0.6% variations, respectively. Motivation for using the term "variation" instead of "error" is the fact that features within a region (in the real-world outdoor scene) do not precisely have the same exact depth.

Table 2 shows a performance comparison in terms of execution time and re-projection error between the proposed method and the one presented in [4]. For this purpose, both

Figure 8. Sample frames from the outdoor real-world dataset.

Figure 9. Regions with similar depths and their corresponding feature sets for a pair of frames.

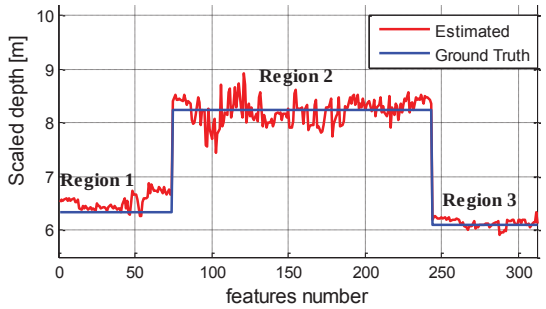


Figure 10. Ground truth and scaled depths where points of region 2 have been used for scaling purposes.

Table 2. Performance evaluation of the proposed algorithm in terms of execution time and re-projection error

Module	Proposed algorithm	[4] algorithm
Repreojection error	0.24 [pixels]	0.09 [pixels]
Detected points	3054	3054
Points for which depth has been defined	1029	1539
Key-point Detection	0.0377 [s]	0.0377 [s]
Tracking	0.0613 [s]	0.0613 [s]
Fundamental Matrix Calculation	0.0094 [s]	0.0067 [s]
Finding Common Points	0.0006 [s]	0.0026 [s]
Triangulation	0.0047 [s]	0.0216 [s]
Bundle adjustment	--	0.0931 [s]
Reprojection	0.0001 [s]	0.0088 [s]
Depth Scaling	0.0002 [s]	0.0032 [s]
Metric upgrade	--	2.2290 [s]
Total time	**0.1141 [s]**	**2.4641 [s]**

implementations were run on a desktop with an Intel i7 2.19GHz processor with 8GB of RAM. Both algorithms were implemented in OpenCV2.4.9 and compiled using the Intel C++ compiler 14.0 in Microsoft Visual Studio 2013. It should be noted that the method of [4] requires several frames (about 6 to 8) for depth estimation as opposed to only 2 frames for the proposed method. In addition, the existing method [4] requires several complex optimization procedures (bundle adjustment and metric upgrade) that are not needed in the proposed method. Since the proposed method does not use bundle adjustment nor metric upgrade, it is substantially

faster while incurring only a minimal penalty in the re-projection error.

6 Conclusion

A low-complexity and low-delay (LOCAD) SfM algorithm that requires as few as two images is proposed for depth estimation. The proposed LOCAD SfM depth estimation method consists of a fast improved multi-scale feature point detector (MFFPD) The improved key-point detector produces better performance in terms of execution speed and reduced false key-point detection and is more robust to edge orientation. Since a single monocular camera is used, the system requires a ground-truth value in order to define the scale factor. Results are over an order of magnitude faster with respect to existing SfM algorithms with just a minimal penalty on the re-projection error.

REFERENCES

[1] J. Fritsch, T. Michalke, A. Gepperth, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko and C. Goerick, "Towards a Human-like Vision System for Driver Assistance," *IEEE Intelligent Vehicles Symposium* , pp. 275-282, 2008.

[2] N. Einecke and J. Eggert, "Stereo Image Warping for Improved Depth Estimation of Road Surfaces," *IEEE Intelligent Vehicles Symposium,* pp. 189-194, 2013.

[3] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge: Cambridge University Press, 2003.

[4] C. D. Prakash, J. Li, F. Akhbari and L. J. Karam, "Sparse Depth Calculation Using Real-Time Key-Point Detection and Structure from Motion for Advanced Driver Assist Systems," in *Advances in Visual Computing*, 2014.

[5] C. Harris and M. Stevens, "A combined corner and edge detector," in *Proceedings of 4th Alvey Vision Conference*, 1988.

[6] J. Shi and C. Tomasi, "Good Features to Track," in *CVPR*, 1994.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV,* pp. 91-110, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, "SURF: Speeded up robust features," vol. 110, no. 3, pp. 346-359, 2008.

[9] S. M. Smith and J. M. Brady, "Susan - a new approach to low level image processing," *IJCV,* vol. 23, no. 1, pp. 45-78, 1997.

[10] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*, 2006.

[11] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, 2011.

[12] E. Mair, G. D. Hager, D. Burshka, M. Suppa and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *European Conference on Computer Vision*, 2010.

[13] P. L. Rosin, "Measuring corner properties," *Computer Vision and Image Understanding,* vol. 73, no. 2, pp. 291-307, 1999.

[14] S. Leutenegger, M. Chli and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *ICCV*, 2011.

[15] A. Alahi, R. Ortiz and P. Vandergheynst, "FREAK: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[16] N. Nain, V. Laxmi, B. Bhadviya, B. Deepak and M. Ahmed, "Fast feature point detector," in *IEEE International Conference on Signal Image Technology and Internet Based Systems*, 2008.

[17] R. A. Newcombe,, S. J. Lovegrove and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," *IEEE ICCV,* pp. 2320--2327, 2011.

[18] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," *6th IEEE and ACM International Symposium on Mixed and Augmented Reality,* pp. 225-234, 2007.

[19] H. C. Longuet-Higgins, "A computer algorithm for reconstruction a scene from two projections," *Nature,* vol. 293, pp. 133-135, 1981.

[20] C. Tomasi and T. Kanade, "Shape and motion from image streams under Orthography: a Factorization Method," *International Journal of Computer Vision,* vol. 9, no. 2, pp. 137-154, 1992.

[21] D. Lowe, "Local feature view clustering for 3D object recognition," *IEEE CVPR,* pp. 682-688, 2001.

[22] J.-Y. Bouguet, "Camera Calibration Toolbox for Matlab," December 2013. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/. [Accessed 5 2014].

[23] B. Triggs, P. F. McLauchlan, R. I. Hartley and A. W. Fitzgibbon, "Bundle Adjustment — A Modern Synthesis," *Vision Algorithms,* pp. 298-372, 2000.

[24] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *J Commun. ACM,* vol. 24, no. 6, pp. 381-395, 1981.

[25] J. Ponce, "On Computing Metric Upgrades of Projective Reconstructions under the Rectangular Pixel Assumption," *Lecture Notes in Computer Science,* vol. 2001, pp. 52-67, 2001.

[26] J. Li and L. J. Karam, "Sparse depth estimation using multi-view 3D modeling," in *IEEE International Conference on Emerging Signal Processing Applications*, 2012.

Human Tracking Using Delphi ESR-Vision Fusion in Complex Environments

T. Wang, R. Aggarwal, and A. K. Somani
Electrical and Computer Engineering, Iowa State University, Ames, Iowa, USA, 50010
{tengw, rka, arun} @iastate.edu

Abstract—*A variety of UGV (Unmanned Ground Vehicle) applications pose the challenge that UGVs need to handle human detection and tracking in complex environments that include dusty, smoky and foggy conditions. These environments make a vision-based human tracking-by-detection system ineffective. To cope with this challenge, we build a radar-vision fusion system, utilizing a 77GHz 2D Delphi ESR (Electronically Scanning Radar) and a CCD camera. Our fusion system utilizes radar returns to generate ROIs (Regions of Interests) and then employs a vision-based human detection technique to validate ROIs. Considering that human are weak targets for a 77GHz radar sensor due to their smaller sizes and weaker reflectivity, we develop a human tracking approach to recover from intermittent human misses. This improves the accuracy of our multi-sensor system. We design experiments to study the behavior of Delphi ESR for human detection. We also characterize Delphi ESR's measurement noise. Using the derived Gaussian noise model parameters, we develop a novel human tracking approach using Kalman filter. We also describe, in detail, an approach to map radar returns to image plane for generating ROIs. A set of real-world experiments show the effectiveness of our approach in human tracking and radar-vision registration.*

Keywords: Radar-Vision System, Delphi ESR, Human Tracking, Kalman Filter, ROI Generation

1. Introduction

An UGV is defined as a vehicle that operates on ground without an onboard human operator presence. Currently, UGVs are being widely deployed for many smart video surveillance systems and a variety of other applications including farming and mining. For such engineering applications of UGVs, real-time and reliable human detection and tracking in their surroundings is required in order to operate them safely and securely.

Human detection and tracking with a monocular camera has been studied widely. In recent years, HOG (Histogram of Oriented Gradients) descriptor [1] and DPMs (Deformable Part-based Models) [2] are widely used in human detection applications, and show excellent performance in static images. Shu et al. [3] and Kim et al. [4] explored DPMs and HOG descriptors to multiple people tracking in crowed scene, respectively: Shu et al. proposed a robust part-based framework to handle partial occlusion and Kim et al. developed a MAP-based online data association approach to address new detection and missing detection problems.

Performances of these vision-based human tracking-by-detection systems are affected by accuracy of their human detection algorithms. Main ideas of these human detection algorithms are to characterize human by pre-defined features and utilize these discriminative features to recognize human targets in images.

It is becoming common for UGV systems to work in weather conditions, such as foggy and cloudy. In addition, the UGV system is focusing more on highly unstructured complex environments, like off-road scenarios. The off-road scenarios offer dusty and smoky environments. One typical example is autonomous surface miners. Surface miners generate dusts when they are operating. Fig. 1 depicts an example of this. These dusty, smoky environments degrade image quality and destroy the discriminative features for human recognition to a great extent. This leads to a significant performance degradation of tracking-by-detection systems.

Fig. 1: Dusty environments where two human targets are visible but not detectable by vision-based human detection algorithms.

1.1 A Solution

Due to the complex nature of the working environment, UGV systems need to be equipped with many different types of sensors to deal with the issues of environmental perception and recognition. A millimeter wave (MMW) radar is an attractive choice due to the fact that it possess all-weather operating capability and can thus penetrate fog, rain, and dust. In recent years, commercial 2D MMW radars have become more available and affordable due to their adoption in automotive applications. However, a 2D radar sensor is

limited to providing the information on radar returns which can be used for detecting ROIs. It does not have the ability to provide object discrimination.

Given the challenges of vision or radar based human tracking systems alone, we develop a radar-vision fusion system to deal with human detection and tracking in dusty, foggy and similar complex environments. Our system deploys a 2D Delphi ESR and a CCD camera, with the CCD camera put directly above the Delphi ESR to maximize their overlapping FOV (Field of View) as shown in Fig. 2. During operation, our radar-vision fusion system is synchronized to receive image and radar data in real time. For each data frame, our system utilizes radar returns to generate ROIs and employs additional local contrast enhancement techniques [5] to enhance features in ROIs. Next, we employ a pre-trained classifier with tunning parameters to validate ROIs. It is worth mentioning that our system allows for a weaker classifier for human detection compared to a vision-based human tracking-by-detection system, as radar data processing helps reduce most of the possible false positives. This is attractive as these dusty and foggy environments destroy some discriminative human features in images and weaker modules are needed to be used.

A fusion system driven by radar needs a set of very reliable radar returns, as any radar misses cannot be recovered by the vision system. Since human are weak targets for 77GHz Delphi ESR, and deliver a discrete set of responses, we develop a human tracking approach to obtain reliable radar returns from human targets.

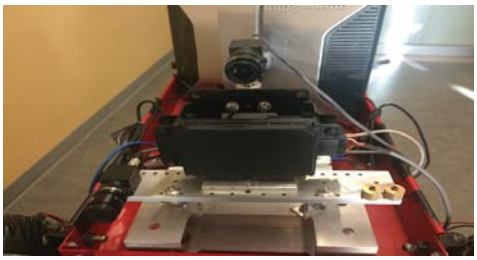


Fig. 2: Our radar-vision fusion system with camera put directly above Delphi ESR (i.e., black rectangular module).

1.2 Paper Organization

This paper is organized as follows. Section 2 presents our experiments with Delphi ESR to understand its behavior for the human detection problem. Section 3 explains our approach to model Delphi ESR's measurement noise. Next we describe our human tracking approach and develop the need for the use of Kalman filter. The mapping between 2D radar space and 2D image plane to generate ROIs is discussed in Section 4. Performances of our human tracking approach as well as radar-vision registration method are analyzed in Section 5. The paper is concluded in Section 6.

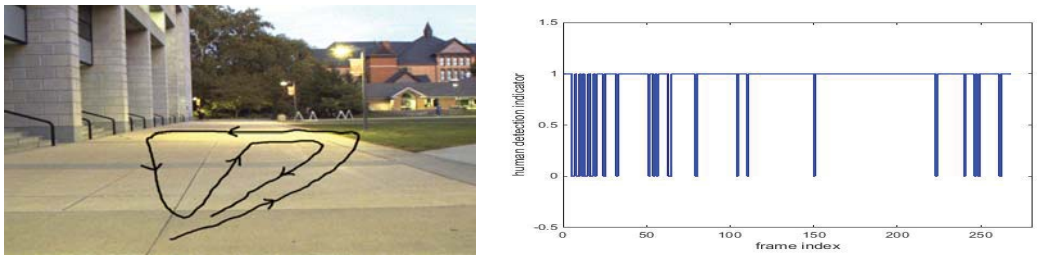
2. Behaviors of Delphi ESR in Human Detection

There are a number of commercially available automotive radar sensors produced by several manufactures. We chose a 77GHz Delphi ESR for human detection purpose. This is because it provides mid-range (60 m) measurements with a wide field of view ($+/-45°$). This characteristics allows human targets across the width of the equipped UGV to be detected. In this project, the ESR unit is programmed to run in 64-Track mode. In this mode the ESR attempts to identify and take measurements of 64 targets every 50 milliseconds. For each track, Delphi ESR provides the range R and azimuth θ information of the target.

We conduct experiments to establish how Delphi ESR performs in applications requiring human detection. The outcomes of one experiment are shown in Fig. 3. The black line in Fig. 3a represents a predefined walking trajectory of a human. During experiment, our system is positioned at a fixed location, approximately two meters away from the start of the trajectory. The human target walks at a constant speed following the depicted path.

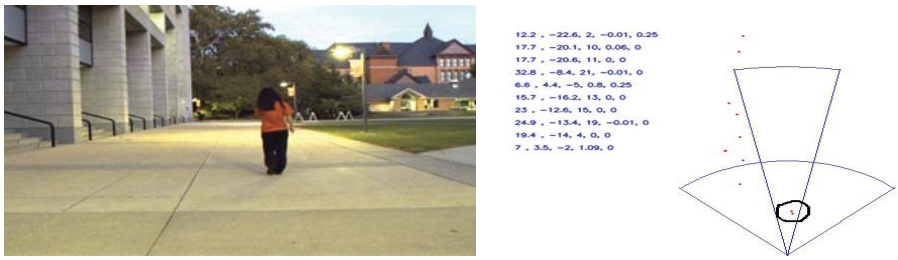
The sequence of ESR returns from the human target is presented in Fig. 3b, where $indicator = 0$ or 1 represents that Delphi ESR misses or successfully detects the human target for a specific image frame, respectively. We make the following observations from the sequence of radar returns.

- The whole experiment process outputs a total of 270 data frames, and the radar missed the human target 26 times. A person is only visible to a radar system when the person scatters back enough signal for the radar transmitter. This implies that the bigger the object is, the better the chances are that the target is detected. The radar misses can be explained by the fact that humans are bordering on the size of "small" objects for a 77GHz radar system. They are not prohibitively small, but small enough to be missed sometimes.
  It is worth mentioning that most of the human misses occurred at the beginning or at the end of the experiment when the human target was close to the radar sensor. For short range, due to ESR's narrow elevation FOV, the portion of the human body that is illuminated by the radar is small. Therefore, the effective radar cross-section is small, which makes the human target even harder to detect. In case human are slightly farther than 2 meters, the detection probability will increase.
- Human misses are intermittent, rather than continuous. For this specific experiment, we noticed that there were at most two consecutive misses. The "intermittent" property of misses suggests that the misses could be recovered with help of additional tracking techniques.
- It is possible that there may be multiple radar returns from a single human target. An example of this phenomena is shown in Fig. 4. Fig. 4a and Fig. 4b show



(a) (b)

Fig. 3: An example of human detection with Delphi ESR



(a) (b)

Fig. 4: One specific example of multiple radar returns from a single human target. Radar returns in the black circle (i.e., $(6.6, 4.4)$ and $(7.0, 3.5)$ expressed in (R, θ) format) are both from the human target, and others are from left buildings.

the image frame and their corresponding radar detection map, respectively. In the radar detection map, both radar returns in the black circle are from the human target. This phenomena can be explained by the fact that a single human in clean background glows to the radar like a lightbulb, which may be interpreted by the radar as multiple targets, each with slight offset. This represents multipath errors. As a result, we need to incorporate a clustering algorithm that can treat each radar return cluster as a single target, when we design a tracking algorithm to recover human misses.

3. Human Tracking Using Delphi ESR

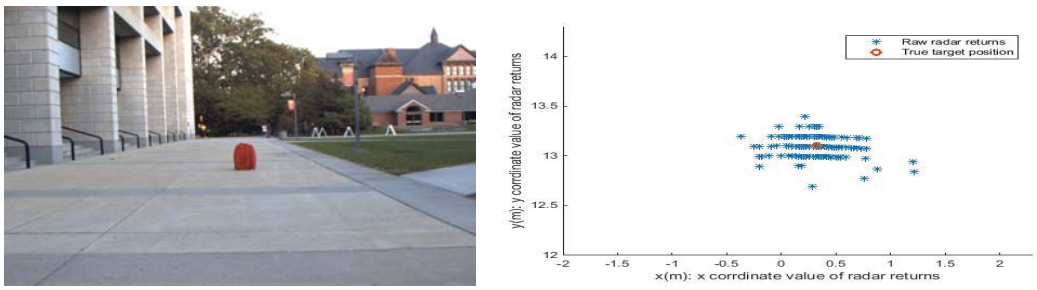
Considering the fact that humans are weak targets for radar detection as well as that human misses are intermittent, we design an efficient tracking model for Delphi ESR to recover from the human misses. We choose to deploy Kalman filtering to recover from the missed data. For the effective use of this approach, we design an experiment to characterize Delphi ESR's measurement noise, as it is an integral component of the tracking approach. In the following, we describe, in detail, our human tracking approach using Kalman filter.

3.1 Measurement Noise of Delphi ESR

To model the measurement noise, we utilize the target object as shown in Fig. 5a to design our experiment. The figure shows a red coat as an object. Inside the red coat is a rectangular metal sheet, which is an ideal target for radar detection due to its great reflectivity. We use the coat on the metal sheet to introduce some measurement noise. During experiment, our radar-vision system and the target object are both positioned at fixed locations. The target object is set at approximately 13 meters (40 feet) away from our system. The distance is chosen such that the effective radar cross-section is large enough for the target object to be detected.

The experiment collects approximate 1200 data frames. This number is large enough to characterize Delphi ESR's noise property based on the experimental observations. The sequence of radar returns from the target object is depicted in Fig. 5b. In this figure, red (circle) point represents the true position of the target object relative to our radar-vision system, and blue (star) points are consecutive radar returns from the target object. Note that we convert radar returns from Polar to Cartesian space by using $x = R * sin(\theta)$ and $y = R * cos(\theta)$, where R, θ are range and azimuth of the target, respectively.

We utilize these radar return samples to model Delphi ESR's measurement noise in X and Y directions (i.e., n_x and n_y), separately. Take n_y as an example. We compute Y-offsets of these radar returns relative to the true position point, and build a frequency histogram using the Y-offset samples. The histogram is plotted in Fig. 6a. We make an observation that the histogram is kind of symmetric with most of the frequency counts bunched in the middle bin centered at 0 and with the counts dying off out in the tails. From the observation, we conclude that it is reasonable to model n_y using normal distribution with mean $\mu_y = 0$. We then compute the variance σ_y^2 and yield that $\sigma_y^2 = 0.0815^2$. The fitted normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$ is plotted



(a) (b)

Fig. 5: Raw radar returns from the target objects for more than 1200 consecutive data frames.

in red in Fig. 6a. The same calculations are also done to build frequency histogram from X-offset samples, and model n_x using normal distribution with mean $\mu_x = 0$ and variance $\sigma_x^2 = 0.1346^2$. The frequency histogram and the estimated normal distribution $\mathcal{N}(\mu_x, \sigma_x^2)$ of n_x are depicted in Fig. 6b, in blue and red respectively. Mean squared errors of estimator $\mathcal{N}(\mu_y, \sigma_y^2)$ and $\mathcal{N}(\mu_x, \sigma_x^2)$ are equal to 0.0527 and 0.0895, respectively.

3.2 Human Tracking with Kalman Filter

Delphi ESR can report as many as 64 pairs of range-azimuth (R, θ) observations in each scan. Each range-azimuth pair represents the location of a scattering center (SC). As mentioned in Section 2, there might be several SCs from a single human target. Therefore, we first employ K-means clustering approach to put radar returns into different target clusters (TCs), with each TC representing a single target. The number of target clusters K is determined by Silhouette Coefficients method. We employ Kalman filtering technique to help radar track its TCs, given that measurement noises of Delphi ESR in X and Y direction both follow the Gaussian distribution.

The state vector of each TC at frame k is defined by the following equation.

$$X_k = [x_k, y_k, vx_k, vy_k]^T \tag{1}$$

In Eq. 1, T stands for the transpose operation; (x_k, y_k) and (vx_k, vy_k) are the TC's center location and velocity in the radar cartesian coordinate system, respectively. The constant-velocity model in Eq. 2 is used to describe the dynamics of the TC.

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ vx_{k+1} \\ vy_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_k \\ y_k \\ vx_k \\ vy_k \end{bmatrix} \tag{2}$$

In Eq. 2, Δt is the time between consecutive frames, and is assumed to be constant. As Delphi ESR outputs the relative position of the target, the measurement state vector is defined by the following equation.

$$Z_k = [x_k, y_k]^T \tag{3}$$

The measurement equation can then be described by the following equation.

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} x_k \\ y_k \\ vx_k \\ vy_k \end{bmatrix} + \mathbf{n} \tag{4}$$

In Eq. 4, $\mathbf{n} = [n_x, n_y]^T \sim N(0, R)$ is two dimensional measurement noise vector, with n_x and n_y representing measurement noise in X and Y direction, respectively. Based on our measurement noise model, we obtain R as follows.

$$R = \begin{bmatrix} \sigma_x^2 & cov(n_x, n_y) \\ cov(n_x, n_y) & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 0.1346^2 & -0.0026 \\ -0.0026 & 0.0815^2 \end{bmatrix} \tag{5}$$

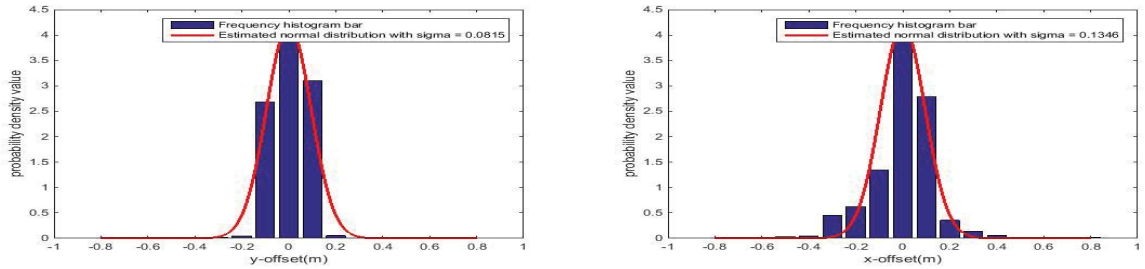
The state transition and measurement equations, as shown in Eq. 2 and Eq. 4 respectively, join together to form the Kalman filter-based tracking model of a TC. The state transition equation can also be treated as a predictor when there are no radar returns from the corresponding target object, and therefore help radar track the target object.

3.3 Multiple Human Tracking Based on MAP Association

In complex environments with multiple human targets, we need to associate current TC observations with recently updated tracking models in order to update the state information of each tracking model. To achieve this, we employ a MAP-based association approach [4]. Its main idea is to encode the problem of multiple people tracking as a node matching problem. The approach adds an extra node for the automatic tracking initialization and formulate the MAP problem to associate detection observations in current time step and tracking models from the last frame.

Before proceeding, we define the following notations to facilitate explanation.

- $tc_k^i = (cx_k^i, cy_k^i)$: ith TC observation from kth radar frame where (cx_k^i, cy_k^i) is the TC's center location.
- $v_k^j = (x_k^j, y_k^j, vx_k^j, vy_k^j)$: jth tracking model after kth data frame, where (x_k^j, y_k^j) and (vx_k^j, vy_k^j) are position and velocity of the tracking model, respectively.



(a) (b)

Fig. 6: Measurement noises of Delphi radar in X an Y directions.

- $P(tc_k^i,\ v_{k-1}^j)$: similarity score between ith current TC observation and jth existing tracking model.

For the radar tracking problem, we define the similarity score as follows to apply the MAP-based association approach.

$$P(tc_k^i,\ v_{k-1}^j) = 1/Z * exp(-dist(tc_k^i,\ (v_{k-1}^j)_{pos} + (v_{k-1}^j)_{vel} * \Delta t) \quad (6)$$

In Eq. 6, Z is the normalizing term, and $dist$ is the Euclidean distance. The subscripts pos and vel stand for the position and velocity of tracking model in Cartesian coordinates, respectively.

In addition, a score is assigned to each tracking model to evaluate its accurate status. When the tracking model is created at frame time of k, its accurate score $L(k)$ is initialized by Eq. 7.

$$L(k) = C \quad (7)$$

In Eq. 7, C is a constant. $L(k)$ is updated as follows.

$$L(k) = \begin{cases} C, & \text{if matching TC observation exists} \\ L(k-1) - 1, & \text{otherwise} \end{cases} \quad (8)$$

Once we obtain the evolution curve of the accurate score, a tracking model is deleted when the accurate score $L(k)$ is decreased to 0. Note that C refers to the maximum number of allowable consecutive human misses. We set C to be 5 in our experiments. This is because we observed from experiments that consecutive misses for human target was smaller than 5 with probability as high as 95%.

Algorithm 1 describe, in detail, our human tracking approach using Delphi ESR based on Kalman filter. At the start of the program (i.e., $k = 1$), we cluster radar returns into TCs and initiate a new tracking model for each TC using the TC's center position information (step 1 & 2). In the following frames (i.e., $k > 1$), we create TCs from raw radar returns, and associate current TC observations with existing tracking models based on MAP-based association approach (step 3 & 4). In case that Delphi ESR successfully detects an existing target in current frame, we update the state vector of the target's tracking model based on new TC observation (case 1); otherwise we predict the new position of the target using its existing tracking model (case 2). In another case that a new target appears, we initiate a new tracking model for the target (case 3)

Algorithm 1: Human Tracking with Kalman Filter

At frame k = 1
  Step1: Employ K-means clustering method to cluster radar returns into TCs. Denote TC set as $U_1 = \{tc_1^i\}$.
  Step2: Create a Kalman filter $v_1^i = (x_1^i,\ y_1^i,\ vx_1^i,\ vy_1^i)$ for each TC $tc_1^i = (cx_1^i,\ cy_1^i)$ by setting $x_1^i = cx_1^i$, $y_1^i = cy_1^i$, $vx_1^i = 0$, $vy_1^i = 0$ and $L^i(1) = C$. The tracking model set is denoted as $V_1 = \{v_1^i\}$;

for *each frame* $k > 1$ **do**

  Step 3: Cluster radar returns into TCs, and denote the TC set as $U_k = \{tc_k^i\}$;

  Step 4: Associate current TCs in U_k with existing tracking models in V_{k-1} using MAP based method;
    Case 1: Tracking model v_{k-1}^i and current TC tc_k^j form a matching pair (normal matching).
    Update state vector of the tracking model based on the position of tc_k^j, and set $L^i(k) = C$.
    Case 2: No current TC associated with tracking model v_{k-1}^i (missing detection).
    Update $L^i(k)$ based on Eq. 8. If $L^i(k) = 0$, delete the tracking model; Otherwise predict the target's position based on state transition equation of v_{k-1}^i.
    Case 3: Current TC tc_k^j is associated with the added node D (new detection).
    Create a new tracking model as step 2 does.

4. Radar-Vision Registration

As the CCD camera is situated directly above Delphi ESR in our system, 3D radar and camera coordinate systems are related via a rotation $\mathcal{R}$, swapping Y and Z of radar system, and a translation T, with displacement only in Y direction. As a result, the problem of registration between 3D radar

space and 2D image plane can be easily converted to the equivalent of the camera calibration problem for generating ROIs. Note that Z coordinate values of radar returns in 3D radar space are always equal to 0, as Delphi ESR is a 2D radar sensor.

We employ the simple pinhole camera model to estimate the linear mapping A between 3D camera coordinate frame and a 2D image plane. The mapping A is in the form as follows.

$$\mathcal{A} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

In Eq. 9, (f_x, f_y) and (c_x, c_y) are camera focal length and optical canters, respectively. With lens model available, radar return (R, θ) will be projected to pixel (u, v, w) in homogeneous coordinate by the following equation.

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \mathcal{A} \times [\mathcal{R} \mid T] \times \begin{bmatrix} R\sin\theta \\ R\cos\theta \\ 0 \\ 1 \end{bmatrix} = \mathcal{M} \times \begin{bmatrix} R\sin\theta \\ R\cos\theta \\ 0 \\ 1 \end{bmatrix} \tag{10}$$

For our specific setup, we have $T = [0, -0.1, 0]^T$ and $\mathcal{R} = [1, 0, 0; 0, 0, 1; 0, 1, 0]$. We employ 2D chessboard pattern to calibrate the CCD camera [6] (i.e., computing $\mathcal{A}$), and finally obtain $\mathcal{M}$ as follows.

$$\mathcal{M} = \begin{bmatrix} 1611.9 & 638.5 & 0 & 0 \\ 0 & 425.9 & 1607.1 & -160.7 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{11}$$

5. Experiments

We consider two scenarios for our experiments to test the effectiveness of the developed human tracking approach. Then we present the results of our experiment to show the performance of our registration approach in locating ROIs.

5.1 Single Human Tracking

We described a human detection experiment using Delphi ESR in Section. 2. In that experiment, Delphi ESR missed the human target 26 times during the experiment. In this section, we use the human tracking approach, described in Algorithm 1, to recover from the human misses.

True human walking trajectory (red) as well as estimates from Kalman filters (blue) are both depicted in Fig. 7. We make the following observations from the result.

- Human misses of Delphi ESR are completely recovered by our tracking approach. We also calculated the mean squared error of Kalman filter estimates for this experiment. The mean squared error is equal to 0.3119m. The estimation error can be explained as follows. When the human target changes its walking direction, the Kalman filter needs some time to catch up with the target. Therefore, one conclusion we draw is that our human tracking approach has the capability to track the human target during the process.
- Our Kalman filter converges with time. That means that our measurement noise model as described in Section. 3.1 is accurate enough to describe Delphi ESR's measurement noise.

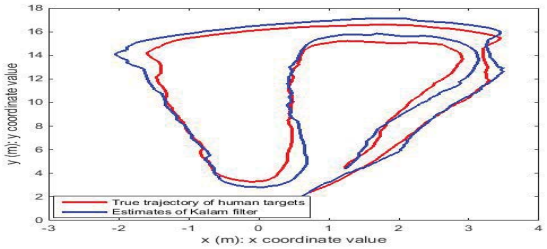


Fig. 7: One example of human tracking using Delphi ESR. Note that the true human walking trajectory is computed by smoothing raw radar returns in a consecutive frame window.

5.2 Multiple Human Tracking

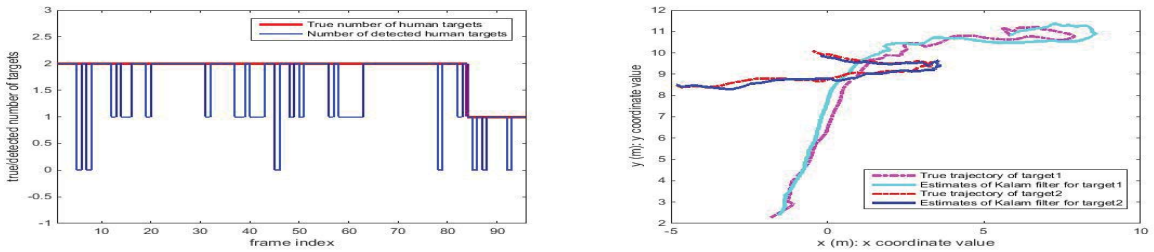
Our second experiment is conducted in dusty environment as shown in Fig. 1. During the experiment, we keep our fusion system at a fixed position. Two human targets walk in different directions, and at a specific time they meet each other and then depart away. This was done to test the effectiveness of MAP-based association approach.

Fig. 8a shows raw radar returns from human targets. We noticed that radar missed the human targets 30 times. We apply the new tracking approach, described in Algorithm 1, to recover from the human misses. Estimates of Kalman filters (i.e., cyan and blue solid lines) as well as true human walking trajectories (i.e., magenta and red dotted lines) are depicted in Fig. 8b. In this figure, black arrows indicate walking directions of two human targets. The two target meet at the bottom intersection point and then depart away from each other. From the result, we observe that the tracking approach can effectively track the targets. The mean squared estimation errors of Kalman filters for the two targets are equal to 0.4200m and 0.2591m, respectively. These are mainly caused by the sudden change in the walking directions.

Another observation is that even when two human targets are close enough, our tracking approach has the capability to differentiate and track them correctly. This can be explained by the fact that MAP-based association method ensures one-to-one mapping between current TCs and existing tracking models.

5.3 Mapping radar returns to image plane

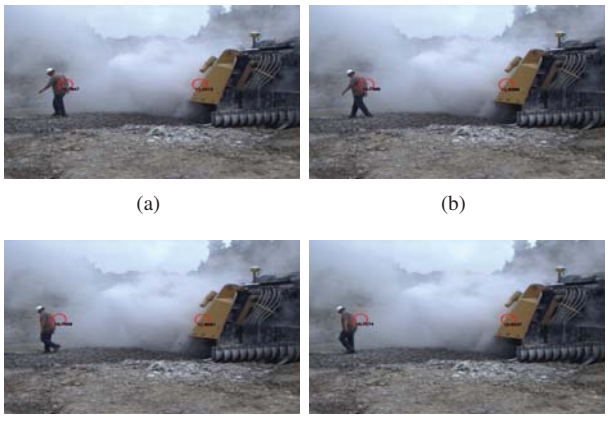
We present an example in Fig. 9 to show the accuracy of our registration approach. This experiment is carried out in an environment where a surface miner is working in field. The human target is walking side by side at average 0.75



(a) (b)

Fig. 8: An example of multiple human tracking using Delphi ESR in dusty environments.

m/s. In this experiment, we firstly employ range filtering to filter out radar returns beyond 20m. This is because we set a goal that we are only interested in human targets within 20m. We then utilize the approach of Algorithm 1 to track each TC. Estimated positions of each TC from Kalman filters are finally mapped to images based on Eq. 10 to generate ROIs.



(a) (b)

(c) (d)

Fig. 9: An example of radar-vision registration for generating ROIs. Locations of ROIs are indicated by red circles.

We present 4 consecutive snapshots in Fig. 9. Locations of ROIs mapped from Delphi ESR returns are indicated by red circles. Registration results correspond to what we expect: radar returns from the human target and surface miner are mapped to these targets in images correctly after registration, although with small offsets. It is worth mentioning that we make the observation that the lateral position error of Delphi ESR for a typical human targets is small and consistent until the target's lateral speed reaches close to 1.0 m/s, at which point the lateral position error gets large. This is due to beam scanning pattern of Delphi ESR. Therefore, one need to compensate the offset for ROI generation purpose when lateral velocity is larger than 1.0m/s, which can be done by formulating the relationship between lateral position error and lateral speed. Once ROI location is identified, we use range information to determine the size of each ROI, and then employ a vision-based human detection techniques to validate each ROI. These are part of our on-going work.

6. Conclusions

We build a radar-vision fusion system to handle human detection and tracking in dusty and smoky environments. Our system utilizes radar returns to generate ROIs and then employs a vision-based human detection technique to validate each ROI. Based on experimental observations that human misses from Delphi ESR are intermittent, and Delphi ESR's measurement noise approximately follows Gaussian distribution, we develop a novel human tracking approach using Kalman filter to recover from human misses. In addition, we develop a radar-vision registration approach via camera calibration to map radar returns to image plane for generating ROIs. A set of real-world experiments showed that our tracking approach has the capability to accurately track human targets, and our registration approach can locate ROIs correctly when lateral velocity is smaller than 1.0 m/s.

7. Acknowledgment

The authors would like to thank Dr. Koray Celik for building the real-time data acquisition system, allowing us to use the system for our research, and providing valuable guidances on our work. Besides, the authors acknowledge Philip and Virginia Sproul Professorship funds from Iowa State University.

References

[1] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Proc. Int. Conf. Computer Vision and Pattern Recognition, pp. 886- 893, 2005.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," IEEE. Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, July 2010.

[3] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based Multiple-Person Tracking with Partial Occlusion Handling, " Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, pp. 1815-1821, 2012.

[4] S.W. Kim, M. Byeon, K. Kim and J.Y. Choi, "MAP-based Online Data Association for Multiple People Tracking in Crowded Scenes," Proc. IEEE Int. Conf. Pattern Recognition, pp. 1212-1217, 2014.

[5] X. Yuan, X. Wei, and Y. Song, "Performance Improvement of Edge-based Human Detection Using Local Contrast Enhancement," Avanced Materials Research, pp. 615-620, 2012.

[6] C. Ricolfe-Viala, and AJ. Sanchez-Salmeron, "Correcting Non-Linear Lens Distortion in Cameras without Using A Model," Optics and Laser Technology, pp. 628-639, 2010.

Block Motion Estimation and Potential Games

Manal K. Jalloul and Mohamad Adnan Al-Alaoui
Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon

Abstract - *In this paper, the problem of inter-prediction block motion estimation (ME) is cast in a game theoretic framework. We model the optimization problem of ME using a network of players in a game theoretic approach. First, a global objective function that captures the notion of consensus is established. Next, it is shown that local objective functions can be assigned to a network of players, so that the resulting game is proven to be a potential game. This game can then be solved in a distributed manner where each player optimizes its own utility function. Consensus strategies can be employed to allow all the players to reach the common minimizer of the global objective function. The resulting scheme is an accurate and highly parallel ME algorithm that decreases the computational burden of the full-search scheme while preserving the quality of the produced motion information.*

Keywords: *Block Motion Estimation, Game Theory, Potential Game, Distributed Algorithm*

1 Introduction

In the video coding industry, ME plays an important role in reducing the temporal redundancy between frames. Among a large number of ME approaches, block-based ME such as the block matching algorithm (BMA) has been adopted in a number of international video coding standards. In the block-matching algorithm, frames are divided into blocks and one motion vector is associated with each block. For each block in the current frame, the motion estimation searches for a motion vector which points to the best match block in the reference frame. The best match block is then used as the predictor for the current block. The full search (FS) block-matching algorithm is the simplest, but computationally very intensive. It provides an optimal solution by exhaustively evaluating all the possible candidates within the search range in the reference frame.

There is a growing need to decrease the computational burden of the FS scheme while preserving the quality of the produced motion information. The motion estimation process needs to be accelerated to meet the real time processing requirements of several cutting-edge multimedia applications.

Block matching motion estimation can be formulated into an optimization problem where one searches for the optimal matching block within a search region which minimizes a certain block distortion measure (BDM), which usually taken as the sum of absolute difference (SAD). Such a problem is classified as non-convex since the objective function is multimodal and has many local minima. Several fast search methods [1-4] were proposed in the literature to speed up the ME process, but they all fall in the problem of local minima. Noticing that, researches turned to global optimization methods to solve the problem of ME. Algorithms based on the genetic algorithm (GA) [5] and simulated annealing (SA) [6] have been proposed. Recently, biologically-inspired optimization algorithms have gained a lot of attention and researches have been trying to apply them to ME. Many variants of particle swarm optimization (PSO) [7-9], artificial bee colony optimization (ABCO) [10], and differential evolution (DE) [11] were used for locating potential global optimum within an arbitrary search space. These algorithms are population-based and mimic the swarming behavior of flocks of birds and herds of fish adapting to their environment. Even though these algorithms have very powerful global optimization capabilities, they are, however, highly centralized and require a central processor to continuously communicate with all the members of the population during the iterative search process. This makes these algorithms very hard to parallelize and thus cannot be accelerated using the available parallel processing technologies. In [12-13], the authors have proposed schemes based on PSO to parallelize motion estimation.

Game theory is a formal framework with a set of mathematical tools to study the complex interactions among interdependent rational players (not necessarily humans). In a game theoretic framework, a set of players each with a set of actions and pay-offs is defined. Game theory is the study of the ways in which strategic interactions among rational players produce outcomes with respect to the preferences (or utilities) of those players. In general the theory provides a structured approach to many important problems arising in signal processing and communications, notably resource allocation and distributed transceiver optimization. Recently, some attempts [14-16] have been done to apply game theory tools to the field of video processing. Nevertheless, to our knowledge, there haven't been any attempts in the literature to model the problem in a game theory framework and try to solve it.

The main idea of our proposed approach is to cast the problem of block motion estimation in a game-theoretic framework. The problem is modeled as an exact potential game that can be solved in distributed manner by a network of players. This approach would be highly parallel since it is distributed (multiple agents or players) and thus suitable for a parallel implementation.

The rest of the paper is organized as follows. In section 2, a review of block motion estimation and game theory are

provided. In section 3, the proposed approach is presented. Section 4 concludes this paper.

2 Background

2.1 Motion estimation and block matching

For motion estimation through a block matching (BM) algorithm, the current frame of an image sequence I^t is divided into non-overlapping blocks of $N \times N$ pixels. For each template block in the current frame, the best matched block within a search window (S) of size $(2W + 1) \times (2W + 1)$ in the previous frame $I^{t-1}$is determined, where Wis the maximum allowed displacement. The position difference between a template block in the current frame and the best matched block in the previous frame is called the motion vector (MV) (see Fig. 1). Under such perspective, BM can be approached as an optimization problem aiming for finding the best MV within a search space.

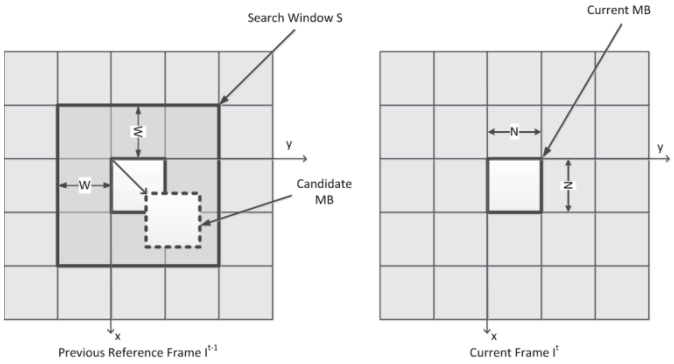


Figure 1 Block motion estimation

The most well-known criterion for BM algorithms is the sum of absolute differences (SAD). It is defined in Eq. (1) considering a template MB at position (x, y) in the current frame and the candidate MB at position $(x + \hat{u}, y + \hat{v})$ in the previous frame I^{t-1}:

$$SAD(\hat{u}, \hat{v}) = \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} |g_t(x + i, y + j) - g_{t-1}(x + \hat{u} + i, y + \hat{v} + j)|, \quad (1)$$

Where $g_t(.)$is the gray value of a pixel in the current frame I^t and$g_{t-1}(.)$ is the gray level of a pixel in the previous frame I^{t-1}. Therefore, the MV $w = (u, v)$is defined as follows:

$$w = (u, v) = arg_{(u,v)\in S} \min SAD(\hat{u}, \hat{v}) \quad (2)$$

Where

$$S = \{(\hat{u}, \hat{v}) | -W \leq \hat{u}, \hat{v} \leq W \; and \; (x + \hat{u}, y + \hat{v}) is \; a \; valid \; pixel \; position \; in \; I^{t-1}\}$$

In the context of BM algorithms, the FSA is the most robust and accurate method to find the MV. It tests all possible candidate blocks from I^{t-1}within the search area to find the block with the minimum SAD. For the maximum displacement of W, the FSA requires $(2W + 1)^2$search points. For instance, if the maximum displacementW is ± 7, the total search-points are 225. Each SAD calculation requires $2N^2$additions and the total number of additions for the FSA to match a 16 × 16 block is 130,560. Such computational requirement makes the application of FSA difficult for real time tasks.

2.2 Game theory

Game theory is a branch of mathematics aimed at the modeling and understanding of resource conflict problems. Essentially, the theory splits into two branches: non-cooperative and cooperative game theory. The distinction between the two is whether or not the players in the game can make joint decisions regarding the choice of strategy. Non-cooperative game theory is closely connected to minimax optimization and typically results in the study of various equilibria, most notably the Nash equilibrium. Cooperative game theory examines how strictly rational (selfish) actors can benefit from voluntary cooperation by reaching bargaining agreements. Another distinction is between static and dynamic game theory, where the latter can be viewed as a combination of game theory and optimal control. In general, the theory provides a structured approach to many important problems arising in signal processing and communications, notably resource allocation and robust transceiver optimization. Recent applications also occur in other emerging fields, such as cognitive radio, spectrum sharing, and in multihop-sensor and adhoc networks [17-18].

3 The proposed game theoretic approach for block motion estimation algorithm

The problem addressed in this paper is how to solve the optimization problem of ME in a distributed way through a network of players in a game theoretic framework.

3.1 The definition of the game

3.1.1 Global objective function:

$$J_{glob}(\hat{u}, \hat{v}) = \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} |g_t(x + i, y + j) - g_{t-1}(x + \hat{u} + i, y + \hat{v} + j)| \quad (3)$$

Where $g_t(.)$ is the gray value of a pixel in the current frame I^t and$g_{t-1}(.)$ is the gray level of a pixel in the previous frame I^{t-1} as shown in Fig.1. That is our global objective is to find one motion vector for the whole block (of dimension NxN). To do that using game theory, we propose to decompose the block into K subblocks and then associate each subblock to a player. Each player would be trying to find the motion vector for its subblock. We want the players at the end of the game to reach consensus that is they should all agree on a common motion vector which is the minimizer of the global objective function. The game should allow the players to communicate during the search process.

3.1.2 Players

We have a set of players$P = (1,2, \ldots\ldots, K)$. Each player is associated to a subblock. The cost function of a player k is the Sum of Absolute Difference of the subblock of dimension LxL at position (x_k, y_k)as shown in fig.2:

$$J_k(\hat{u}, \hat{v}) = \sum_{j=0}^{L-1} \sum_{i=0}^{L-1} |g_t(x_k + i, y_k + j) - g_{t-1}(x_k + \hat{u} + i, y_k + \hat{v} + j)| \quad (4)$$

3.1.3 Utility function

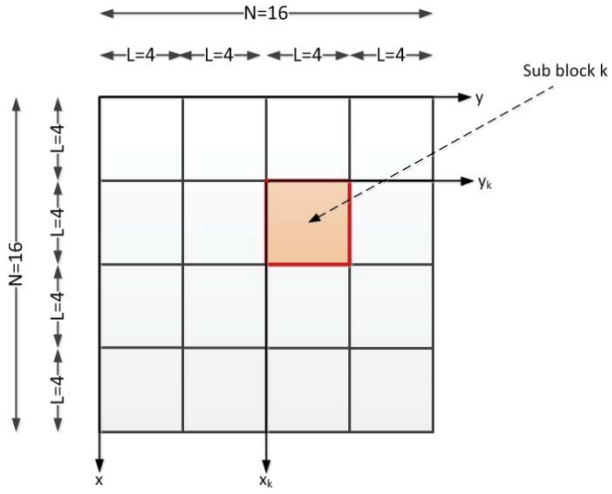
$J_k(\hat{u}, \hat{v})$ cannot be used as a utility function for agent k because it doesn't depend on the action profile of the other agents. In order to introduce such dependency, the utility function of agent k can be chosen as:

$$U_k(\hat{u}, \hat{v}) = J_k(\hat{u}, \hat{v}) + \alpha * \sum_{i \in N_k} ((\hat{u} - u_i)^2 + (\hat{v} - v_i)^2)^{1/2} \quad (5)$$

The utility function includes a regularization term which is the Euclidean distance to the motion vectors of the neighboring subblocks. In other words, the objective of a player is not only to minimize the SAD of its subblock but also to find a motion vector that is in high correlation with the motion vectors of the neighboring subblocks.

3.1.4 Action set

The action set of player k is the set of motion vectors $(\hat{u}, \hat{v})$ within a specified search window as shown in Fig.1.



. Figure 2 Macroblock decomposition into sub-blocks.

3.2 Modeling the problem as an exact potential game

Definition of Potential Games [19]:

Player action sets $\{A_i\}_{i=1}^n$ *together with player objective functions* $\{U_i: A \to \mathbb{R}\}_{i=1}^n$ *constitute a potential functions, constitute a potential game if, for some potential function* φ:

$$U_i(a_i^1, a_{-i}) - U_i(a_i^2, a_{-i}) = \varphi(a_i^1, a_{-i}) - \varphi(a_i^2, a_{-i})$$

For every player $P_i \in P$ *, for every* $a_i^1,\ a_i^2 \in A_i$*, and for every* $a_{-i} \in A_{j \neq i}$.

A potential game, as previously defined, requires perfect alignment between the global objective and the players' local objective functions in the following sense: If a player unilaterally changed its action, the change in its objective function would be equal to the change in the potential function.

The proposed multi-agent block motion estimation problem can be modeled as a potential game by appropriately defining the players' utilities. First, we establish a global objective function that captures the notion of consensus. Next, we show that local objective functions can be assigned to each player, so that the resulting game is, in fact, a potential game.

Consider a consensus problem with n-player set P, where each player $P_i \in P$has a finite action set A_i. A player's action set could represent the finite set of locations that a player could select. We will consider the following potential function for the consensus problem:

$$\varphi(a) = \sum_{i=1}^n J_i(a_i) + \sum_{P_i \in P} \sum_{P_j \in N_i} \frac{\|a_i - a_j\|}{2} \quad (6)$$

Where

$J_i(a_i)$ is the local cost function (SAD) of subblock i which is assigned to player i. This cost function depends only on the action a_i (motion vector) of player i.

N_i is the neighbor set of player i.

Now, the goal is to assign each player an objective function that is perfectly aligned with the global objective:

$$U_i(a_i, a_{-i}) = J_i(a_i) + \sum_{P_j \in N_i} \|a_i - a_j\| \quad (7)$$

The utility function includes a term which is the distance to the motion vectors of the neighboring subblocks. In other words, the objective of a player is not only to minimize the SAD of its subblock but also to find a motion vector that is in high correlation with the motion vectors of the neighboring subblocks.

Now, each player's objective function is only dependent on the actions of its neighbors.

Claim: Player objective functions (7) constitute a potential game with potential function (6), provided that the time invariant interaction graph induced by neighbor sets $\{N_i\}_{i=1}^n$is undirected, i.e.,

$$P_i \in N_j \Leftrightarrow P_j \in N_i$$

Proof: Since the interaction graph is time invariant and undirected, the potential function can be expressed as

$$\varphi(a) = J_i(a_i) + \sum_{k=1,k\neq i}^{n} J_k(a_k) + \sum_{P_j \in N_i} \|a_i - a_j\| + \sum_{P_k \neq P_i} \sum_{P_j \in N_k \setminus P_i} \frac{\|a_k - a_j\|}{2} \quad (8)$$

The change in the objective function of player P_i by switching from action a_i^1 to action a_i^2, provided that all other players collectively play a_{-i}, is

$$U_i(a_i^1, a_{-i}) - U_i(a_i^2, a_{-i}) = J_i(a_i^1) - J_i(a_i^2) + \sum_{P_j \in N_i} \|a_i^1 - a_j\| - \sum_{P_j \in N_i} \|a_i^2 - a_j\| = \varphi(a_i^1, a_{-i}) - \varphi(a_i^2, a_{-i}) \quad (9)$$

This is an exact potential game.

3.3 Learning algorithms to solve the problem

So far, the problem of block motion estimation has been modeled as an exact potential game where a network of players will try to reach consensus on the common minimizer of their local objective functions which is also the minimizer of the global objective function. Learning algorithms for potential games have been extensively studied in the game theory literature [20-21]. Most of these learning algorithms for potential games guarantee convergence to a (pure) Nash equilibrium.

4 Conclusions

In this paper, we study the distributed optimization of block motion estimation in a game theoretic framework. The problem of block motion estimation is formulated as a consensus game. The global objective and utility functions of the players are chosen so that the problem is modeled as an exact potential game. The proposed game theoretic framework of the problem is distributed and non-centralized. Existing ME algorithms based on optimization techniques like SA, PSO, GA, are centralized. The proposed game theoretic formulation is solved using a network of players that use simple update strategies to try, in a cooperative yet distributed manner, to minimize their local utility functions and at the same time reach consensus on the common minimizer of the global objective function. Moreover, the algorithm is highly parallelizable. Because of the distributed nature of the proposed scheme, it is suitable for a parallel implementation. So far, proposed ME algorithms were either serial or had only partial parallelism. The algorithm presented in this paper exhibits high parallelism and thus can exploit the advance in the hardware industry. Future research would investigate the suitable learning algorithm, with high accuracy and fast convergence, to be used to solve this potential game.

5 Acknowledgment

This work was supported by the American University Research Board (URB).

6 References

[1] R. Li, B. Zeng, M.L. Liou, "A New Three Step Search Algorithm For Block Motion Estimation," IEEE Trans. Circuits Syst. Video Technol.,vol.4, no.4, pp. 438–442, 1994

[2] L.M. Po, W.C. Ma, "A Novel Four-Step Search Algorithm for Fast Block Motion Estimation," IEEE Trans. Circuits Syst. Video Technol., vol.6, no.3, pp. 313–317, 1996.

[3] S. Zhu, K. K. Ma, "A New Diamond Search Algorithm for Fast Block-Matching Motion Estimation," IEEE Transactions on Image Processing, vol. 9, pp. 287–290, 2000

[4] C. H. Cheung, L. M. Po, "A novel cross-diamond search algorithm for fast block motion estimation," IEEE Transactions on Circuits and Systems for VideoTechnology 12 (12) (2002) 1168–1177.

[5] A. El Ouaazizi, M. Zaim, & R. Benslimane, "A Genetic Algorithm for Motion Estimation," IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.4, April 2011

[6] Z. Shi, W.A.C. Fernando, and A. Kondoz, "Simulated Annealing for Fast Motion Estimation Algorithm in H.264/AVC," Simulated Annealing - Single and Multiple Objective Problems, Marcos de Sales Guerra Tsuzuki (Ed.), ISBN: 978-953-51-0767-5.

[7] R. Ren, M.M. Manokar, Y. Shi, B. Zheng, "A Fast Block Matching Algorithm for Video Motion Estimation Based on Particle Swarm Optimization and Motion Prejudgement," 2006.

[8] Jing Cai, W. David Pan, "On Fast And Accurate Block-Based Motion Estimation Algorithms Using Particle Swarm Optimization," Information Sciences, vol. 197, pp. 53–64, 15 August 2012.

[9] M. K. Jalloul, M. A. Al-Alaoui, "A Novel Hybrid Dynamic Particle Swarm Optimization Algorithm for Motion Estimation in High Resolution Video", International Conference on Engineering of Reconfigurable Systems and Algorithms, OPT-I 2014, Kos Island, Greece, June 4-6, 2014.

[10] Erik Cuevas, Daniel Zaldívar, Marco Pérez-Cisneros, HumbertoSossa, ValentínOsuna, "Block matching algorithm for motion estimation based on Artificial Bee Colony (ABC)," Applied Soft Computing, Volume 13, Issue 6, June 2013, Pages 3047-3059.

[11] Erik Cuevas, Daniel Zaldívar, Marco Pérez-Cisneros, Diego Oliva, "Block-matching Algorithm Based on Differential Evolution for Motion Estimation," Engineering Applications of Artificial Intelligence, Volume 26, Issue 1, January 2013, Pages 488-498.

[12] M. K. Jalloul, M. A. Al-Alaoui, "A Novel Parallel Computing Approach for Motion Estimation Based on Particle Swarm Optimization", International Conference on

Engineering of Reconfigurable Systems and Algorithms, ERSA 2013, Las Vegas, USA, July 22-15, 2013.

[13] M. K. Jalloul, M. A. Alaoui, “A Novel Parallel Motion Estimation Algorithm Based on Particle Swarm Optimization”, International Sysmposium on Signals & Systems, ISSCS 2013, Romania, July 11-12, 2013.

[14] A. Albarelli, E. Rodola, and A. Torsello, "A Game-Theoretic Approach to Fine Surface Registration Without Initial Motion Estimation," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on , vol., no., pp.430,437, 13-18 June 2010

[15] K. Roy, P. Bhattacharya, and C.Y. Suen, "Iris Segmentation Using Game Theory", presented at Signal, Image and Video Processing, 2012, pp.301-315.

[16] A. Chakraborty and J. Duncan, “Game-theoretic Integration for Image Segmentation,” IEEE Trans. Pattern Analysis Machine Intelligence, vol. 21, no. 1, pp. 12 – 30, Jan 1999.

[17] Y. Wu and K. Liu, “An Information Secrecy Game in Cognitive Radio Networks,” IEEE Trans. Information Forensics and Security, vol. 6, no. 3, pp. 831–842, Sep. 2011.

[18] B. Wang, K. Liu, and T. Clancy, “Evolutionary Cooperative Spectrum Sensing Game: How to Collaborate?” IEEE Trans. Commun., vol. 58, no. 3, pp. 890–900, Mar. 2010.

[19] Jason R. Marden, Gürdal Arslan, and Jeff S. Shamma, “Cooperative Control and Potential Games”, IEEE transactions on Systems, Man, and Cybernetics—part b: Cybernetics, vol. 39, no. 6, December 2009

[20] D. Monderer and L. Shapley, "Potential games," Games Econom. Behav., vol. 14, no. 1, pp. 124-143, 1996.

[21] D. Monderer and L. Shapley, "Fictitious Play Property for Games with Identical Interests", J. Econ. Theory, vol. 68, no. 1, pp. 258-265, 1996.

EFFICIENT VIDEO CODING IN REGION PREDICTION IN ONLINE VIDEO SURVEILLANCE

Shivprasad P. Patil[1], **Rajarshi Sanyal**[2], **and Prof. Ramjee Prasad**[1]
[1] Center for TeleInFrastruktur, Aalborg University, Aalborg, Denmark.
[2] Belgacom International Carrier Services, Brussels, Belgium.

Abstract – *There is a recent trend to implement video coding techniques for video surveillance. Various coding techniques had been proposed to enhance estimation accuracy in progressive video coding. As the conventional coding approaches have constraints to optimize image processing for moving bodies whilst the background is typically static. The problem magnifies in case of rotating cameras. The challenge in this case is to segregate the dynamic entities while the background rotates at a fixed pace. This paper presents an approach for the improvement of error free coding in video surveillance by employing rotating cameras. A least mean estimator approach is used and the recurrent full search motion estimator logic is defined for the prediction of foreground moving elements from the video sequence, which comes from a rotating sensor. The benefits that we derive from this method are more accurate detection of the actual moving object, thereby, reducing data redundancy by eliminating the nonessential background information.*

Keywords: video surveillance, least mean estimator, motion estimator, rotating camera, FS-BMA.

1 Introduction

The development of digital video technology has made it possible to use digital video coding in various applications such as teleconferencing, digital broadcast codec, video telephony, and video surveillance etc. In purview of surveillance related applications, video coding finds its use in particularly traffic video surveillance, where there is an urge to optimize image processing and frame reduction given the channel capacity constraints of a typical city traffic surveillance network. With the state of the art method, the background is also considered as moving. There is no process to discriminate the background from a truly dynamic object (moving people, cars). We propose a new technique by, which we can segregate/discriminate the background from dynamic objects in the video frames. We treat dynamic objects as the intelligence and remove the non essential background information from the subsequent video frames. Thus, the background information is not transmitted repeatedly, but only the useful information pertaining to moving objects is conveyed. With this method we can leverage substantial benefits like reducing transmission overheads, and storage etc. of a video surveillance system by employing rotating cameras. To improve the true detection rate of moving objects and thereby, reducing the bandwidth of such a new application, in this work, a new coding technique has been proposed. Most video surveillance systems rely on the ability to detect moving objects in the video stream. Therefore, object detection remains an important information extraction step in a wide range of computer vision applications. Each image is segmented by automatic image analysis techniques. This should be done in a reliable and effective way in order to cope with unconstrained environments, non stationary background, and different object motion patterns. Furthermore, different types of objects are manually considered e.g., persons, vehicles, or groups of people. Many algorithms have been proposed for object detection in video surveillance applications. They rely on different assumptions e.g., statistical models of the background [1,2,3], minimization of Gaussian differences [4], minimum and maximum values [5], adaptivity [6, 7] or a combination of frame differences and statistical background models [8]. Two approaches have been recently considered to characterize the performance of video segmentation algorithms: pixel-based methods, and template based methods or object-based methods. In Pixel based methods we thrive to detect all the active pixels in a given image. Therefore, the problem of object detection is formulated as a set of independent pixel detection. The algorithms can therefore, be evaluated by standard measures used in the Communication theory e.g., misdetection rate, false alarm rate, and receiver operating characteristic (ROC) [9]. Several proposals have been made to improve the computation of the ROC in video segmentation problems e.g., using a perturbation detection rate analysis [10] or an equilibrium analysis [11]. The usefulness of pixel-based methods for surveillance applications is questionable since we are interested in the detection of object regions (in our

case, a moving car is an object region), and not in independent pixel detection. The computation of the ROC can also be performed using rectangular regions selected by the user, with and without moving objects [12]. This improves the evaluation strategy since the statistics are based on templates instead of isolated pixels. As far as the object based segmentation concept is concerned, we do the evaluation of the object of interest. In this approach, most of the works aim to characterize the object on the basis of colour, shape, [13, 14, 15] or area based performance evaluation [16]. This approach is instrumental to measure the performance of image segmentation methods for video coding and synthesis, but it is not usually used in surveillance applications.

It is found that, in video surveillance applications, stationary cameras are often employed to form a network. Due to technological advancement and the cost factor, the rotating camera is finding its place in these applications. Employment of a rotating camera reduces the number of cameras to be installed, thereby, reducing installation and maintenance cost. Also, bandwidth requirement overheads are reduced.

It is imperative that video coding techniques are required for video surveillance especially, for city traffic surveillance where we have channel bandwidth restrictions and processing resource constraints. The video coding technique that we propose, aims to reduce the redundancy and hence, can be termed as channel coding. Motivation for the use of the video coding approach for segmentation purposes lies in more accurate detection of false motion arising due to the rotation of the camera than the other existing methods, including a statistical one. In the conventional approach of the coding technique, two successive frames are compared for estimation of motion. This is referred to as Full Search Block Matching (FSBMA). As our sensor is rotating, road side buildings also seem to be moving, which we call as false motion. The application of FSBMA does not deal with the rejection of false motion, thereby, reducing the accuracy of true motion detection. Hence, there is scope to improve the existing method. To achieve the objective of improvement in the existing coding algorithm, in this work the Recurrent Full Search Block Matching Algorithm(R-FSBMA) approach is proposed.

Basically, in video coding, for finding moving elements, two successive frames are compared, which is also called as the block matching algorithm (BMA). The pixels, which are not matching are taken as moving elements or motion vectors (MV). Obtained MV's are considered for further processing. However, in our case a stable background is also appearing as moving due to camera rotation. Also, we have to look into the movement pattern of moving objects, which can be linear or nonlinear. Hence, to find the actual moving object, conventional two successive frame comparisons are not effective. Hence, the direct application of BMA will not result in correct MV estimation. So, we have to go for comparing frames in a recursive manner, where in we have compared the current frame with its successive frame to detect MV. Here, we record the variations as linear and non- linear. As stable objects will have linear variation with frame, we can reject such coefficients, thereby, eliminating stable pixels falsely taken as moving.

The objective of our project is as follows:

1. Develop a new recurrent block matching approach for more effective and accurate detection of a moving object.
2. Apply an adaptive filtration method to attenuate the noise of the video sample generated from multi-segmented intersection.

The novelty of the proposed work is, rather than searching the successive video frames for motion, we go for searching the motion component in a set of frames, thereby, presenting a new recursive coding technique. Also, we are able to extract a moving object when a video file contains actual and false motion. It is worth to mention here, that our algorithm is not only supporting for a complex scene, where there are multiple segments at the intersection, but also for jitter in a rotating sensor.

The rest of the paper is outlined into six sections. In section 2, a conventional video coding system and its application to video surveillance is presented. In section 3, the error estimator logic is presented. Motion prediction technique is outlined in section 4. The proposed recurrent-FSBM algorithm is described in section 5. The observations obtained for the proposed work are presented in section 6. The conclusion of the developed work is outlined in section 7.

2 Video coding and application

Many applications have been proposed based on the assumption that an acceptable quality of video can be obtained for a bandwidth of about 1.5 Mbits/second (including audio).Computer vision systems often depend on the ability to distinguish or describe a moving object in an image space. An algorithm is designed to segment a moving foreground based on the block-matching motion

method and recursive tracing of the resulting motion Vectors. The objective of this project is the creation of an algorithm that will separate moving foreground from a stationary background in a given video sequence. The separation of true motion associated with foreground, can further be utilized for sending the null values for stationary background, thereby reducing transmission bandwidth. The selection of a motion estimator model represents the first step in the problem. Gradient-based methods such as optical flow have shown high performance, but generally come with an increased computational overhead than block-based matching. The disadvantage of block methods is an expected loss of sharpness at the edge regions marking the boundary between foreground and background. Regardless of the motion estimator, careful attention must be paid to noise effects when estimating motion. Faulty motion vectors due to image noise can lead to visually unpleasant effects such as isolated background blocks in the resulting segmented image. Noise-reduction filters may be used to alleviate this problem. Another method is to examine the resulting mean-squared error of the known zero-motion vector regions. If any error exists, it must be due to the presence of noise in a particular image sequence. Accurate knowledge of all the motion vectors in a sequence theoretically, provides the means to segment the images into pixels associated with a moving object and pixels associated with a rigid background. The algorithm for tracing motion vectors throughout the sequence is highly recursive and can be computationally expensive, depending on the number of non-zero motion vectors present, which could be optimized in future works.

The moving frames are generally represented as a sequence of multiple frames. These frames are static in nature when isolated. All these frames together create a moving image as shown in figure 1. On a closer observation it can be seen that most of the moving frames have got correlated pixels among the successive frames. The transmission of these correlated pixels for low bit rate application is a very difficult task. To overcome this difficulty the moving image can be isolated from the stationary elements and can be transmitted isolately for more efficient low bit rate application.

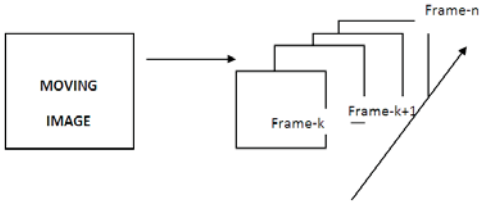


Figure 1: multi-frame representation of a video sample.

Generally, an image has two layers, namely- foreground and background. In case of a moving image, there are three possibilities:

1) Foreground and background moving
2) Foreground stationary and background moving
3) Background stationary and foreground moving

In the proposed work, the first case of both foreground and background moving is considered.

Normally, in video surveillance, vehicles and people are moving, whereas the other objects are stable. So, vehicle and people motion can be considered as true motion. But, in our case camera rotation gives a false motion to the hoardings, building etc. So, wherein conventional video coding is proposed for true motions only, a mixed model of true and false motion estimation has to be devised.

As, it is difficult to apply the conventional coding for video processing, in this paper a new coding is presented, which is briefed in the earlier section. Prior to the estimation approach, de-noising of a noisy video sample is required. In this work, an adaptive filtration based on the LMSE approach is used.

3 Denoising using LMS algorithm

The Least Mean Square (LMS) algorithm is an adaptive algorithm, which uses a gradient-based method of steepest decent. The LMS algorithm uses the estimates of the gradient vector from the available data. LMS incorporates an iterative procedure that makes successive corrections to the weight vector in the direction of the negative of the gradient vector, which eventually leads to the minimum mean square error [17]. Compared to other algorithms the LMS algorithm is relatively simple; it does not require the correlation function calculation nor does it require matrix inversions.

From the method of steepest descent, the weight vector equation is given by;

$$w(n+1)=w(n)+\mu[-\nabla (E\{e^2(n)\}] \quad (1)$$

Where μ is the step-size parameter and controls the convergence characteristics of the LMS algorithm; $e^2(n)$ is the mean square error between the output y(n) and the desired output, which is given by,

$$e^2(n)=[d(n)-w(n)x^T(n)]^2 \quad (2)$$

The gradient vector in the above weight update equation can be computed as

$$\nabla (E\{e^2(n)\})= 2Rw(n) -2r \quad (3)$$

Where R is an autocorrelation of input signal x(n) and r is a cross correlation between the desired response and input. In the method of steepest descent the biggest problem is the computation involved in finding the values r and R matrices in real time. The

LMS algorithm simplifies this problem by using instantaneous values;

$$R = x(n)x^T(n) \quad (4)$$

$$r = d(n)x(n) \quad (5)$$

Therefore, the weight update can be given by the following equation,

$$w(n+1)= w(n)+\mu x(n)[d(n)–x^T(n)w(n)]$$
$$= w(n)+\mu x(n)e(n) \quad (6)$$

The LMS algorithm is initiated with an arbitrary value w(0) for the weight vector at n=0.

The successive corrections of the weight vector eventually leads to the minimum value of the mean squared error.

Therefore, the LMS algorithm can be summarized in the following equations;

$$\left.\begin{aligned} y(n)&=w^T x(n)\\ e(n)&=d(n)-y(n)\\ w(n+1)&=w(n)+\mu x(n)e(n)\end{aligned}\right\} \quad (7)$$

This computed weight provides an optimal value for noise elimination. Over this de-noised video sample a new motion estimation approach is proposed. This approach is an extension to the FS-BMA approach.

4 Motion prediction

The motion estimation and compensation technique has been widely used in video compression due to its capability of reducing the temporal redundancies between frames. Most of the algorithms developed for motion estimation so far are block-based techniques, called the block-matching algorithm (BMA). In this technique, the current frame is divided into a fixed size of blocks, and then each block is compared with candidate blocks in a reference frame within the search area [18,19]. The widely used approach for the BMA is the full search BMA (FSBMA), which examines all the candidate blocks within the search area in the reference frame to obtain a motion vector (MV). The MV is a displacement between the block in the current frame and the best matched block in the reference frame in horizontal and vertical directions. The motion estimation algorithm is performed with a variable size of search area depending on block types varying from an 8x8 block to the complete frame. The video sequences for low bit-rate video coding applications such as videophone and video-conferencing have some restrictive motion characteristics. A block in a specific region in the previous frame can belong to the same region at that position in the current frame; in other words a block in the background region may lie in the background region in the current frame. The changing block shows the percentage of the difference from the background to the active region or vice versa. The other labels mean that the block types are the same in successive frames. In all video sequences, the percentage of background blocks in the successive frames is very high. The changing blocks occupy only 30% below, meaning that the motion field of each block is very high in the successive frames for the other blocks. Also, the pattern of distribution is very similar without regard to video sequences. It is shown that the temporal correlation between the successive frames is very high, that is, if a block in the previous frame belongs to background regions or active regions, the block, which is located in the same position in the current frame may be classified as a background block or active moving block, respectively, with a strong probability.

The basic idea of block matching is depicted in the figure 2, where the displacement for a block (LxL) in frame K (the present frame) is determined by considering a window of size [(L+2W) x (L+2W)] in frame K+1 (the search frame) for finding the location of the best-matching block of the same size. The search is usually limited to $(L+2W)^2$ region called the search window.

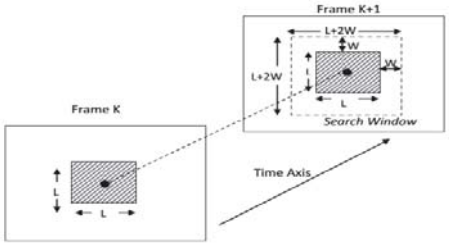


Figure 2: Matching approach.

Block matching algorithms differ in

- The matching Criteria
- The search strategy
- The determination of block size

Matching criteria:

The matching of the blocks can be quantified according to various criteria of, which the most popular and less expensive is mean absolute difference (MAD), given by equation (8). Another criteria mean square error (MSE), is given by equation (9).

$$MAD = \frac{1}{L^2}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\left|P_{ij} - S_{ij}\right| \quad (8)$$

$$MSE = \frac{1}{L^2}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(P_{ij} - S_{ij})^2 \quad (9)$$

Where L is the side of the block, and Pij and Sij are the pixels being compared in the block from the present frame and the block from the search frame, respectively. Normally, MSE is not used, as it is difficult to realize the square operation in hardware.

The determination of block size:

The selection of an appropriate block is essential for any block-based motion estimation algorithm. There are conflicting requirements on the size of the search blocks. If the blocks are too small, a match may be established between blocks containing similar gray level patterns, which are unrelated in the sense of motion. On the other hand, if the blocks are too large, then the actual motion vectors may vary within a block, violating the assumption of a single motion vector per block. The block size for the proposed design is calculated by performing continuous testing, taking a different combination of frame sizes with different frame skips.

5 Recurrent estimation logic

Ideally, the tracer recognizes this and segments the region over all the frames, and not just the frames in which it moved. In general, this stage forms the computational bottleneck of the overall algorithm.

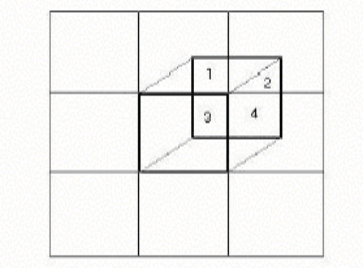


Figure 3: Recurrent searching of an overlapped pixel.

Tracing motion vectors lend itself naturally to a recursive solution. Each block with non-zero motion vectors in each frame represents a "seed" call to the tracing function. A moving block will, in general, translate into a region corresponding to four blocks. The tracing algorithm begins with a seed call. This seed block will move into as many as four other blocks, and each of these blocks is recursively called by the tracing function. The purpose of the tracing function is simply to identify the appropriate moving pixels based on the motion vectors and block regions, and then to seed further calls to it. Motion tracing has a straightforward solution only in one direction temporally. In other words, tracing must be done in both the forwards and reverse temporal directions for best segmentation results.

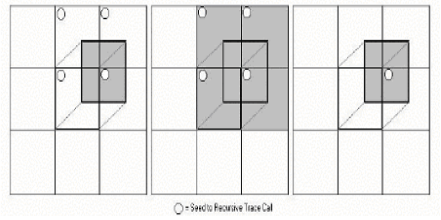


Figure 4: The process of searching in the frames using R-FSBMA.

For any moving block only the pixels corresponding to that moving block are associated with motion, but all four regions impinged by the block are seeded to the successive tracing call. This is the most accurate approach, but also the most computationally burdensome. The second approach is to seed all four blocks as well, but to treat all pixels within the four seeded blocks as having moved rather than just the actual moving pixels. This approximation greatly simplifies the tracing algorithm, and also increases the algorithm efficiency dramatically, since a block that is seeded to the tracing function need not be ever seeded again.

A final approach is to mark all moving pixels as in the general case, but to only seed the block corresponding to maximum overlap. If there are equal overlaps, then multiple blocks are seeded. Although this variation only approximates the tracing problem, it can be much faster since each trace call usually, only seeds one recursive call rather than four. In the most general case, the tracing algorithm runs slow. For improved speed, motion vectors are computed not between each frame, but between every n frames and tracing is done on this smaller set of motion vectors.

6 Simulation observation

To observe the developed work a video sequence is read, wherein a set of video frames is selected and the tracing algorithm is applied. The obtained results are as shown below:

The video file is captured at an elevated location at the center of a cross road, and the sensor is rotated for 360 degrees to capture the traffic images. The video sequence shows the vehicle movement and other static regions in the vicinity. The video sample is captured at 25fps, with a resolution of 272x 352.

Figure 5: Extracted Video frames from the video file.

A set of successive frames is extracted from the captured video sequence. Further, they are used for processing. The extracted frames are illustrated in figure 5.

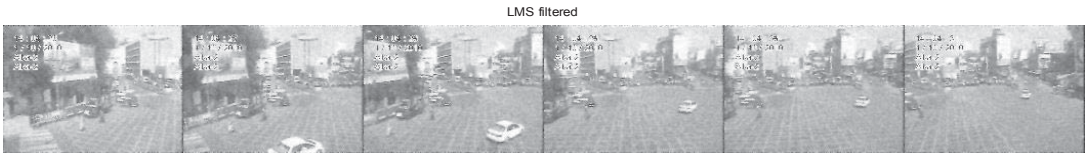


Figure 6: De-noised sample after LMS filtration.

It is required to eliminate the noises so as to achieve higher accuracy in the estimation of moving objects.
To achieve this, a conventional adaptive LMS filter is applied to denoise the affected sample. The obtained result for such filtration is given in figure 6. It is observed that a higher visual quality is achieved with this approach.

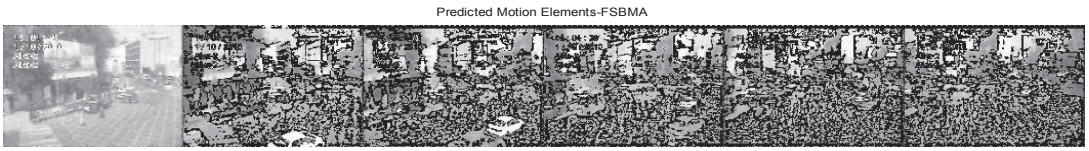


Figure 7: Predicted region by FSBMA approach.

Over the filtered sample, a full search block matching algorthm is applied to compute the moving element. It is observed that as the camera is in a rotating position, the background objects will also change their corresponding position for each frame. Hence, such components are also detected as moving elements in predicted video frames.

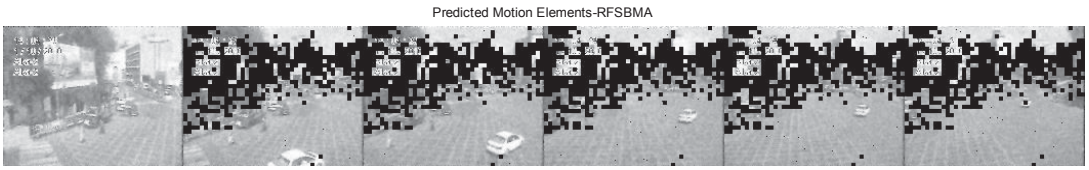


Figure 8: Predicted region after recurrent tracing.

In the case of the proposed Recurrent FSBMA approach, due to successive computation of Motion vector in both inter and intra frames, the elimination of a background element is possible. Hence, this approach detects the moving elements more accurately than the FSBMA approach, which is shown in figure 8.

7 Conclusion

A new coding approach for video surveillance is presented. The incorporation of new coding algorithms for denoising using the least mean error estimator results in higher estimation probability. This denoising approach is a dynamic model and hence, is suitable for all type of system interface. The proposal of recurrent motion estimation logic results in an improvement in the detection of a moving object in a video sequence, generated from a rotating camera. In the sequel, the authors would like to conclude that, the proposed work based on the

recurrent block matching approach is found to be more effective and accurate in the field of video surveillance, by employing rotating sensors. It is worth to mention that, our proposal has applicability in the metropolitan surveillance network with wired or wireless rotating camera implementation, where bandwidth and processing resource optimization are the key challenges.

Acknowledgements

The datasets of the traffic at the intersection is kindly provided by Pune Municipal Corporation, India.

7 Reference

[1] Y. Ren, C. Chua, and Y. Ho, "Statistical background modeling for non-stationary camera," Pattern Recognition Letters, 24(1-3):183– 196, January 2003.
[2] C. Stauffer, W. Eric, and L. Grimson, "Learning patterns of activity using real-time tracking," IEEE Trans. Pattern Anal. MachineIntell., vol. 22, no. 8, pp. 747–757, August 2000.
[3] T. Bouwmans, F. Baf and B. Vachon. "Background modeling using mixture of gaussians for foreground detection - A survey." in Recent Patents on Computer Science 1, 3, pp 219-237,2008.
[4] N. Ohta, "A statistical approach to background suppression for surveillance systems," in Proceedings of IEEE Int. Conference onComputer Vision, pp. 481–486, 2001.
[5] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Who? when? where? what? a real time system for detecting and tracking people,"in IEEE International Conference on Automatic Face and Gesture Recognition, pp. 222–227, April 1998.
[6] M. Seki, H. Fujiwara, and K. Sumi, "A robust background subtraction method for changing background," in Proceedings of IEEE Workshop on Applications of Computer Vision, pp. 207–213, 2000.
[7] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel, "Towards robust automatic traffic scene analysis in real-time," in Proceedings of Int. Conference on Pattern Recognition, pp. 126–131,1994.
[8] R. Collins, A. Lipton, and T. Kanade, "A system for video surveillance and monitoring," in Proc. American Nuclear Society (ANS) Eighth Int. Topical Meeting on Robotic and Remote Systems, Pittsburgh, PA, pp. 25–29,April 1999.
[9] H. V. Trees, "Detection, Estimation, and Modulation Theory". John Wiley and Sons, 2001.
[10] T. H. Chalidabhongse, K. Kim, D. Harwood, and L. Davis, "A perturbation method for evaluating background subtraction algorithms," in Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2003), Nice, France, October 2003.
[11] X. Gao, T.E.Boult, F. Coetzee, and V. Ramesh, "Error analysis of background adaption," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 503–510, 2000.
[12] F. Oberti, A. Teschioni, and C. S. Regazzoni, "Roc curves for performance evaluation of video sequences processing systems for surveillance applications," in IEEE Int. Conf. on Image Processing, vol. 2, pp. 949–953, 1999.
[13] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), Nice, France, pp. 125–132,2003.
[14] P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in Int. Conference on Image Processing, pp. 308–311, 2000.
[15] C. E. Erdem, B. Sankur, and A. M.Tekalp, "Performance measures for video object segmentation and tracking," IEEE Trans. Image Processing, vol. 13, no. 7, pp. 937–951, 2004.
[16] V. Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer, "Performance evaluation of object detection algorithms," in Proceedings of 16th Int. Conf. on Pattern Recognition (ICPR02), vol. 3, pp. 965–969,2002.
[17]B.Widrow and S.D.Stearns, Adaptive Signal Processing, Prentice-Hall, Englewood Cliffs,NJ, 1985.
[18] Jianhua Lu , Liou, M.L. , "A simple and efficient search algorithm for block-matching motion estimation, " IEEE Trans. Circuits and Systems for Video Technology, vol. 7 , pp.429-433, Apr 1997.
[19]Chun-Ho Cheung, and Lai-Man Po, "A Novel Cross-Diamond Search Algorithm for Fast Block Motion Estimation", IEEE Trans. Circuits And Systems For Video Technology, vol 12., no. 12, pp. 1168-1177, December 2002.

Improving Spatial Saliency Using Affinity Model and Temporal Motion

Manbae Kim
Dept. of Computer and Communications Engineering, Kangwon National University
Chunchon, Gangwondo, Republic of Korea
E-mail: {manbae}@kangwon.ac.kr

Abstract –*Saliency map has been applied in diverse fields such as image segmentation, object detection, image scaling and so forth. Over the past decades, a variety of spatial saliency generation methods for still images have been introduced. Recently, motion saliency has gained much interest where motion data estimated from an image sequence are utilized. In this paper, we propose the saliency generation method that enhances the spatial saliency based on the combination of spatial and motion saliencies without the consideration of motion classification. Further, an affinity model is integrated for the purpose of connecting close-by pixels with different colors and obtaining a similar saliency. In experiment, we performed the proposed method on seven image sequence sets. Our saliency map is compared with Zhai's method for evaluating the saliency improvement. Further, from the objective performance evaluation, we validated that the saliency value increases by +41 per a pixel over the spatial saliency on the average.*

Keywords: Spatial saliency, video motion, motion saliency, motion complexity, affinity model

1 Introduction

Humans search for visual attention on important objects in the image. Detecting such salient objects remains as a significant goal, even though the performance is drastically reduced in complex images. Extracted saliency maps are widely used in many computer vision applications including image segmentation, human detection, image re-targeting and so forth. Saliency originates from visual contrast and is often attributed to variations in image attributes like color and edge boundary.

Except some particular images, it is well known that the extraction of satisfactory saliency maps is a difficult task. For instance, large-size foreground objects that have similar color distribution to that of a background are not guaranteed to achieve good saliency. Due to those problems, video motions have been incorporated into spatial-temporal saliency extraction, which has gained much interest over past decades. Most of conventional methods relying on motions heavily depend on a region segmentation that is costly and time consuming as well as lacks reliable accuracy. This results in less application in the practical fields in terms of real-time processing as well as the performance.

Based on statistical data of an input image, a pixel with low frequency is assgind a large saliency and vice versa. A spatial saliecy method is classfied as local, global, or a combination of both.

The critical constraints that can produce good saliency are two-folds; (1) the size of the foreground needs to be smaller compared with a background and (2) the color distribution of the foreground is not similar to that of the background. Since only particular images satisfy such contraints, video motions can be a good supplment to the saleincy estimation. Further, the color difference inside of an identical object might produce different saleincies, yielding the extraction of an unsatisfctory saleincy map.

This paper proposes the improvement of spatial saliency map utilizing temporal motions as well as an affinity model. For the spatial saliency, the method proposed by Zhai *et al*. [1] is adopted and used as a base method. Even though the method is simple, but yet efficint. Therefore it has been frequently utilized in many other works. The method uses a global contrast of an image.

As motion-based saleincym methods, Xia improved Itti's model [2] by refering multiple frames [3]. Zhu *et al*. combined color, brightness and motion directions for a fused saliency map [4]. Zhai used a SIFT feature descriptor to extract feature points and estimated motion by hormomophic estimation [1]. This homograpy is used for the detection of objects of interest that is combined with spatial saliency. Beside those works, most of works heavily depends on the static or temporal segmentation using motions [6-13].

The reason why a segmentation is used for motion saliency is two-folds: (1) The direct usage of moion does not guarantee the performance due to difficulty in the achievement of reliable motion data and 2) the separation of the two regions belonging to an identical object occurs in most of images. The aim of this paper is

to ehhance the spaial sliency utilzing motion as well as affinity model.

An additional problem of motion utilization is the diversity of motion class and the low realiblity of the motion data. Even optical flow methods do not guarantee the satisfactory outcome. In the motion class, there exist a variety of motions such as object motion, camera panning, object tracking and so forth. Since the motion saliency is largely affected by such motions, the good motion saliency cannot be expected unless an accurate motion class is determined. To solve this, our method is designed to be independent of the motion class as well as the motion relibility. Moreover, the usage of an affinity model lessens the difference of neighboring pixel saliencies, yielding the reduction of region discrepancy.

This paper is organized as folows: In the next section, our proposed method will be presnted. Section 3 presents the experimental results. Finally, we summarize our work in Section 4.

2 Proposed Method

Fig. 1 shows the overall block diagram of the proposed method. Given an RGB image, a grayscale image is made. Usually, RGB or Lab color space is used for saliency construction. But, since the purpose of our method is to improve the spatial saliency, only grayscale is used for the clear comparitive performance evaluation. From the grayscle image, a spatial saliency map is obtained. From the current and previous frames, motion data are estimated. Using this motion, motion saliency is obtained. The two saliency maps are fused into a single map that is refined by an affinity model.

One of factors considerd by previous motion saleincy methods is a motion type that is generally composed of object motion (stationary camera and moving object) and camera motion. In the latter, objects might be stationary or be in motions, which is considered to be of no importance in the proposed method

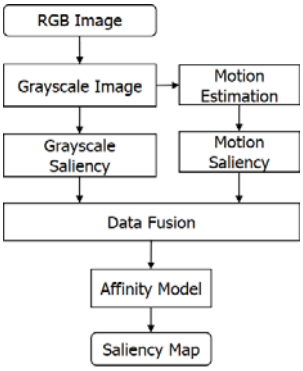


Fig. 1. Flow diagram of the proposed method.

For the spatial saliency map, we adopt Zhai's method [1], which is the global color contrast approach. This method is based on the observation that a biological vision system is sensitive to contrast in visual signal. It defines saliency values for image pixels using color statistics of the input image. Specifically, the saliency of a pixel is defined using its color contrast to all other pixels in the image, i.e., the saliency value of a pixel *i* in image I is defined as

$$S_i = \sum_{j=1}^{N} D(I_i, I_j) = \sum_{j=1}^{N} |I_i - I_j| \qquad (1)$$

where N is the number of pixels in an image I. $D(I_i, I_j)$ is a color or grayscale distance metric between pixels *i* and *j*.

Zhai's method produce saliency maps in Fig. 2. Fig. 2(a) is orginal RGB images and Fig. 2(b) is their salieny maps. Since color or grayscale is a prmary input data, pixels with different colors generate different saleincy values, which is one of disadvantages of the conventinal methods. This problem is clearly observed in most of outputs as shown in Fig. 2(b) except a skier in the third image.

Fig. 2. Saliency map obtained by Zhai's method. (a) input images from [18, 19, 20] and (b) saliency maps.

A) *Grayscale transformation*

Eq. (1) utilizes the color difference between pixels. The larger the difference is, the larger the saliency becomes. The linear equation can be transformed into a

non-linear monotoical increasing function as shown in Fig. 3. The transformation formulae is expressed as follows:

$$S_i = \sum_{j=1}^{N} \left[1 - \left(1 - \frac{|y_i - y_j|}{255} \right)^2 \right] \quad (2)$$

where y_i, y_j are grayscale values of pixel i and j. S_i is normalized to [0, 1].

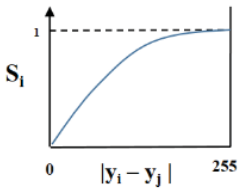


Fig. 3. Saliency transformation using a monotonically-increasing exponential function

B) *Affinity Model*

Unlike the conventional methods, the proximity property of neighboring pixels is utilized in the proposed method in addition to color difference. As mentioned earlier, one of the main drawbacks of the color-based saliency is the appearance of different saliency values between two neighboring pixels belonging to a same object. For exmaple, a part of an object can have a large saliency, but other part is assigned a small saliency. This is observed in most images and results in unnaturalness of visual objects.

An affinity model can soothe this problem, forcing two close-by pixels to have similar values as close as possible. The affinity model is frequenty used in region segmenation, filtering, etc. [14, 15]. In the segmentation, defining an affinity model gained by integrating local grouping cues such as color and boundary is important. The affinity ψ_{ij} between two pixels *i* and *j* is modeled according to the grouping cues. In general, an exponential function is used and the affinity beween two pixels *i* and *j* is expressed by

$$\psi(i,j) = \exp\left(-\frac{\|x_i - x_j\|}{\sigma_a} \right) \quad (3)$$

where x_i, x_j are coordinates of pixel *i* and *j*. σ_a controls the affinity strength.

Then, integrating this to Eq. (2), we have the following refined spatial saliency.

$$S_i = \sum_{j=1}^{N} \psi(i,j) \cdot S_{i,j}$$

$$= \sum_{j=1}^{N} \exp\left(-\frac{\|x_i - x_j\|}{\sigma_a} \right) \cdot \left[1 - \left(1 - \frac{|y_i - y_j|}{255} \right)^2 \right] \quad (4)$$

The effect of the affinity model is depicted in Fig. 4. Zhai's saliency map is shown in Fig. 4(a). The saliency map refined by the affinity model is displayed in Fig. 4(b). The observation indicates that the inner region of the skier has more homogenuous distribution in terms of the saliency. The similar homogenuity can also be observed at the background.

Fig. 4. The effect of affinity model. (a) Zhai's saliency map and (b) saliency map improved by an affinity model.

C) *Motion Support*

Motion vectors are estimated by the comparison of a current frame I_k and a previous frame I_{k-1}. Among diverse motion estimation methods, a block-based motion estimation is adopted for real-time processing. The moton vector $\alpha = (U, V)$ might be derived for a pixel or a block, where *U*, *V* are horizontal and vertical motion components respectively. The motion magnitude is computed by

$$|\alpha| = \sqrt{U^2 + V^2} \quad (5)$$

The utilization of motion data for the saliecy improvement has gained much interests. Most methods implement a segmentation using motion information. After segmentation is done, segments are classified as either foreground or background. Then, a region-based approach is carried out. If the accuracy of the segmentation is high, the results would be reliable. However, most of images contain complex scenes, yielding the achivement of unsatisfactory performance. This results in inconsistency and uncomfortable saliency maps because mis-segmentation can produce the different saliency values in an idental object or a background. Therefore, out proposed method does not employ any segmentation, but seeks the direct utilization of motions.

For relate works, Lie *et al*. derive region information from the motion segments as well color segmenataion based on Gaussian Mixture Model, (GMM) [16]. This result is combined with a spatial saliency. Chang *et al*.

search a moving pixel trajectory and seperate the foreground and background using support vector machine (SVM). Then pixels considered as a camera motion are removed. Other works estimate optical-flow based motion vectors and separate an image into disjoint regions and obtain the saleincy data.

As mentioned earlier, we perform the motion data directly on the image. Further, the complex and unreliable classfication of motion types is not considerd. We make two following assumptons; (1) the motion magnitude of objects is high for the object motion. The saliency is then proportional to the motion magnitude, and (2) in the camera motion, the motion magnitude is strong being inversely propertional to the scene distance from a camera lens. Then, the assumption that the saliency is proportial to the motion magnitude is made.

From the two assumptions, the saliency of pixel *i*, s_i^M is derived from $\| \alpha \|$ of Eq. (5). Unlike the spatial saliency using the color difference, the motion salency can not directly use the motion difference. The reason is because diverse motion magnitude and orientation spread on a motion map. For instance, in the case of a fixed camera and object motion only, the motion diffence might work using a formulae $S_i = \sum_{i=1}^{N} |\alpha_i - \alpha_i|$. On the contrary, this formulae is not appropriate to the camera motion due to diverse motion magnitudes and orientations.

The observation that the saliency is proportional to motion magitude enables to directly apply the magnitude for the saleincy estimation using this formulae

$$S_i^M = \left[1 - \left(1 - \frac{\|\alpha_i\|}{\alpha_{max}}\right)^2\right] \quad (6)$$

where α_{max} is the maximum magnitude of motion vectors that is determined by a search window in the block-based motion estimation.

It is well known that a block-based motion estimation generates unreliable motion vectors. This can be also observed from the optical flow (OF) methods even if the OF is more close to a true motion. Rather that the direct usage of motion vectors, we employ the *motion complexity* representing the degree of motion distribution in the image. The primary purpose is to increase or enhance the spatial saleincy. The motion complexity will be large for a camera motion and relatively small for a object motion. Since the camera motions spread over an entire image, it is not easy to find a solution that can be appied to the motion saliency. Further, the accurate separation of two motion types is also not an easy task as mentioned earlier. On the contrary, motion complexity is proportional to motion distributon. If the motion complexity is low, it is close to the objec motion and otherwise, it can represent a camera motion. Based on the observation, this motion complexity is appropriately multiplied to the motion saliency.

Motion complexity of a motion map is measured by calculating a standard deviation of pixel motion values. The reason behind using the standard deviation for the measurement of motion complexity is that it is the measure of the dispersion or variability of a set of values around the mean of that set. Thus, if the motion map has high motion complexity, the standard deviation of the pixel motion values is expected to be high. Motion vectors take values ranging from 0 to a maximum search range. 0 represents no moton, while the maximum search range corresponds to the highest motion. Given a motion map, the mean pixel motion is computed as follows: Firstly, we obtain a differenc image ΔI^k between current and previous images.

$$\Delta I^k = \left| I^k - I^{k-1} \right| \quad (7)$$

The motion complexity σ^k expressed by the variance of the difference image is computed by

$$\mu^k = \frac{1}{N} \sum_{i=1}^{N} \left| I_i^k - I_i^{k-1} \right| \quad (8)$$

$$\sigma^k = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\Delta I_i^k - \mu^k \right)^2} \quad (9)$$

where mean μ^k is the mean of the difference image.

To control the weight of motion saliency, we compute τ.

$$\tau = \frac{\sigma_{max} - \sigma}{\sigma_{max}} \quad (10)$$

σ_{max} is set to a value that is derived from test images.

The motion complexity is then multiplied to the motion saliency of Eq. (6) as follows :

$$S_i^M = \tau \cdot \left[1 - \left(1 - \frac{\|\alpha_i\|}{\alpha_{max}}\right)^2\right] \quad (11)$$

If σ is large, then τ is decreased and vice versa. For the camera motion with high complexity, τ is small, thereby producing a low weight. On the contrary, object motion's weight increases.

A final saliency incorporating the multiple saliencies is expressed by

$$S_i = \sum_{j=1}^{N}\left(S_{i,j}^{Y} + S_i^{M}\cdot\tau\right)\cdot S_{i,j}^{P}$$
$$= \sum_{j=1}^{N}\left\{\left[1-\left(1-\frac{|y_i-y_j|}{255}\right)^2\right] + \left[1-\left(1-\frac{\|\alpha_i\|}{\alpha_{max}}\right)^2\right]\cdot \left(\frac{\sigma_{max}-\sigma}{\sigma_{max}}\right)\right\}\cdot e^{\frac{\|x_i-x_j\|}{\sigma_a}} \quad (12)$$

The performance evaluation will be presented in the next section.

3 Experimetnal Results

We consider seven different video sequences from three different datasets ([18]–[20]) in our experiments. Four out of the seven videos (*redbird*, *horse*, *ski*, *girl*) are selected from [19]. *birdfall2* is from [18]. *walking* and *street men* are from [20].

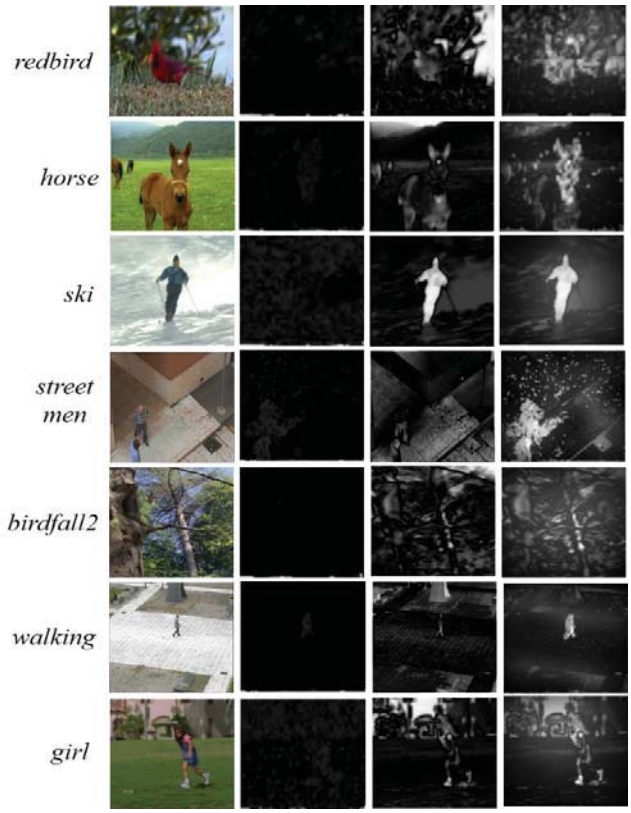


Fig. 5 The results of saliency maps. (a) test images, (b) motion maps, (c) Zhai's saliency maps, (d) saliency maps of the proposed method.

Fig. 5(a) shows one of video frames from each test set and Fig. 5(b) shows motion maps. The search window size is set to 40 in the block-based motion estimation. The saliency maps by Zhai's method are shown in Fig. 5(c) and subsequently ours are shown in Fig. 5(d). Fig. 5 shows that *hen*'s saliency is improved. The top sky region belonging to a background has high saliency, but our method reduces the saliency in that region. Further, the saliency of a foreground object increases. *ski* has similar saliency maps in the two methods. The usage of the affinity model reduces the saliency differences between different pixels with a same object or background. *Street men* clearly displays improved saliency for the pedestrians. In *birdfall2*, a bird is rapidly falling down to the ground (e.g., very high motion). This results in wrong tracking of the bird so that the tree's saliency relatively increases.

Fig. 6 Ground truth data frequently used for performance evaluation.

To quantitatively evaluate the performance, the conventional motion saliency methods heavily depend on ground-truth data provided with the original data (i.e., label information at the pixel level) (see Fig. 6) due to the usage of the segmentation. Then they evaluate the performance based on subjective tests without any objective measurement metrics due to the lack of objective metrics.

We adopt a new objective measurement to evaluate the performance of the proposed method. A pixel is classified as a static or a motion pixel based on motion data. Then, we investigate the increment and decrement of each pixel saliency, where Zhai's method acts as a baseline. The ratio of increased saliency of static and motion pixels is evaluated. If the increment ratio of motion pixels is relatively large compared to that of static pixels, it is expected that the proposed method is satisfactory.

Definitions:

P_M^+ : the ratio of motion pixels with saliency increment

δ_M^+ : the average of saliency increment of motion pixels

P_M^- : the ratio of motion pixels with saliency decrement

δ_M^- : the average of saliency decrement of motion pixels

P_S^+ : the ratio of static pixels with saliency increment

δ_S^+ : the average of saliency increment of static pixels

P_S^- : the ratio of static pixels with saliency decrement

δ_S^- : the average of saliency decrement of static pixels

Eight measurements defined in *Definition* are computed as follows:

$$\mathrm{P_M^+} = 100 \times \frac{\mathrm{N_M^+}}{N_M} \quad \text{and} \quad \mathrm{P_M^-} = 100 \times \frac{\mathrm{N_M^-}}{N_M}$$

$$\mathrm{P_S^+} = 100 \times \frac{\mathrm{N_S^+}}{N_S} \quad \text{and} \quad \mathrm{P_S^-} = 100 \times \frac{\mathrm{N_S^-}}{N_S}$$

$$\delta_{\mathrm{M}}^{+} = \frac{1}{\mathrm{N_M}} \sum_{\mathrm{i=1,i\in I_M}}^{\mathrm{N}} [\mathrm{MAX}(\mathrm{S_{PS}(i)} - \mathrm{S_{ZS}(i)}, 0)]$$

$$\delta_{\mathrm{M}}^{-} = \frac{1}{\mathrm{N_M}} \sum_{\mathrm{i=1,i\in I_M}}^{\mathrm{N}} [\mathrm{MIN}(\mathrm{S_{PS}(i)} - \mathrm{S_{ZS}(i)}, 0)]$$

$$\delta_{\mathrm{S}}^{+} = \frac{1}{\mathrm{N_S}} \sum_{\mathrm{i=1,i\in I_S}}^{\mathrm{N}} [\mathrm{MAX}(\mathrm{S_{PS}(i)} - \mathrm{S_{ZS}(i)}, 0)]$$

$$\delta_{\mathrm{S}}^{-} = \frac{1}{\mathrm{N_S}} \sum_{\mathrm{i=1,i\in I_S}}^{\mathrm{N}} [\mathrm{MIN}(\mathrm{S_{PS}(i)} - \mathrm{S_{ZS}(i)}, 0)]$$

where N_M is the number of motion pixels and N_S is the number of static pixels. N_M^+ , N_M^- are the number of motion pixels with increment and decrement, respectively. N_S^+ and N_S^- are the number of static pixels with increment and decrement, respectively. S_{ps} and S_{zs} are saliency maps generated by the proposed method and Zhai's method.

Statistical values obtained by each test sequence is presented in Table 2 with additional data information in Table 1. The baseline is Zhai's result and then we examine the performance improvement of the proposed method over the baseline. Here, we expect that motion pixels' saliency is incremented and static pixels' saliency is decremented.

As shown in Table 2, *hen* has P_M^+ of 61%, P_M^- of 18%, P_S^+ of 64%, and P_S^- of 15%. On average, P_M^+ , P_M^- , P_S^+ , P_S^- are 73%, 10%, 61%, and 21%. More important measurement is the amount of increment and decrement values. For *hen*, *ski*, *horse*, *street men*, *birdfall2*, *walking*, *girl*, ($\delta_M^+ - \delta_S^+$) are +23, +13, +30, +54, +63, +60. The relative increment of motion pixels is larger than that of static pixels. ($\delta_M^- - \delta_S^-$) are –20, +2, +3, +8, +23, -20. On average, ($\delta_M^+ - \delta_S^+$) is +41 and ($\delta_M^- - \delta_S^-$) is -1. According to this outcome, it is validated that our method utilizing motion information improves the spatial saliency.

Table 1. Data of test sequences

Test Sequence	Image resolution	Motion type	No. of frames
Redbird	384x212	Object motion	42
Ski	352x288	Camera motion	66
Horse	360x288	Camera motion	71
Street men	704x576	Object motion	22
Birdfall2	260x320	Object motion	30
Walking	704x576	Object motion	40
Girl	400x320	Camera motion	22

Table 2. Objective performance comparison of the proposed method with Zhai's method

Test Sequence	Motion pixel		Static pixel	
	$\mathrm{P_M^+}$ (δ_{M}^{+})	$\mathrm{P_M^-}$ (δ_{M}^{-})	$\mathrm{P_S^+}$ (δ_{S}^{+})	$\mathrm{P_S^-}$ (δ_{S}^{-})
Redbird	61 (+55)	18 (-61)	64 (+21)	15 (-41)
Ski	80 (+33)	2 (-7)	62 (+20)	20 (-9)
Horse	71 (+48)	8 (-37)	72 (+18)	7 (-40)
Street men	71 (+73)	8 (-26)	41 (+19)	38 (-34)
Birdfall2	79 (+87)	4 (-3)	63 (+24)	18 (-26)
Walking	79 (+81)	21 (-44)	67 (+21)	32 (-22)
Girl	66 (+39)	7 (-30)	70 (+18)	4 (-42)
Average	72 (58)	10 (-29)	62 (20)	18 (-30)

4 Conclusion

In this paper, we proposed an enhancement method of spatial saliency which utilizes an affinity model and motion data. The research motivation underlying the proposed method is not to use any complex region segmentation, which poses a difficulty in achieving correct outcome. We used the affinity model to make the saliency values of close-by pixels own similar values. Motion complexity acts as a tool that is free of image motion classification. Compared with Zhai's spatial saliency method, our approach improves the original saliency. A major advantage of our proposed method is that we do not require any object segmentations that are

used for the prior knowledge of the objects of interest. Experiments performed on a variety of videos with object or/and camera motions verified the effectiveness and robustness of our proposed method. A final contribution is the presentation of objective performance evaluation tool, which has been ignored in conventional saliency evaluation.

5 Acknowledgement

This research was supported by Kangwon National University UICF.

6 References

[1] Y. Zhai, and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," Proceedings of the 14th annual ACM Int' Conf. on Multimedia, pp. 815-824, 2006.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[3] Y. Xia, R. Hu, Z. Huang and Y. Su, "A novel method for generation of motion saliency," Proc. of IEEE 17th Int' Conf. on Image Processing, Sep. 2010.

[4] Y. Zhu, N. Jacobson, H. Pan, and T. Nguyen, "Motion-decision based spatiotemporal saliency for video sequences," IEEE Int' Conf. on Acoustics, Speech and Signal Processing, pp. 1333-1336, 2011.

[5] X. Yang, R. Hu, Z. Huang and Y. Su, "A novel method for generation of motion saliency," IEEE Int' Conf. on Image Processing, pp. 4685-4688, Sep. 2010.

[6] R. Pan, X. Lin, C. Huang and L. Wang, "A novel vehicle flow detection algorithm based on motion saliency for traffic surveillance system," IEEE 9th Int' Conf. on Computational Intelligence and Security, 2013.

[7] D. Liu and M. Shyu, "Semantic retrieval for videos in non-static background using motion saliency and global features," IEEE 7th Int' Conf. on Semantic Computing, 2013.

[8] C. Huang, Y. Chang, Z. Yang and Y. Lin, "Video saliency map detection by dominant camera motion removal," IEEE TSVT, 2012.

[9] A. Rahman, D. Houzet, D. Pellerin and L. Agud, "GPU implementation of motion estimation for visual saliency," IEEE Int' Conf. on Design and Architectures for Signal and Image Processing, 2010.

[10] A. Hiratani, r. Nakashima, K. Matsumiya, I. Kuriki, and S. Shitori, "Considerations of self-motion in motion saliency," 2nd IAPR Asian Conf. on Pattern Recognition, 2013.

[11] R. Achanta, A. Shaji, and K. Smith, A. Lucchi, P. Fua and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans on. Pattern Analysis and Machine Intelligence, 34(11), 2274-2282, 2012.

[12] A. Hiratani, R. Nakashima, K. Matsumiya, I. Kuriki, and S. Shitori, "Considerations of self-motion in motion saliency," 2nd IAPR Asian Conf. on Pattern Recogniotion, 2013.

[13] J. Li, Y. Tian, T. Huang, and W. Gao "A dataset and evaluation methodology for visual saliency in video," IEEE Int' Conf. on Multimedia and Expo, pp. 442-445, June 2009.

[14] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.

[15] Y. Boykov and G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," Int J. Computer Vision, vol. 70, no. 2, pp. 109-131, 2006.

[16] W. Li, H. Chang, K. Lien, H. Chang, and Y. F. Wang. "Exploring Visual and Motion Saliency for Automatic Video Object Extraction,"IEEE Trans. on Image Processing, Vol. 22, No. 7, July 2013.

[17] C. Huang, Y. Chang, Z. Yang, and Y. Lin, "Video Saliency Map Detection by Dominant Camera Motion Removal," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 24, Issue 8, pp. 1336-1349, 2014.

[18] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in Proc. Brit. Mach. Vis. Conf., 2010.

[19] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in Proc. IEEE Int. Conf. Multimedia Expo, pp. 638–641, June–July, 2009.

[20] D. Baltieri, R. Vezzani and R. Cucchiara, "3DPes: 3D People Dataset for Surveillance and Forensics," in Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects, Scottsdale, Arizona, USA, pp. 59-64, Nov-Dec, 2011. (http://imagelab.ing.unimore.it/visor/3dpes.asp)

A Hybrid Indoor Localization System Based on Infra-red Imaging and Odometry*

Suat Karakaya, Hasan Ocak, Gürkan Küçükyıldız and Orkun Kılınç
Mechatronics Engineering Department, Kocaeli University, Kocaeli, Turkey

Abstract - *In this study, a real-time indoor localization system was developed by using a camera and passive landmarks. A narrow band-pass infra-red (IR) filter was inserted to the back of the camera lens for capturing IR images. The passive landmarks were placed on the ceiling at pre-determined locations and consist of IR retro-reflective tags that have binary coded unique ID's. An IR projector emits IR rays at the tags on the ceiling. The tags then reflect the rays back to the camera sensor creating a digital image. An image processing algorithm was developed to detect and decode the landmarks in captured images. The proposed algorithm successfully estimates the position and the orientation angle based on relative position and orientation with respect to the detected tags. To further improve the accuracy of the estimates, extended Kalman filter (EKF) was adapted to the measurement algorithm. The proposed method initially estimates the position of a mobile robot based on odometry and kinematic model. EKF was then used to update the estimates given the measurement obtained from the image processing system. Real time experiments were performed to test the performance of the system. The results prove that the proposed indoor localization system can effectively estimate position with an error less than 5cm.*

Keywords: Video and image processing, Extended Kalman filter, Indoor localization, Differential drive

1 Introduction

Mobile robot motion planning is one of the most attractive studies for researchers. Motion planning includes sub-tasks such as path planning, localization and path traversal. Each sub task has various challenges and proposed solutions. Localization problem can be assumed as two types which are indoor and outdoor localization. Outdoor localization systems usually utilize GPS data, IMU and encoder units. Requirement of accuracy and changing conditions (e.g. walls, moving obstacles, humans, fragile objects etc.) can explain the diversity in indoor localization studies [6-15]. Hazas and his team have studied on ultrasonic indoor localization system called as active bat system [1]. They have mounted ultrasonic receivers on the ceiling to measure the distance between active ultrasonic tags. Each tag transmits ultrasonic pulse and the receiver measures the distance based on time of flight of the received data. They claimed to achieve 2 cm accuracy by the proposed technique. However, the study has challenges such as requirement of large number of very accurately mounted ultrasonic receivers. Biswas and his team has utilized depth image of a mobile robot's environment to localize the mobile robot [2]. They used Kinect sensor as depth camera and proposed fast sampling plane filtering (FSPF) to reduce the computational cost while processing the multi-dimensional measurement data in order to eliminate undesired time delays in real time. Laser imaging detection and ranging system (LIDAR) has also been used for indoor localization [3]. Distance and angular measurements can be obtained from a LIDAR in various ranges. LIDAR sends laser beam to its environment and gives a distance measurement of the surface from which the laser beam is reflected. Highly accurate measurements can be obtained with LIDAR sensors at the cost of high price. Radio frequency identification (RFID) tags were used for indoor localization in [4]. Wireless sensors based localization can also be applied in indoor environments [5]. A multi-sensor system measures received signal strength and estimates the location of moving objects. For determination of location, signal strengths are compared with a reference value kept in a database. However, such systems do not provide satisfactory performance for applications where highly accurate positioning is required. Only a few meters of accuracy can be reached by this method except for specific applications. In this study, we developed an image processing based indoor localization system and combined it with odometry by using extended Kalman filter (EKF). While the mobile robot was moving on a random path, exact measurements were taken from a highly-accurate laser point measurement sensor. Meanwhile, the position of the mobile robot was continuously measured by developed localization system and odometry. EKF based position estimates was also recorded in real time. By making comparison between estimation and exact location, it was seen that EKF estimates were more accurate than both odometry and the developed indoor localization system. A considerable improvement in indoor localization accuracy was achieved through EKF.

*This study was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under the grant TUBITAK-113E777.

2 IR imaging based indoor position and orientation measurement system

2.1 Tag detection

Undesired light effects, shadows, over and under lightening and environmental disturbances pose challenges in image processing applications. To achieve robustness against such effects, we preferred studying in infra-red wave length. The proposed image processing system consists of a USB camera with a narrow band-pass filter inserted lens, an IR projector and passive IR retro-reflective tags. Each tag has a binary coded unique ID that was designed in a 3x4 size matrix format seen in Figure 1. Each element of the tag matrix contains logical 1 or 0. The corner elements of the tag matrix are reserved to fixed values for orientation detection and the remaining elements encode the tag by constructing a specific binary number. The description of the notation shown in Figure 1 is given Table 1.

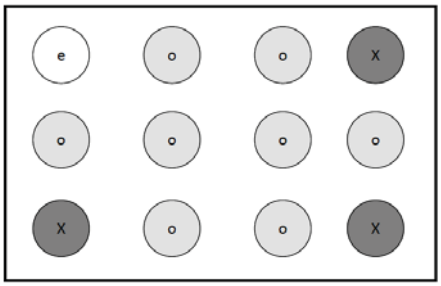


Figure 1. Tag design

Table 1. Description of the tag matrix notation

Notation	Description
e	Always will be empty and not permitted to place on a reflective circle (0)
X	Always will be placed on a reflective circle (1)
o	Can be empty or occupied (0 or 1)

The elements of the tag matrix which were represented with logical 1 corresponds a circular IR reflective sticker with 25 mm diameter. The occupied regions on the rectangular tag reflect the IR rays emitted by the projector which makes them brighter in captured images while the rest of the regions on the tag and the other pixels seem to be darker. Therefore, all pixels except for the reflective circles are filtered out by thresholding.

In the image processing step, various eliminations were applied to input frames for deciding whether a white pixel cloud is a valid tag or not. The input frame was converted to black and white (BW) image and connected component analysis algorithm was applied on the binary image. The centroid of each component is computed. As more than one tag can be in the camera's field of view, a clustering algorithm was applied to the component centroids. Each centroid was assigned to a cluster and a minimum bounding rectangle (minBR) was found for each cluster. The clusters were filtered based on median size, collinearity of the occupied regions' center points, touching of the minBR to edges of the image, tag design (check the reserved regions on the cluster), number of elements (a cluster can contain maximum 11 elements), diagonal length, ratio of long and short edges of the minBR. Each elimination step checks whether a given cluster satisfies a specific property of the designed tags or not.

The elimination steps are as follows: The absolute difference between a component in a cluster and the median area of the components in that cluster should not be more than a certain threshold in pixels. The minBR of a cluster should not touch the image edges. The three corners of the minBR should correspond to white pixels in the BW image based on the tag design depicted in Figure 1. The ratio of the long and the short edges of the tag should be consistent with the tag design. The clusters that pass all the elimination steps are assumed to be candidate tags. The candidate tags were decoded and the corresponding tag matrices were computed. Finally, the number of 1's in the tag matrix was required to be the same as the number of components in the corresponding cluster. Once the final requirement is satisfied, the decoded tag is checked if registered or not. Unregistered tags were discarded even if they satisfy all criteria.

It is possible for the camera to see multiple tags at a time. Localization accuracy improves with increasing number of detected tags. Tags were placed on the ceiling of the mobile robot work space. The real world coordinates of all inserted tags were recorded to a database. Each correctly-detected tag gives location information for the mobile robot. Location measurements gathered from multiple tags were clustered based on the measurements. The final measurement was computed as the measurement averages of the tags in the cluster with the maximum number of elements.

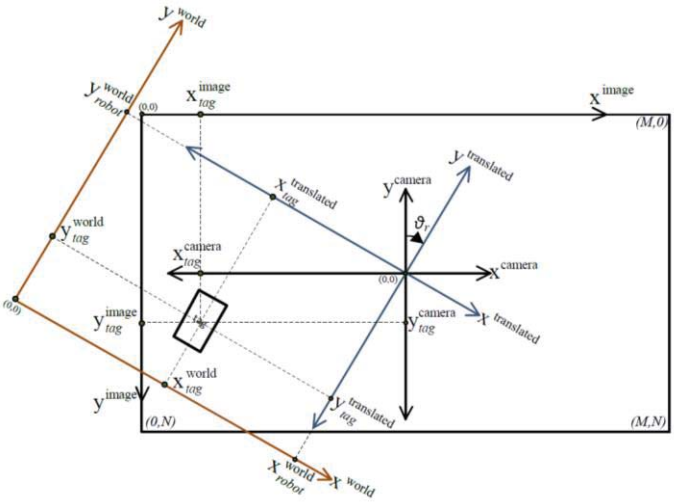


Figure 2. Image, camera, real world and translated real world coordinate

2.2 Position estimation for a detected tag

Let x_{tag}^{image} and y_{tag}^{image} be the image coordinates of the tag center as illustrated in Figure 2. The tag center coordinates with respect to the camera coordinate system (x_{tag}^{camera},

y_{tag}^{camera}) were calculated by translating the image coordinates to the center of the frame, rotating 90° in counter-clock-wise and reversing the x-axis. Then, the coordinates of the mobile robot with respect to the world coordinate system translated to the origin of the camera coordinate system can be computed as;

$$\begin{bmatrix} x_{\text{tag}}^{\text{translated}} \\ y_{\text{tag}}^{\text{translated}} \end{bmatrix} = \mathbf{R}_\theta \begin{bmatrix} -(N - x_{\text{tag}}^{\text{image}}) \times c_{\text{x}} \\ (M - y_{\text{tag}}^{\text{image}}) \times c_{\text{y}} \end{bmatrix} \quad (1)$$

Camera intrinsic parameters c_x and c_y were calculated through a calibration process. Tags were placed on the ceiling with a 0° orientation. Therefore, orientation angle of the mobile robot (θ) was assumed to be the orientation angle of the minBR for the tag. N and M are the number of rows and columns in the image and $\mathbf{R}_\theta$ is the rotation matrix around z-axis. Finally, coordinates of the mobile robot with respect to world coordinate system is calculated as;

$$\begin{bmatrix} x_{\text{world}}^{\text{robot}} \\ y_{\text{world}}^{\text{robot}} \end{bmatrix} = \begin{bmatrix} x_{\text{tag}}^{\text{world}} \\ y_{\text{tag}}^{\text{world}} \end{bmatrix} - \begin{bmatrix} x_{\text{tag}}^{\text{translated}} \\ y_{\text{tag}}^{\text{translated}} \end{bmatrix} \quad (2)$$

3 Kinematic model and EKF equations

Differential drive kinematic model of a mobile robot consists of two drive wheels and a common axis. The motion of the mobile robot is supplied by varying the velocities of each independent wheel. The robot rotates around a point which lies on the robot's wheel axis. The kinematic model of the mobile robot is illustrated in Figure 3.

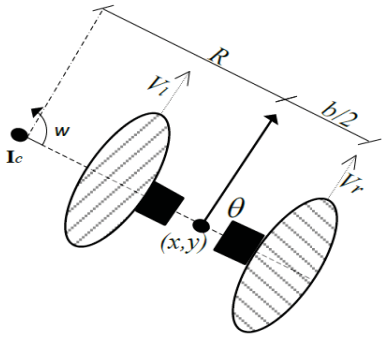


Figure 3. Differential drive kinematics

Let v_r, v_l, b, w, θ, R, $\mathbf{I}_c$, x, y be the right and left wheel velocities, wheel base, angular velocity, orientation angle, rotation radius, center point of rotating motion and current location of the mobile robot, respectively.

Rotation radius (R), angular velocity (ω) and center point of rotating motion ($\mathbf{I}_c$) are given by,

$$R = \frac{b}{2}\frac{v_r + v_l}{v_r - v_l}, \quad \omega = \frac{v_r - v_l}{b}, \quad \mathbf{I}c = [x - R\sin\theta, y + R\cos\theta] \quad (3)$$

Distances travelled by right and left wheels (s_r and s_l), change in orientation angle (∂_θ) of the robot during a small time interval d_t are then given by,

$$s_l = V_l d_t, \; s_r = V_r d_t, \; R = \frac{b}{2}\frac{s_r + s_l}{s_r - s_l}, \; \partial_\theta = w\partial_t \quad (4)$$

The state-space model based on the wheels' velocities can be written as,

$$\begin{bmatrix} x_{t+\partial t} \\ y_{t+\partial t} \\ \theta_{t+\partial t} \end{bmatrix} = \begin{bmatrix} \cos(\omega\partial t) & -\sin(\omega\partial t) & 0 \\ \sin(\omega\partial t) & \cos(\omega\partial t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x - Ic_x \\ y - Ic_y \\ \theta \end{bmatrix} + \begin{bmatrix} Ic_x \\ Ic_y \\ \omega\partial t \end{bmatrix} \quad (5)$$

The system vector can be obtained by using Equations 3-5 as,

$$\begin{bmatrix} x_{t+\partial t} \\ y_{t+\partial t} \\ \theta_{t+\partial t} \end{bmatrix} = \mathbf{f}(x_t, y_t, \theta_t, s_r, s_l) = \begin{bmatrix} x_t \\ y_t \\ \theta_t \end{bmatrix} + \begin{bmatrix} \frac{b}{2}\frac{(s_r + s_l)}{(s_r - s_l)}(\sin(\theta + \partial_\theta) - \sin(\theta)) \\ \frac{b}{2}\frac{(s_r + s_l)}{(s_r - s_l)}(-\cos(\theta + \partial_\theta) + \cos(\theta)) \\ \partial_\theta \end{bmatrix} \quad (6)$$

The state transition and the input matrices, **A** and **B**, for the linearized model can be calculated by taking Jacobian of the system vector **f**,

$$\mathbf{A} = \begin{bmatrix} \frac{\delta f_x}{\delta x} & \frac{\delta f_x}{\delta y} & \frac{\delta f_x}{\delta \theta} \\ \frac{\delta f_y}{\delta x} & \frac{\delta f_y}{\delta y} & \frac{\delta f_y}{\delta \theta} \\ \frac{\delta f_\theta}{\delta x} & \frac{\delta f_\theta}{\delta y} & \frac{\delta f_\theta}{\delta \theta} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \frac{(s_r + s_l)}{(s_r - s_l)}(\cos(\theta) - \cos(\theta + \partial_\theta)) \\ 0 & 1 & \frac{(s_r + s_l)}{(s_r - s_l)}(\sin(\theta) - \sin(\theta + \partial_\theta)) \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

$$\mathbf{B} = \begin{bmatrix} \frac{\delta f_x}{\delta s_r} & \frac{\delta f_x}{\delta s_l} \\ \frac{\delta f_y}{\delta s_r} & \frac{\delta f_y}{\delta s_r} \\ \frac{\delta f_\theta}{\delta s_r} & \frac{\delta f_\theta}{\delta s_l} \end{bmatrix}$$

Assuming that all three states (positions x, y, and the orientation θ) can be measured, then the output matrix **C** is a 3x3 identity matrix. Extended Kalman equations are given in Table 2.

Table 2. Extended Kalman equations

Prior estimate:	Posterior estimate:
	$K_k = \mathbf{P}_k^- \mathbf{C}^T (\mathbf{C}\mathbf{P}_k^- \mathbf{C}^T + \mathbf{R})^{-1}$
$\mathbf{x}_k^- = f(\mathbf{x}_{k-1}, s_r, s_l)$	$\hat{\mathbf{x}}_k = \mathbf{x}_k^- + K_k(\mathbf{y}_k - \mathbf{C}\mathbf{x}_k^-)$
$\mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^T + \mathbf{Q}$	$\mathbf{P}_k = (\mathbf{I} - K_k\mathbf{C})\mathbf{P}_k^-$

We used diagonal process and measurement error covariance matrices (**R** and **Q**) assuming the errors for each state are uncorrelated. The error covariance matrices are given as;

$$\mathbf{Q}=\begin{bmatrix} w_x^2 & 0 & 0 \\ 0 & w_y^2 & 0 \\ 0 & 0 & w_\theta^2 \end{bmatrix},\ \mathbf{R}=\begin{bmatrix} v_x^2 & 0 & 0 \\ 0 & v_y^2 & 0 \\ 0 & 0 & v_\theta^2 \end{bmatrix} \tag{8}$$

, where w_x, w_y, w_θ are the measurement error variances and v_x, v_y, v_θ are the process error variances for x, y and θ, respectively.

4 Experimental results and conclusion

In this study, a mobile robot platform was developed as a test platform for the proposed scheme. The binary coded tags were placed on the ceiling and a camera was inserted to top side of the mobile robot facing towards the ceiling. The robot was steered with a radio transmitter on a random path and was stopped periodically to measure its actual position by a highly-accurate laser distance sensor. Since an external real-time absolute localization system was not available, measurements were manually taken at certain intervals for comparison. In EKF Equation 8, the measurement and process error variances were taken as;

$$w_x = w_y = 2\ [\frac{cm}{s}],\ w_\theta = 1\ [\frac{\deg}{s}]$$
$$v_x = v_y = 10\ [\frac{cm}{s}],\ v_\theta = 1[\frac{\deg}{s}] \tag{9}$$

During the movement of the robot, position measurements, odometry readings and EKF based position estimates were continuously recorded. Measurements performed by the image processing system and the EKF estimates were compared along with the corresponding actual position of the mobile robot. Comparative experimental results are illustrated in Figure 4.

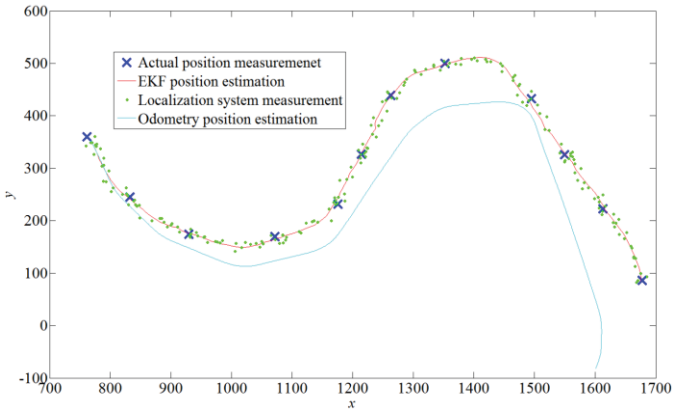


Figure 4. The position estimates of the mobile robot

In Figure 4, the green dots represent the measurements performed by the image processing system, the red line correspond to EKF based position estimates and the cyan line shows the robot position computed based on the odometry alone. The blue crosses depict the actual position of the mobile robot measured at certain intervals with a highly accurate laser distance sensor. The measurement and the estimate errors for the image processing system and the EKF-based localization system are given in Table 3.

Table 3. The measurement errors

Actual Position (cm)		EKF Estimate (cm)		Measurement System (cm)		Estimation Error (cm)		Measurement Error (cm)	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
762	360	759	356	767	354	-3	-4	5	-6
833	245	835	241	825	240	2	-4	-8	-5
930	174	932	175	937	179	2	1	7	5
1072	170	1075	167	1079	165	3	-3	7	-5
1176	232	1173	235	1172	237	-3	3	-4	5
1215	327	1216	331	1210	334	1	4	-5	7
1263	439	1265	440	1268	443	2	1	5	4
1353	500	1350	498	1357	496	-3	-2	4	-4
1495	433	1492	430	1490	429	-3	-3	-5	-4
1550	326	1554	322	1554	332	4	-4	6	4
1613	223	1617	219	1617	215	4	-4	4	-8
1678	86	1675	83	1687	93	-3	-3	9	8

It can be observed from Figure 4 that the robot position computed based on odometry alone is highly inaccurate because odometry error accumulates over time. Accumulative error present in odometry is mostly caused by wheel slippage, surface roughness, measurement error in rotation and distance between wheels, and non-continuous sampling of wheel increments. The image processing system measures the position with an error under 10 cm. On the other hand, position error for the EKF based estimates are under 5 cm. Experimental results prove that the proposed scheme a is very effective in indoor localization system, and can estimate the position with a very low error that should be satisfactory for most indoor applications.

5 References

[1] Hazas, M., Hopper, A., "A Novel Broadband Ultrasonic Location System for Improved Indoor Positioning", IEEE Transactions on Mobile Computing, 2006, 5, 5, 536 - 547

[2] Biswas, J., Veloso, M., "Depth Camera Based Indoor Mobile Robot Localization And Navigation", Proceedings of IEEE International Conference on Robotics and Automation, 2012, 1697-1702.

[3] Rongbing L., Jianye L., Zhang, L., Hang Y., "LIDAR/MEMS IMU Integrated Navigation (SLAM) Method For A Small UAV In Indoor Environments", Proceedings of IEEE Inertial Sensors and Systems Symposium (ISS), 2014 , 1-15

[4] Mautz, R., "Overview of Current Indoor Positioning Systems", Journal of Geodesy and Cartography, 2009, 35, 1, 18-22

[5] Youngsu P., Je Won, L., SangWoo K., "Improving Position Estimation on RFID Tag Floor Localization Using RFID Reader Transmission Power Control", IEEE International Conference on Robotics and Biomimetic (ROBIO) 2008, 1716-1721

[6] Surrécio, A., Nunes, U., Araújo R., "Fusion of Odometry with Magnetic Sensors Using Kalman Filters and Augmented System Models for Mobile Robot Navigation", IEEE International Symposium on Industrial Electronics, 2005, 1551 – 1556

[7] Chen, S.Y., "Kalman Filter for Robot Vision: A Survey", IEEE Transactions on Industrial Electronics, 59, 11, 4409 - 4420

[8] Kharidia, S.A., Qiang, Y., Sampalli, S., Cheng, J., Hongwei, D., Wang, L., "HILL: A Hybrid Indoor Localization Scheme", 10th International Conference on Mobile Ad-hoc and Sensor Networks (MSN), 2014, 201 – 206

[9] Chuenurajit, T., Phimmasean, S., Cherntanomwong, P., "Robustness Of 3D Indoor Localization Based on Fingerprint Technique in Wireless Sensor Networks", 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Tech-nology (ECTI-CON), 2013, 1 – 6

[10] Torok, A., Nagy, A., Kovats, L., Pach, P., "DREAR - Towards Infrastructure-Free Indoor Localization via Dead-Reckoning Enhanced With Activity Recognition", Eighth International Conference on Next Generation Mobile Apps, Services and Technologies (NGMAST), 2014, 106 – 111

[11] Lemic, F., Busch, J., Chwalisz, M., Handziski, V., Wolisz, A., "Infrastructure for Bench-Marking RF-Based Indoor Localization under Controlled Interference", Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2014, 26 – 35

[12] Li, R., Fang, Z., Hao, B., Yang, F., "Research on Indoor Wireless Localization System for Radioactive Sources Based On ZigBee", International Conference on Computing, Control and Industrial Engineering (CCIE), 2010, 359 – 362

[13] Chen, Y., Lymberopoulos, D., Liu, J., Priyantha, B., "Indoor Localization Using FM Signals", IEEE Transactions on Mobile Computing, 2013, 12, 8, 1536-1233

[14] Kim H., S., "Advanced Indoor Localization Using Ultrasonic Sensor and Digital Compass", International Conference on Control, Automation and Systems, 2008, 223 – 226

[15] Daum, F., "Nonlinear Filters: Beyond the Kalman Filter", IEEE Aerospace and Electronic Systems Magazine, 2005, 20, 8, 57-69

SESSION

PATTERN RECOGNITION AND FEATURE DETECTION, EXTRACTION AND CLASSIFICATION

Chair(s)

TBA

Evaluation of Synthetically Generated Airborne Image Datasets using Feature Detectors as Performance Metric

Georg Hummel[1], **and Peter Stütz**[1]
[1]Institute of Flight Systems, University of the Bundeswehr, Munich, Germany

Abstract - *The use of synthetic datasets to develop, prototype and qualify new computer vision algorithms is currently not widely accepted, though highly sought after by the industry. This is due to lack of knowledge on how the results acquired with such datasets will transfer to real live performance. Therefore, this paper introduces an approach to evaluate modelled synthetic datasets against their real counterparts. In a use case, the performance of common feature detectors is evaluated using the repeatability metric against real and synthetic datasets. Based on resulting performances; general usability, rendering techniques and modelling efforts for generation of synthetic datasets are discussed.*

Keywords: synthetic datasets; dataset evaluation; feature detectors; homography; performance analysis;

1 Introduction

Today the development of new CV-algorithms often depends on the quality of design, training or test datasets. However, when it comes to applications striving to process data from aircraft mounted sensors, public availability of datasets is rare and available data are homogeneous or fragmented. Therefore, the resulting algorithms are often limited to the operational conditions available in the used datasets to perform as intended. Datasets such as VIVID[1] or NGSIM[2] are providing good means for prototyping of specific algorithm but cannot cover the complexity of weather and lighting influences in aerial imagery due to their recording at one specific date and location.

In 1995 [3] already discussed the concept of using a synthetic environment to develop CV-algorithms. In the last 20 years computer graphic technologies experienced a technology leap allowing to model weather conditions, illumination or shadowing in photo-realistic qualities. In [4] an airborne object algorithm designed on real datasets was evaluated on its performance with synthetic data. It has become a common procedure to use abstract synthetic datasets for initial development of new computer vision algorithms [5] followed by further steps using real datasets. Several image processing benchmarks [6]–[8] use synthetic data due to the easy to access ground truth for quantitative performance measurements. Still, in the final stage, the computer vision domain seeks to extract information from real (recorded) imagery, which is much more complex than its synthetic representations. Thus the acceptance using synthetic data for evaluation of algorithms is low, since while the content of the scene can be the same, the image structure may be fundamentally different (e.g. texture, color, contrast, etc.) [9].

This paper details the CV-algorithm evaluation step suggested in the concept presented in [10] using geo-referenced airborne image datasets of real and synthetic nature. Therefore, the general concept is briefly introduced in the following section.

2 General concept

The general evaluation concept is intended to allow investigation of essential image properties and influencing rendering technologies and to identify a trade-off between modelling detail and algorithm performance. It further aims to provide suggestions and design guidelines towards a benchmark simulation system. This shall be achieved by evaluating basic CV-algorithms against datasets consisting of sequential images by varying rendering techniques and to compare their performance. Thus, we can derive conclusions on the suitability of conducted modelling and rendering efforts of synthetic datasets.

The multi-level concept consists of four different levels as depicted in Fig. 1. The bottom layer (layer 1) contains the datasets and their corresponding ground truth. These datasets consist of aerial imagery and aircraft telemetry derived from test flights performed in either the real ("real datasets") or the synthetic environment ("synthetic datasets"). The ground truth contains the camera movement between compared images as geometric transformation. Level two analyzes the image structure of evaluated datasets using image descriptors (MPEG7) usually deployed for image queries to search engines or image databases. This mechanism is explained in detail in [10]. It allows the direct comparison of image properties. Level three uses computer vision algorithms as test

Table 1: Aircraft parameters provided by telemetry

Recorded Telemetry			
Parameter	***Accuracy***	***Unit***	***Update Frequency***
Latitude (WGS84)	0.06"	degree	5 Hz
Longitude (WGS84)	0.06"	degree	5 Hz
Altitude (AGL)	0.1	meter	100 Hz
Yaw (Euler)	1°	degree	100 Hz
Pitch (Euler)	1°	degree	100 Hz
Bank (Euler)	1°	degree	100 Hz

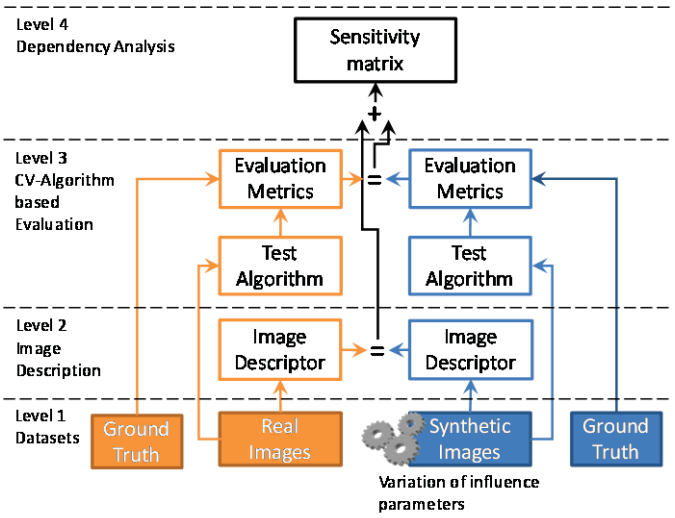


Fig. 1. Evaluation concept of datasets against known computer vision algorithms. This concept is part of the more general concept presented in [10].

algorithm to extract the performance differences among datasets. For clarity, only widely used metrics measuring the quality of algorithms are selected (time based metrics are not considered). The last level performs dependency analysis using the results from level two and three, which allows correlation and weighing between resulting performance and identified image properties. Thus, conclusions in level four shall allow to identify rendering techniques suitable for computer vision algorithm testing and evaluation. This paper, discusses level three (CV-algorithm based evaluation) and level one (dataset generation).

3 Dataset generation

Special interest in this work has been laid on the generation of datasets. First, we had to ensure the scenic correlation between synthetic and real datasets. Therefore, the test flight area was modelled in a virtual environment to create snapshots with identical scenery. Secondly, we had to record telemetry data representing the sensors pose and location (e.g. location, attitude and altitude of the aircraft at which images are taken in flight). This enables us to position the camera in the virtual environment equivalently. The following subsection explains the dataset in detail.

3.1 Test flight dataset "Real"

The taxiway of a former airport on the premises of the University of the Bundeswehr Munich was selected as test site, because it was easy to access, free from unauthorized persons, allows small aircraft operation and had changing terrain (e.g. field, woods, buildings, etc.). As sensor platform, a Multicopter equipped with eight 350W motors and 13" propellers was selected due to its payload, in air stability and low vibrations. This platform has a maximum take-off weight (MTOW) of 6kg allowing 2.2kg payload at max. The aircraft can be navigated via waypoints at a fixed above ground altitude. The camera has been mounted perpendicular to the aircraft frame using a fixed rigid non-stabilized mount. The deployed camera, a XIMEA MQ042CG-CM has been configured to a resolution of 2048x2048 at 30 Hz. A detachable C-Mount lens from Myutron, achieving a total field of view of 25.4° has been deployed. The telemetry was received directly from the flight control system via serial interface at 100 Hz and containing several aircraft parameters detailed in Table 1. Telemetry and image data were recorded on-board in sync using Linux based distributed data services [11], running on a Commel LS-37B Single-board computer.

The actual test flight was conducted on March 18, 2015 at noon on sunny weather leading to crisp shadows and some reflections on buildings. The altitude has been fixed to 75 meters above ground. During the flight, 1000 meter of terrain have been covered that were categorized in nine classes of which three are presented in this paper later on. Each category was reduced to 11 sequentially taken images at 1 Hz to reduce data while retaining sufficient overlap. The images have been resized and cropped to 1024x768 pixels for comparison with synthetic datasets. Due to automatic white balancing the images of the real dataset had a slightly green tint. In future tests manual color calibration may minimize this effect.

3.2 Synthetic dataset

At first, a virtual environment (engine) suitable for geo-referenced dataset generation was selected. VBS3 from Bohemia Systems was preferred as it is widely used in tactical military simulation, capable to reproduce high ground detail, wide terrain areas, providing a resource database and tools for geo-referenced map generation. The virtual database was modelled in four different quality levels as can be seen in Table 2. The raw data used to model the database variants comprised satellite images, elevation data, geo-spatial vector data and 3D-Objects. The department of geo-information of the Bundeswehr provided orthographic satellite images used in various resolutions and a digital surface model (3D altitude mesh) in 15 meters per pixel (mpp) resolution. Rasterized Shapefiles are used as masks populating the area with different detail maps (e.g. concrete, grass high, etc.) for high-resolution texture details at low altitude. Finally, the terrain was populated using geo-referenced and geo-specific 3D-Objects either provided by VBS3 or created using Blender. All Buildings were modelled after their blueprints to ensure accurate dimensions.

Table 2: Generated terrain databases detailed with raw data used for modelling in meters per pixel (mpp).

Terrain Databases				
Surface Detail (Database Name)	***Resolution Satellite Images***	***Resolution Digital Surface Model***	***Resolution Rasterized Shapefiles***	***Objects***
Low	5 mpp	15 mpp	5 mpp	Yes
Mid	1 mpp	15 mpp	1 mpp	Yes
High	0.2 mpp	15 mpp	0.2 mpp	Yes
High no Building	0.2 mpp	15 mpp	0.2 mpp	No

Facade textures have been photographed and applied after rectification using perspective transformation. Roofs are most prominent in aerial images therefore after identification of type, material and color; their textures have been modeled precisely using free texture databases. Each 3D model consists of geometry, texture map, normal map, specular map and material definition (setting the lighting behavior).
Common industrial tool chains and efforts have been employed in generating the virtual database and its 3D objects. However, the additional requirement of a geo-referenced database necessary for real- and synthetic- dataset comparison increased the development time significantly.
To create synthetic imagery correlating to the test flight, the virtual camera had to be positioned according to recorded telemetry data. Thus, the telemetry data were replayed and used as a trigger to synchronize the image extraction of the virtual environment. Lighting was adjusted using a hemi-spherical lighting model to adjust day light color and strength as well as length and orientation of shadows. The implement-ted camera model of VBS3 was employed which enables the parametrization of focus, aperture, field of view and zoom. These were set to comparable values of its real counterpart. The image resolution has been fixed to 1024x768. The focus was set to infinity equivalent to the real camera.

3.3 Specific dataset categories

The test flight route was separated in nine different classes concerning the nature of the scene. For each of these classes synthetic datasets were extracted. In this paper, the dataset classes Field, Woods and Concrete are discussed.
Field designates a regularly mowed meadow on even terrain. There are no objects in the scene and it is homogenous without any sharp edges or specific high contrast textures. This dataset is intended to demonstrate the differences in terrain image quality between real and synthetic datasets.
Woods designates a dense forest, hiding the ground texture almost completely. In the synthetic datasets, trees have been placed approximately and geo-typical tree models have been used. This dataset was used with caution for two reasons: Firstly, the virtual environment has been modelled using aerial images taken in summer, meaning all trees are in full bloom, while during real test flight they were leafless. Secondly, the height of trees (up to 15 meters) violates the homography constraint, which states that all features shall be in a plane. This reduces the overall results of real and synthetic datasets. However, the highly heterogenic, diverse and cluttered textures are demanding for feature detectors.
Concrete designates a concrete area with transport containers, concrete plates, a mobile bridge and a silver car. This dataset provides sharp edges on several man-made objects as well as a highly textured surface. The object heights are not exceeding two meters, which is small compared to the aircraft altitude. Thus, the homography error was considered negligible.

4 CV-algorithm based evaluation

To investigate the usability of synthetic data for CV-algorithm development and prototyping it was important to use well-known algorithms to allow a comprehensible assessment of acquired results. Therefore, the feature detectors SIFT [12], SURF [13] and MSER [14] were selected as test algorithms. Feature detectors in particular are interesting, because they filter the image domain for recognizable locations, which were used to extract information from the image domain to the feature domain. Often, further processing is solely working on the feature domain (e.g. stereo vision, image stitching). Thus, the performance of feature detectors influences the performance of many specific algorithms and implementations.
Performance of these algorithms is measured using evaluation concept and metric presented by Mikolajczyk in [15]. This *repeatability* metric measures the number of detected corresponding regions in image pairs. It assumes that all features found in Image *I*, mapped to a plane, experience a global geometric trans-formation and can again be found on the transformed plane in Image *J*. The homography matrix H^{ij} describes this geometric transformation and allows reprojection of Features R_J to Image *I*. Features are described as regions *R* on the image with a location and a radius. A feature pair is corresponding when the region ${}^{i}R_j = (H^{ij})^T \, {}^{j}R_j$ reprojected into Image *I* overlaps with R_I [15]:

$$1 - \frac{R_I \cap {}^{I}R_J}{R_I \cup {}^{I}R_J} < e_O \qquad (1)$$

This means a pair is accurate when the overlap error e_O smaller is then the intersection of R_I and ${}^{I}R_J$ divided by its union. The overlap error is set to 0.4 ≙ 40%. This metric is scale dependent, thus punishing differences in region size. The resulting number of correspondences against all possible correspondences is the repeatability measure depicted in percentage. The evaluation is performed in MATLAB and based on the benchmark framework "VLBenchmark" [16].
Additionally to the repeatability, the number of resulting correspondences was analyzed to provide an absolute measure that needs to be considered when evaluating the relative repeatability. Thus, it is possible that the performance of a detector, which only detects four regions in an image but identifies them all results in 100% repeatability. The number of correspondences shows that the detector performs poorly on given dataset, since these few features were not sufficient for possible subsequent processing steps. The necessary ground truth (homography matrix between the images) is first calculated using SURF features [13] matched with a brute force SSD matching algorithm as initialization of an iterative RANSAC-Algorithm for optimization [17].
By using homography, it was possible to measure algorithm performance against generated ground truth in real and synthetic datasets. The decision to create the ground truth of the simulation also using homography deems from the intention of having comparable results therefore using the

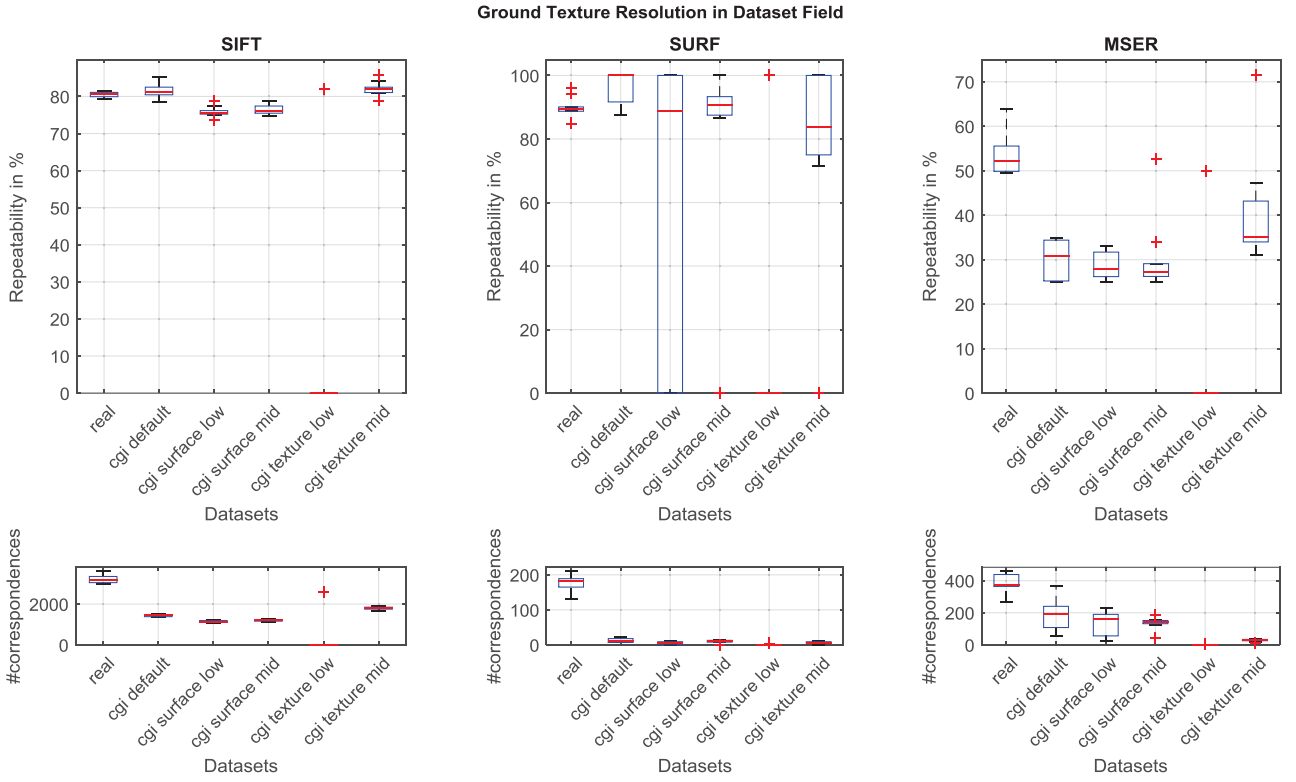


Fig. 2.Evaluation of the dataset class Field using parameter group Ground Texture Resolution. Each box presents the performance of one dataset. The red line inside the box marks the merdian, the upper and lower end mark the 75th and 25th percentile, the black whiskers mark the outmost inliers. Outliers are marked with a red plus.

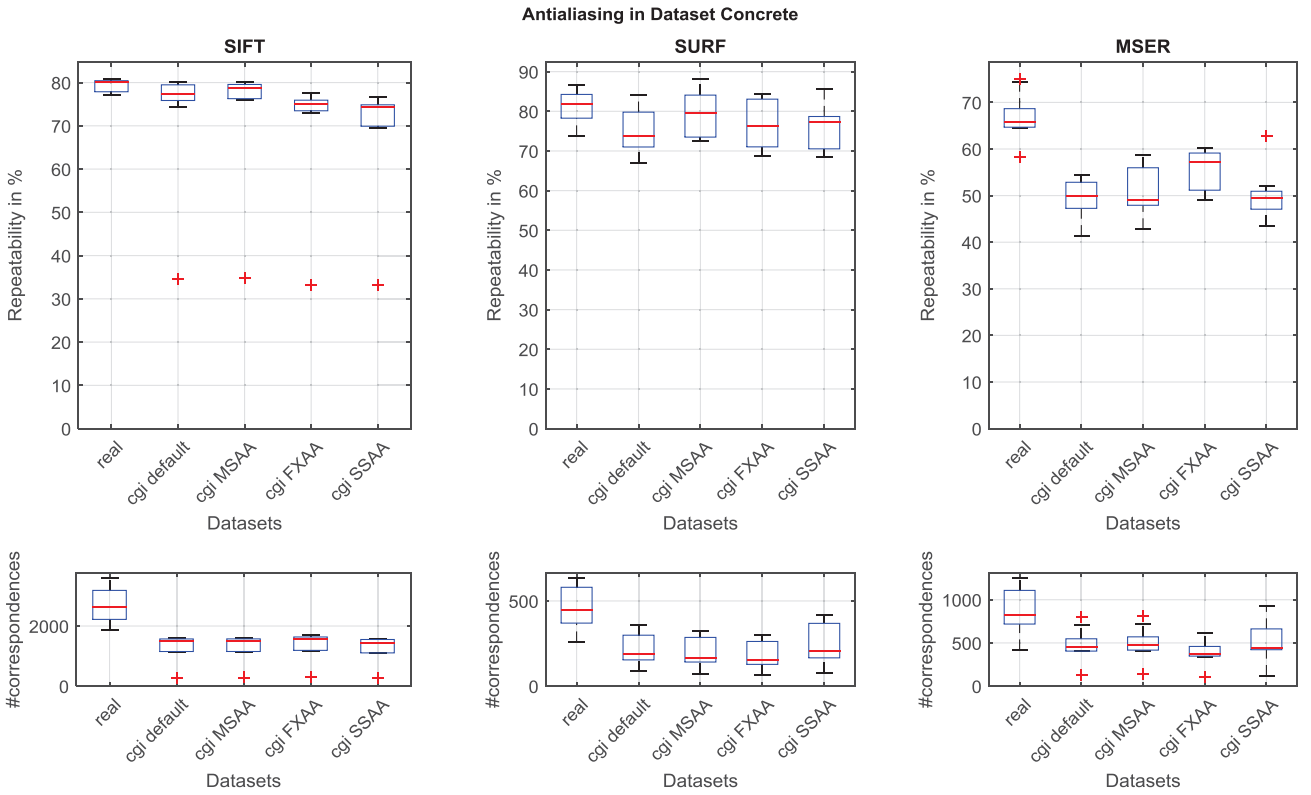


Fig. 3. Evaluation of the dataset class Field using parameter group Ground Texture Resolution. In general synthetic datasets using anti-aliasing increased their performance (except 1.5x SSAA). Using SIFT and SURF synthetic datasets achievied similar performance to real dataset, however finding less corresponding features in total.

same evaluation chain. It needs to be noticed that homography can only be used when either the camera has no or small translation between to images or the displayed surface is planar. Altitude information confirms that the surface of the recorded premises is adequately planar. Due to high aircraft altitudes (75m) and a top down view, small altitude difference of occasional trees and buildings are considered negligible. In the worst case, performance would drop on all datasets and the relative results between tested datasets would not be influenced.

4.1 Evaluated parameter groups

After generation of dataset *Real* and its ground truth, several different synthetic datasets have been created. The first dataset generated was the *Default* dataset, defining the default parameter settings of the rendering engine. Afterwards each additional synthetic dataset was created by modifying a parameter of the engine to identify its influence. Because of space limitations, only a selection of parameters is evaluated in this paper. The parameters have been clustered in two groups. These groups were evaluated against a specific dataset class and results were discussed in detail. Additional significant findings were reviewed without full presentation of the evaluation results.

The first group of parameters was named *ground texture resolution* consisting of surface texture resolution (*surface*) and detail texture resolution (*texture*). The used engine creates ground surfaces by overlaying the geo-specific surface texture with a procedural detail texture. This detail texture emulates a higher resolution of the ground surface but does not provide much contrast due to texture blending. The recorded flight imagery in the real dataset has a ground resolution of 0.03mpp, which corresponded to the highest detail texture resolution setting (see Table 3). This group was tested against dataset class Field that depicts only the ground texture with a repetitive detail texture (meadow).

The second parameter group called *Antialiasing (AA)* embraces three anti-aliasing techniques, namely *Multi Sampling* (*MSAA*), *Fast Approximate* (*FXAA*) and *Super Sampling* (*SSAA*). These techniques all had the goal to reduce jagged nature of sharp edges or lines, which were introduced during rasterization. The reason for different techniques to exist is mostly due the different computation effort necessary. The dataset demonstrating eightfold *MSAA* shows selective sampling depending on polygon-pixel coverage, simple sprites (i.e. tree leaves) are unaffected. The *FXAA* dataset, a post processing antialiasing method, used a high pass filter to detect edges followed by a blur only along those edges. The *SSAA* method simply renders the whole scene in 1.5x of the output resolution and resizes it to its original resolution by averaging. This group was tested against dataset class *Concrete* in detail showing its capabilities on objects in the scene. Each group was additionally tested against *Default* and *Real* dataset to allow absolute comparison.

Table 3: Resolution of parameters in group *ground texture resolution* given in meter per pixel (mpp)

Ground Texture Resolution		
Level	***Surface Texture Resolution***	***Detail Texture Resolution***
default (high)	0.2 mpp	0.03 mpp
mid	1 mpp	0.06 mpp
low	5 mpp	0.12 mpp

4.2 Evaluation and discussion

The first group evaluates the *Field* dataset (empty meadow) against the detectors (SIFT, SURF, MSER) using the datasets *real, computer generated imagery (CGI) default, CGI surface texture low, CGI surface texture mid, CGI texture low and CGI texture mid*. The results are depicted in Fig. 2 using boxplots. Each detector has a separate plot providing its results for each dataset. A red line inside the box marks the median. The upper and lower edges mark the 75th and 25th percentile of the dataset and black lines outside the box mark the maximum and minimum value still considered as inlier. Outliers are marked as a red plus. The results are depicted using the relative repeatability measure indicating the amount of corresponding regions that align with less ten 40% difference of their area. The metric is supported by boxplots (lower row) depicting the absolute number of successfully matched correspondences.

4.2.1 Ground texture resolution

The SIFT detector as shown in Fig. 2 performed quite well (SIFT: 80% repeatability with more than 3000 correspondences on dataset *Real*) on the *Field* dataset class. Comparing the results of dataset *Real* and *Default* results in almost equal performance while the number of correspondences, however was halved. This can be explained by the low contrast of edges due to texture blending. Reducing the resolution of the surface texture results in a drop of repeatability accompanied by a drastically increased standard deviation as indicated by the size of the box. Low surface resolution reduces colored edges since it is smoothing the transition (between background pixels) heavily. This reduces the quality of regions resulting in a lower repeatability rate as shown in the dataset *CGI surface low* but also already indicated in dataset *CGI surface mid*. Reducing the ground surface detail to 0.12 mpp (*CGI Texture Low*) actually disables all detectors. No tested detector was capable to cope the blurring effect of downscaling the detail texture. Interestingly, downscaling the detail texture only once, to a ground resolution of 0.06 mpp was not only allowing the detectors to provide correspondences but to perform even slightly better than dataset *Default*.

Evaluating the SURF detector results showed very high relative repeatability for almost all datasets, but at a very low number of correct correspondences. While the *Default* dataset had 177 correspondences, all CGI datasets only had 12 correspondences on average. Thus, the detector was not providing enough significant features, revealing that the box

filter approximation used to find SURF features could not handle homogenous areas of low structure well. Similar to SIFT, on dataset *texture low* no features could be found.
MSER features tend to be small since they are robustly detected extremal regions by thresholding the image at several thresholds. These extremes however are sparse in the dataset class *Field*, therefore leading to repeatability rates of 53% for dataset *Real* and 28-40% on *CGI* datasets. In addition, the number of features is low in respect to usual values of the MSER detector. The performance between dataset *Default*, *CGI surface mid* and *CGI surface low* dropped by 1% each, demonstrating the robustness of MSER features against surface texture changes on larger scale. However, *CGI texture low* and *CGI texture mid* depict that MSER was strongly influenced by reduction of high frequency details of the image. While *CGI texture mid* provides a median repeatability of 35% it only finds 27 correspondences in total, revealing the low absolute performance of the detector.

4.2.2 Antialiasing

In Fig. 3 the results of group *Antialiasing* in dataset class *Concrete* are depicted. Firstly, all datasets perform well using the SIFT detector. Differences between datasets are only small, since image changes were only minor and mostly limited to edges. Similar to aforementioned evaluation, here the detector found more correspondences on *Real* compared to *CGI* datasets (by factor two). However, repeatability as well as the total number of correspondences on synthetic data demonstrated acceptable performance of the detector. Using *MSAA* increased the repeatability, getting closer to Real performance values. Since this dataset contained no trees (sprites), this technique could perform to its fullest. Using *FXAA* reduces the repeatability by 3% showing a visible blurring effect on object edges. *SSAA* lead to repetitions of irregular aliasing patterns along edges leading to displacement errors of skewed lines, which results in a 5% repeatability drop compared to *Default*.
The SURF detector detected five times fewer features than SIFT but achieved a slightly higher repeatability. With this detector, the *MSAA* dataset performed even better in comparison to dataset *Default* closing in to a performance difference of 3% to the *Real* dataset. *FXAA* also performed better on SURF, showing that box filters could take advantage of Antialiasing. Nonetheless, it should be noted that SURF repeatability rates dispersed much more than on SIFT. *Super Sampling* also showed a slight increase, which however can be considered to be within the error of measurement.
Evaluating the repeatability of MSER on dataset *Real* against SIFT and SURF, displayed a drop in repeatability of 13-15%, making it the least suited region detector for aerial images of this nature. Here, all synthetic datasets heavily dropped in repeatability performance. However, *MSAA* and especially *FXAA* could slightly close the resulting gap. *SSAA* decreased the performance even further.

4.2.3 Objects, Shadows and Lighting

In every dataset class, containing man-made or natural objects the CGI dataset *No Objects* performed best, even better than corresponding *Real* datasets. This is due to the prominent 0.03 mpp pattern together with the 0.2 mpp surface texture along the ground surface. This is mainly due to the homography assumption of a planer surface was fully fulfilled.
During evaluation of dataset class *Woods* the number of correspondences raised extensively for MSER (avg. 6000) and SURF (avg. 1700) even higher than evaluated real datasets (MSER: 2000 and SURF: 900). SIFT however behaved similar to its results in dataset class *Field* or *Concrete*. The difference between real and synthetic datasets lay in the absence of leaves in dataset *Real*. Leaves created a large number of extremes, because of their cluttered and overlapping distribution. Repeatability rates using *Woods* datasets ranged from 34 to 67% indicating violation of the homography ground truth constraint.
Dataset classes *Woods* and *Concrete* where used to evaluate the effect of shadows in CGI on detector repeatability and number of correspondences leading to differences of less than 1% in repeatability and number of correspondences. Even the removal of shadows did not change this effect. Thus, shadow generation is not influencing the performance of feature detectors.

5 Conclusion

In this paper, the next step of a concept to evaluate synthetic datasets using computer vision algorithms has been presented. Here, feature detectors were used to evaluate their performance on real and scene-wise corresponding synthetic datasets depicting airborne reconnaissance imagery. In addition, differences in behavior between the detectors have been discussed. The objective was to investigate the use of synthetic environments for CV-algorithm prototyping and evaluation. Additionally the influence of specific rendering techniques has been investigated.
Therefore, a test flight has been conducted recording airborne imagery and position of the aircraft, which was reproduced in a synthetic environment. To achieve correspondence, the terrain has been modelled in geo-referenced detail (textures, terrain and man-made objects). The recorded images have been separated into three terrain classes. The performance was evaluated using the repeatability metric, which used a homography-based ground truth.
In general, *Real* datasets performed roughly equal for SIFT and SURF detectors and 20% better for the MSER detector then *Default* synthetic datasets. Additionally in total, more feature correspondences have been found in real datasets, due to more extremes in the images (e.g. intensity, edges). It has been identified that a high quality ground texture (at least half of the cameras ground resolution) was mandatory.
These textures could however be procedural and repetitive. For increased performance, a high-resolution satellite image (0.2 mpp) was blended with the procedural texture. Additional

Fig. 4. Example patches of datasets (from left to right). *Field: Real, CGI Default* and *CGI Texure Low. Concrete: Real, CGI Default* and *CGI MSAA*.

rendering methods, such as *Multi Sampling* (for SURF, SIFT) or *Fast Approximate* (for MSER) *Antialiasing* improved the repeatability of synthetic datasets. Synthetic datasets with and without objects have been evaluated resulting in too high performance when objects are missing, due to its planar surface. Shadow generation techniques were also tested showing no influence on repeatability measures.

Aforementioned results lead to the conclusion that used setup demonstrated the usability of synthetic environments. Therefore, feature-based algorithms can be prototyped or evaluated in synthetic environments when mentioned constraints are considered and can be improved using anti-aliasing methods.

The next steps will be a dependency analysis weighing acquired results against numerical distance measures of MPEG7 image retrieval descriptors, intended to identify image parameters influencing the performance of synthetic datasets. Furthermore, a metric allowing evaluation on perspective datasets or without the planar level constraint would increase the range of possible datasets. Moreover, the study could be extended with additional rendering techniques, image descriptors and metrics. In addition, the evaluation of computer vision algorithms could be extended to CV-algorithms that are more complex such as object detectors or trackers.

6 References

[1] R. Collins, X. Zhou, and S. K. Teh, "An open source tracking testbed and evaluation web site," IEEE Int. Work. Perform. Eval. Track. Surveill., pp. 17–24, 2005.

[2] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," ITE J. (Institute Transp. Eng., vol. 74, no. August, pp. 22–26, 2004.

[3] W. Burger and M. Barth, "Virtual Reality for enhanced computer vision," Virtual Prototyp. Virtual Environ. …, 1995.

[4] G. Hummel, L. Kovács, P. Stütz, and T. Szirányi, "Data Simulation and Testing of Visual Algorithms in Synthetic Environments for Security Sensor Networks," in Future Security, 2012, vol. 318, pp. 212–215.

[5] R. Szeliski, Computer vision: algorithms and applications. 2011.

[6] K. Martull, S., Peris, M., & Fukui, "Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps," Int. Conf. Pattern Recognit., pp. 117–118, 2012.

[7] H. Tamura and H. Kato, "Proposal of international voluntary activities on establishing benchmark test schemes for AR/MR geometric registration and tracking methods," in ISMAR. 8th IEEE Int. Symp. on, 2009, pp. 233–236.

[8] M. Berger, J. a. Levine, L. G. Nonato, G. Taubin, and C. T. Silva, "A Benchmark for Surface Reconstruction," ACM Trans. Graph., vol. 32, no. 2, pp. 20:1–20:17, 2013.

[9] J. Ferwerda, "Three varieties of realism in computer graphics," Proc. SPIE Hum. Vis. Electron. …, vol. SPIE 5007, pp. 290–297, 2003.

[10]G. Hummel and P. Stuetz, "Using Virtual Simulation Environments for Development and Qualification of UAV Perceptive Capabilities: Comparison of Real and Rendered Imagery with MPEG7 Image Descriptors," in MESAS: First International Workshop, Rome, Italy, May, 2014, vol. 8906 LNCS, pp. 27-43.

[11]F. Boehm and A. Schulte, "Scalable COTS Based Data Processing and Distribution Architecture for UAV Technology Demonstrators," Eur. Telem. Test Conf. etc …, 2012.

[12]D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[13]H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," Lect. Notes Comput. Sci., vol. 3951 LNCS, pp. 404–417, 2006.

[14]J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Br. Mach. Vis. Conf.*, pp. 384–393, 2002.

[15]K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J.* ..., 2005.

[16]K. Lenc, V. Gulshan, and A. Vedaldi, "VLBenchmarks," 2012. [Online]. Available: http://www.vlfeat.org/benchmarks/. [Accessed: 25-Mar-2015].

[17]M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24. pp. 381–395, 1981.

Efficient Computation of Distribution Kernels and Distances

Hongjun Su[1] **and Hong Zhang**[1]
[1]Department of Computer Science and Information Technology
Armstrong State University, Savannah, GA, USA

Abstract - *Similarity and dissimilarity measures such as kernels and distances are key components of classification and clustering algorithms. We propose an efficient algorithm for computation of kernel and distance functions between two probability distributions. The complexity of the proposed algorithm is insensitive to the dimension of the input space and therefore especially suitable for high dimensional distributions.*

Keywords: Distribution, Distance, Kernel, Density, Algorithm

1 Introduction

A kernel is a similarity measure that is the key component of support vector machine ([4]) and other machine learning techniques. More generally, a distance (a metric) is a function that represents the dissimilarity between objects.

In many pattern classification and clustering applications, it is useful to measure the similarity between probability distributions. Even if the data in an application is not in the form of a probability distribution, they can often be reformulated into a distribution through a simple normalization. For example, a gray scale image can be viewed as a sample from a 2D distribution. A dataset from a flow cytometry experiment is from a high dimensional probability space.

A large number of divergence and affinity measures on distributions has already been defined in traditional statistics. These measures are typically based on the probability density functions. The numerical calculations of these functions could incur significant computational cost especially in high dimensional spaces. In this paper, we propose an efficient algorithm for computation of kernel and distance functions between two probability distributions.

This paper is organized as follows. Section 2 introduces common kernels and distances defined on probability distributions and their computational challenges. In Section 3, an algorithm of linear complexity is proposed for computing the kernel and distance functions. Experimental results on Gaussian mixture distributions are presented in Section 4. In Section 5 provides conclusions and proposals for further improvements.

2 Distance and Similarity Measures Between Distributions

Given two probability distributions, there are well known measures for the differences or similarities between the two distributions.

The Bhattacharyya affinity ([1]) is a measure of similarity between two distributions:

$$B(p,q) = \int \sqrt{p(x)q(x)}dx$$

This is clearly a kernel on the space of density functions, since the feature map is:

$$\Phi(p) = \sqrt{p(x)}$$

In [7], the probability product kernel is defined as a generalization of Bhattacharyya affinity:

$$k^{prob}(p,q) = \int p(x)^{\rho} q(x)^{\rho} dx$$

The Bhattacharyya distance is a dissimilarity measure related to the Bhattacharyya affinity:

$$D_B(p,q) = -\ln\left(\int \sqrt{p(x)q(x)}dx\right)$$

The Hellinger distance ([6]) is another metric on distributions:

$$D_H(p,q) = \sqrt{\frac{1}{2}\int\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}$$

The Kullback-Leibler divergence ([8]) is defined as:

$$D_{KL}(p,q) = \int\left(\ln\frac{p(x)}{q(x)}\right)p(x)dx$$

Related to the distance measures are the statistical tests to determine if two samples are drawn from different distributions. Examples of such tests include the Kolmogorov-Smirnov statistic ([10]) and the kernel based tests ([5]). In [11], distance measures based cumulative distribution functions are introduced.

The kernel and distance measures typically defined in terms of the two density functions. The numerical evaluation of these functions presents computational challenges. Consider the direct histogram representation of the density function. If the sample space is of dimension *d* and each dimension is divided into *k* subintervals, then the number of hypercubes will be k^d which can become intractable as *d* increases.

In practical applications, the exact probability distributions are usually unknown. Typically the available data are in the form of i.i.d. samples from the distributions. Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ be samples from the two distributions $p(x), q(x)$. We would like to obtain an estimate of the kernel or distance value on the two distributions. We propose an algorithm that has a complexity proportional to the sample size and is insensitive to the dimension.

3 An algorithm on Kernel and Distance Functions Between Two Distributions

A classical technique to estimate a density function is the Parzen window method ([9]). Given a sample $x_1, x_2, \ldots, x_n$ and a hypercube window function of width h_n, the density function is estimated as

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^d}\phi\left(\frac{x - x_i}{h_n}\right)$$

With the density estimates for the two distributions, we may compute a kernel or distance by evaluate the appropriate integral. However, as illustrated in the previous section, the computational cost can be prohibitive in multidimensional cases. To avoid processing a large number of boxes, our algorithm will instead proceed along the sample points.

Memory requirement is another challenge. Again the storage of k^d boxes may not be feasible. The sample size *n* may be a much smaller number, so the data is sparse in the entire sample space. We propose to represent the density function with a data structure similar to sparse matrix representation, such as a dictionary of keys or a coordinate list.

The following is a pseudo code of our proposed algorithm.

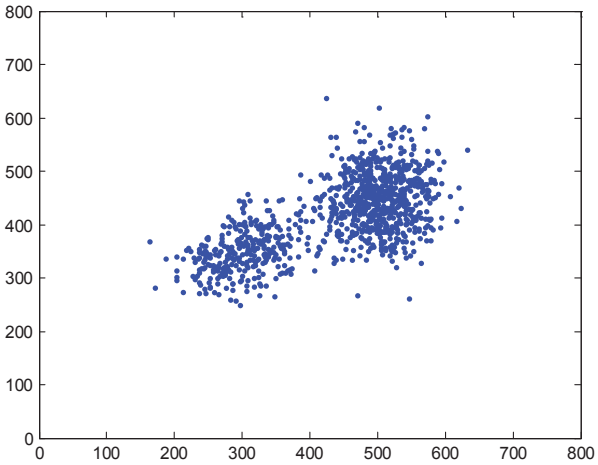
```
1.  create a sparse matrix p
2.  for i = 1 to n
3.       for all lattice points x inside
    the hypercube centered at xi
4.            p[x] = p[x] + 1/(nh^d)
5.  create a sparse matrix q
6.  for i = 1 to m
7.       for all lattice points x inside
    the hypercube centered at yi
8.            q[x] = q[x] + 1/(mh^d)
9.  for all non-zero points x in p and q
10. b = b + sqrt(p[x]q[x])
```

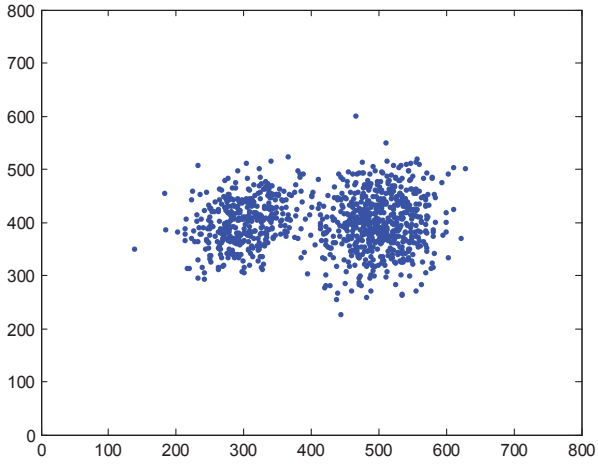
The above code computes the Bhattacharyya affinity. Other measures can be calculated in the same fashion.

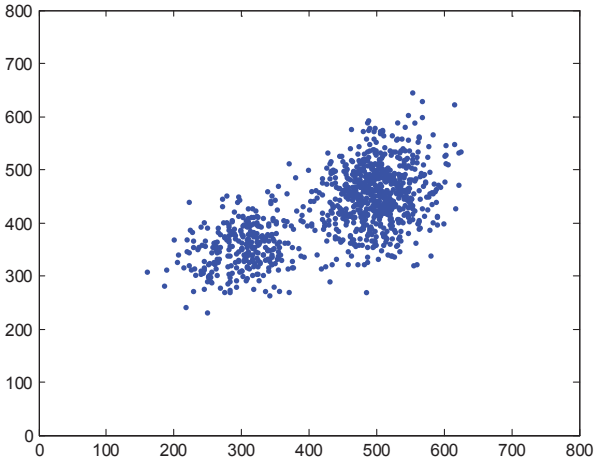
The computational complexity of the algorithm is linear of the sample size $O(n+m)$. The dimension *d* is relevant in the window function. However, it is not significant since the window function typically has a small, fixed support.

4 Experimental Results and Discussions

Two different Gaussian mixture distributions ([2]) and two samples from each distribution are constructed to test the algorithm. The first two figures show the samples from the first distribution and the last two figures from the second distribution. The two Gaussian mixtures have slightly different means, covariance matrices and weights.







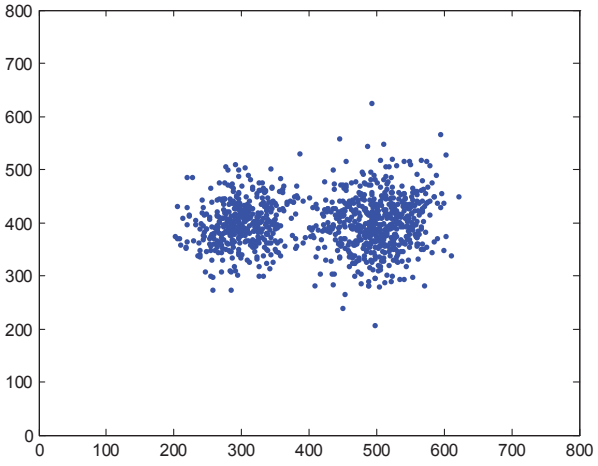


Figure 1. Four samples from two Gaussian mixtures

Using the proposed algorithm, the Gram matrix of the Bhattacharyya kernel for the four data samples is given by:

$$\begin{pmatrix} 1.0000 & 0.9353 & 0.7859 & 0.7903 \\ 0.9353 & 1.0000 & 0.7874 & 0.7896 \\ 0.7859 & 0.7874 & 1.0000 & 0.9311 \\ 0.7903 & 0.7896 & 0.9311 & 1.0000 \end{pmatrix}$$

The matrix clearly shows that the similarities between the two samples of the same distribution are much higher than those between the different distributions.

5 Conclusions and Future Work

In this paper, we presented an algorithm for estimating kernel and distance functions on probability distributions using discrete samples. The algorithm has a complexity of $O(n+m)$ and is insensitive to the dimension of the sample space.

For future work, we propose to improve the algorithm by developing a data structure capable of representing the overlapping windows efficiently.

6 References

[1] Bhattacharyya, A., "On a measure of divergence between two statistical populations defined by their probability distributions". Bulletin of the Calcutta Mathematical Society 35: 99–109, (1943).

[2] Bishop, Christopher, Pattern recognition and machine learning. New York: Springer, (2006).

[3] Bogachev, V.I.; Kolesnikov, A.V. "The Monge-Kantorovich problem: achievements, connections, and perspectives". Russian Math. Surveys 67: 785–890.

[4] Boser, B. E.; Guyon, I. M.; Vapnik, V. N., "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. p. 144, (1992).

[5] Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Scholkopf, B.; Smola, A., "A kernel two-sample test", J. Machine Learning Research, 13, 723-773, (2012).

[6] Hellinger, Ernst, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen", Journal für die reine und angewandte Mathematik (in German) 136: 210–271, (1909).

[7] Jebara, T.; Kondor, R.; Howard, A., "Probability Product Kernels," J. Machine Learning Research, 5, 819-844, (2004).

[8] Kullback, S.; Leibler, R.A., "On Information and Sufficiency". Annals of Mathematical Statistics 22 (1): 79–86, (1951).

[9] Parzen, E., "On Estimation of a Probability Density Function and Mode". The Annals of Mathematical Statistics 33 (3): 1065, (1962).

[10] Smirnov, N.V., "Approximate distribution laws for random variables, constructed from empirical data" Uspekhi Mat. Nauk , 10 pp. 179–206 (In Russian), (1944).

[11] Su, H.; Zhang, H., "Distances and Kernels Based on Cumulative Distribution Functions," Proceedings of the 2014 International Conference on Image Processing, Computer Vision, & Pattern Recognition, 357-361, (2014).

Dissimilarity Representations Using l_p-norms in Eigen Spaces

Sang-Woon Kim
Department of Computer Engineering, Myongji University, Yongin, 449-728 Korea

Abstract— *This paper presents an empirical evaluation on a dissimilarity measure strategy by which dissimilarity-based classifications (DBC) can be implemented efficiently. In DBC, classification is not based on feature measurements of individual objects (a set of attributes), but rather on a suitable dissimilarity measure among the individual objects (pair-wise object comparisons). One problem of DBC is the high dimensionality of the dissimilarity space. To address this issue, two kinds of solutions have been proposed in the literature: prototype selection (PS)-based methods and dimension reduction (DR)-based methods. In this paper, instead of utilizing the PS-based or DR-based methods, we study a way of performing DBC in Eigen spaces (termed as EDBC), spanned by the subset of principal eigenvectors, extracted from the training dataset through a principal component analysis. Specifically, in EDBC, we use lp-norms in combination with a rotation to eigenvectors to compute distances in a vector space, for constructing a dissimilarity-based classifier. The experimental results, obtained with artificial and real-life benchmark datasets, demonstrate that when the dimensionality of the Eigen spaces has been appropriately chosen, the classification accuracy of DBC can be improved.*

Keywords: statistical pattern recognition, dissimilarity-based classification, dissimilarity representation, dissimilarity measures

1. Introduction

Dissimilarity-based classification (DBC) [1] is a way of defining classifiers among the classes, in which the process is not based on feature measurements of individual objects (a set of features), but rather on a suitable dissimilarity measure among the individual objects (pair-wise object comparisons). The advantage of this strategy is that it offers a different way to include expert knowledge on the objects in classifying them [2]. A major task we encountered when dealing with DBC is that we need to measure the inter-pattern dissimilarities for all the training data to ensure there is no zero distance between objects of different classes. Consequently, the classes do not overlap, and therefore, the lower error bound is zero. Thus, the classification performance of DBC relies heavily on how well the dissimilarity space, which is determined by the dissimilarity matrix, is constructed. To improve the performance, therefore, we need to ensure that the dissimilarity matrix is well designed. In order to address this issue, two things should be considered: the representation set (prototypes) and the distance metric. The latter is used for measuring the dissimilarity between the pair-wise objects, while the former, to which other objects are compared, is used to generate a dissimilarity representation.

First, regarding the representation set, a dissimilarity representation is constructed in which each dimension corresponds to the distances of all prototypes from an object. A problem that we encounter when the training dataset (and also the prototype subset) is sufficiently large is the high dimensionality of the dissimilarity space. To address this problem, two kinds of solutions have been proposed in the literature: prototype selection (PS)-based methods [1], [3] and dimension reduction (DR)-based methods [4], [5]. The PS-based method works by directly choosing a small set of prototypes from the training samples. On the other hand, the DR-based method consists of building the dissimilarity matrices using all the available training samples and subsequently applying some of the standard DR schemes.

Next, with regards to the distance metric, investigations have focused on measuring the appropriate dissimilarity using various l_p distances, including the Manhattan distance (l_1 distance) and the Euclidean distance (l_2 distance) [1], the Hausdorff and modified Hausdorff distances [6], [7], and traditional measures, such as those used in template matching and correlation-based analysis [1], [8]. Furthermore, in order to find a better distance measure involving a training dataset, various metric learning algorithms have been proposed in the literature [9], [10], [11]. In one case, when given certain kinds of dissimilarities, new dissimilarity measures defined using the given ones were proposed and evaluated to optimize the measures for nearest neighbor classification [11].

On the other hand, subspace methods of pattern recognition [12] belong to a technique in which the object classes are not primarily defined as bounded regions in a feature space, but rather given in terms of linear subspaces defined by the principal component analysis (PCA) [13]. The length of a vector projected onto a class subspace, or the distance of the vector from the class subspace, is a measure of similarity or degree of membership for that particular class, and serves as the discriminant function of these methods. For face recognition, a typical PCA-based subspace approach, called Eigen face, has been proposed in the literature [14], [15], and has undergone continuous development [17], [18].

From this point of view, instead of utilizing a PS-based or DR-based method, in this paper we report on our study of a way of carrying out DBC in Eigen spaces (termed EDBC), spanned by the subset of principal eigenvectors, extracted

from the training dataset through the PCA. In particular, in EDBC, we use l_1-norms in combination with a rotation to eigenvectors to compute distances in a vector space, thus constructing a dissimilarity-based classifier. The major task of our study was to determine how the dissimilarity measure can be effectively computed.

The main contribution of this paper is to present an empirical evaluation of the three approaches of reducing the dimensionality of dissimilarity spaces for optimizing DBCs. This evaluation shows that DBCs can be improved by measuring the dissimilarity with l_1-norm in Eigen spaces after constructing them using a PCA. Here, the aim of using the Eigen space approach, instead of the PS or DR based method, is to accommodate some useful information for discrimination and to avoid the problem of finding the optimal number of prototypes. Our experimental results, obtained with the three approaches for artificial and real-life benchmark datasets, demonstrate that when the dimensionality of the Eigen spaces has been appropriately selected, and the dissimilarity has been effectively measured in the subspaces, the classification accuracies of the DBC can be improved.

The remainder of the paper is organized as follows. In Section 2, after providing a brief introduction to DBC, we present an explanation of the PS and DR based methods and an improved DBC. In Sections 3 and 4, we present the experimental set-ups and the results obtained from the work using artificial and real-life datasets. Finally, in Section 5, we present our concluding remarks, as well as some additional features that deserve to be studied further.

2. Related Work

2.1 Dissimilarity Representation [1]

A dissimilarity representation of a set of object samples, $T = \{\boldsymbol{x}_i\}_{i=1}^{n} \in \mathbb{R}^d$, is based on pair-wise comparisons and is expressed, for example, as an $n \times m$ dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, where $P = \{\boldsymbol{p}_j\}_{j=1}^{m} \in \mathbb{R}^d$, a prototype set, is extracted from T, and the subscripts of D represent the set of elements of which the dissimilarities are evaluated. Thus, each entry, $D_{T,P}[i,j]$, corresponds to the dissimilarity between the pairs of objects, $\boldsymbol{x}_i$ and $\boldsymbol{p}_j$, where $\boldsymbol{x}_i \in T$ and $\boldsymbol{p}_j \in P$. Consequently, when given a distance measure between two vectors, $\rho(\cdot,\cdot)$, an object, $\boldsymbol{x}_i$, is represented as a column (or a row) vector, $\delta(\boldsymbol{x}_i, P)$, as follows:

$$\begin{aligned} \delta(\boldsymbol{x}_i, P) &= [\rho(\boldsymbol{x}_i, \boldsymbol{p}_1), \rho(\boldsymbol{x}_i, \boldsymbol{p}_2), \cdots, \rho(\boldsymbol{x}_i, \boldsymbol{p}_m)], \quad (1) \\ & \quad 1 \le i \le n. \end{aligned}$$

Here, the dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, defines vectors in a *dissimilarity space* on which the d-dimensional object, $\boldsymbol{x}$, given in the input-feature space, is represented as an m-dimensional vector, $\delta(\boldsymbol{x}, P)$. In this paper, the vector is simply denoted by $\delta(\boldsymbol{x})$. As explained previously, the dissimilarity based approach is originally developed for objects, not for feature vectors. However, the dissimilarities are now used as features and can be replaced without any problem by similarities. Thus, it should be noted that an approach defined for arbitrary distances between full objects is used for distances measured in a feature space [19].

Based on the brief explanation, a conventional algorithm for DBC is summarized as in the following:

1) Select the prototype subset, P, from the training set, T, using one of the prototype selection methods described in the literature [1].
2) Using Eq. (1), compute the dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, in which each dissimilarity is computed on the basis of the distance measures described in the literature [1].
3) For a test sample, $\boldsymbol{z}$, compute a dissimilarity vector, $\delta(\boldsymbol{z})$, using the same measure used in Step 2.
4) Achieve the classification by invoking a classifier built in the dissimilarity space and operating it on the dissimilarity vector $\delta(\boldsymbol{z})$.

2.2 Prototype Selection (PS) Methods [4]

The intention of selecting prototypes is to guarantee a good tradeoff between the recognition accuracy and the computational complexity, when DBC is built on $D_{T,Y}[\cdot,\cdot]$ rather than $D_{T,T}[\cdot,\cdot]$. Various PS methods (e.g., *Random*, *RandomC*, *KCentres*, *ModeSeek*, *LinProg*, *FeatSel*, *KCentres-LP*, and *EdiCon*) have been proposed in the literature [1], [3]. In the interest of compactness, the details of these methods are omitted here, but can be found in the related literature.

The DBCs summarized in Section 2.1, in which the representation prototypes are selected with a PS, are referred to as PS-based DBCs, or simply as PS-based methods. An algorithm for PS-based DBCs is summarized in the following:

1) Select the representation subset, Y, from the training set, T, by resorting to one of the prototype selection methods.
2) Using Eq. (1), compute the dissimilarity matrix, $D_{T,Y}[\cdot,\cdot]$, in which each individual dissimilarity is computed based upon the measures described in the literature.
3) For a test sample, z, compute a dissimilarity vector, $\delta_Y(z)$, using the same measure used in Step 2.
4) Achieve the classification by invoking a classifier built in the dissimilarity space and by operating the classifier on the vector, $\delta_Y(z)$.

2.3 Dimension Reduction (DR) Schemes [4]

In DBCs, a good selection of prototypes seems to be crucial to succeed with the classification algorithm. The prototypes should avoid redundancies in terms of selection of similar samples, and include as much information as possible. However, it is difficult for us to find the optimal number of prototypes. Furthermore, when selecting them,

some useful information for discrimination is likely to be lost. To avoid these problems, an alternative approach where *all* of the available data are selected as prototypes, and, subsequently, a dimension reduction scheme, such as PCA, can be applied to the reduction of dimensionality. That is, we prefer not to directly select the prototypes from the training data; rather, we employ a way of using the dimension reduction scheme after computing the dissimilarity matrix with the entire set of training data.

DBCs, in which the dimensionality of dissimilarity spaces is reduced with a DR, are referred to as DR-based DBCs or DR-based methods. An algorithm for DR-based DBCs is summarized in the following:

1) Select the entire set of training samples, T as the representation set, Y.
2) Using Eq. (1), compute the dissimilarity matrix, $D_{T,T}[\cdot,\cdot]$, in which each individual dissimilarity is computed based on the measures described in the literature. After computing the $D_{T,T}[\cdot,\cdot]$, reduce its dimensionality by invoking a DR.
3) This step is the same as Step 3 in the conventional DBC.
4) This step is the same as Step 4 in the conventional DBC.

2.4 DBC in Eigen Spaces [18]

To overcome the limitation caused by the variations and the outlier data, and consequently to improve the performance of DBC, in this paper, we measure the dissimilarity between paired objects with l_2 and l_1-norms in a transformed subspace, rather than in the original input-feature space. The basic strategy of the proposed technique is to solve the classification problem by first mapping the input-feature space to Eigen spaces, and then constructing a dissimilarity matrix with the distance measures in the Eigen spaces. Finally, DBC is performed on the dissimilarity space to reduce the classification error rates.

The proposed approach, which is referred to as an Eigen space DBC(EDBC), is summarized in the following:

1) Select the entire training set T as the prototype subset P.
2) For each object vector, $\boldsymbol{x}_i \in T$, $(i = 1, \cdots, n)$, after computing a transformation matrix, A (i.e., the arranged principal eigenvectors), and a mean vector, $\boldsymbol{m}$ transform the input-feature vector, $\boldsymbol{x}_i$, into the corresponding q-dimensional feature vector, $\boldsymbol{y}_i$, using a transformation formula, $\boldsymbol{y}_i = A^T(\boldsymbol{x}_i - \boldsymbol{m})$.
3) Using Eq. (1), compute the dissimilarity matrix, $D_{T,T}[\cdot,\cdot]$, in which each individual dissimilarity, $\rho(\boldsymbol{x}_i, \boldsymbol{x}_j)$, is measured with the dissimilarity computed in the subspace of A, $\rho(\boldsymbol{y}_i, \boldsymbol{y}_j)$, where $\rho(\cdot,\cdot)$ denotes an l_2 or l_1 metric.
4) This step is the same as Step 3 in the conventional DBC.
5) This step is the same as Step 4 in the conventional DBC.

The rationale of this strategy is presented in a later section together with the experimental results.

3. Experimental Setup

In order to evaluate the effectiveness of l_2 and l_1-norms for performing DBCs in the Eigen space, we compared the classification accuracies achieved with them against certain kinds of PS-based and DR-based methods. This was done by performing experiments on two artificial datasets [20] and on other multivariate datasets cited from the UCI Machine Learning Repository [21].

In all the experiments, the datasets were first randomly split into training sets and test sets at a ratio of 75:25 (%). Then, the training and testing procedures were repeated 30 times and the results obtained were averaged. For the PS-based method, the representation set was randomly selected from the training dataset, while, for the DR-based method, the representation set was defined by a subset of principal eigenvectors, extracted from the training dataset through a principal component analysis.

To evaluate the classification accuracies of DBCs designed using these two methods, and those of the principal Eigen spaces; different classifiers, such as the k-nearest neighbor classifier and the linear Bayes normal classifier, were employed and implemented with PRTools. In subsequent sections, these will be denoted *knnc* (where k = 1) and *ldc*, respectively.

In summary, DBC was carried out differently in six ways (named PS-DBC-l_2, PS-DBC-l_1, DBC-DR-l_2, DBC-DR-l_1, Eig-DBC-l_2, and Eig-DBC-l_1). The details of these settings for DBC are itemized as follows:

1) PS-DBC-l_2: the representation set P is randomly selected from the training dataset T (i.e., $P \subseteq T$) and the dissimilarity between the pairwise objects, $\delta(\cdot, P)$, is measured in l_2-norm (refer to Section 2.2).
2) PS-DBC-l_1: similarly done as in PS-DBC-l_2, but $\delta(\cdot, P)$ is measured in l_1-norm.
3) DBC-DR-l_2: the P is a subspace defined by a subset of principal eigenvectors, extracted from $(n \times n)$-dimensional dissimilarity matrix through a PCA, and $\delta(\cdot, P)$ is measured in l_2-norm (refer to Section 2.3).
4) DBC-DR-l_1: similarly done as in PS-DBC-l_2, but $\delta(\cdot, P)$ is measured in l_1-norm.
5) Eig-DBC-l_2: the P is randomly selected from the Eigen space generated from T through a PCA and $\delta(\cdot, P)$ is measured in l_2-norm (refer to Section 2.4).
6) Eig-DBC-l_1: similarly done as in Eig-DBC-l_2, but $\delta(\cdot, P)$ is measured in l_1-norm.

4. Experimental Results

4.1 Experiment # 1 (Highleyman/Difficult Data)

First, the experimental results obtained with the two classifiers trained in the six ways for two artificial data, namely, Highleyman and Difficult Data [20], were checked. Fig. 1 shows a comparison of the classification error rates obtained with the PS and DR-based methods, and EDBC, for Highleyman and Difficult. Here, Highleyman and Difficult Data, which are 2-dimensional 2-class datasets of the positive and negative examples of $[200, 200]$, are defined by a Gaussian distribution, $Gauss(\boldsymbol{m}, \boldsymbol{S})$, and the class priors $P(1) = P(2)$. Moreover, the x and y-axes represent the cardinality of the representation set (and the reduced dimensionality of the Eigen space) and the estimated error rates, respectively.

From these plots as shown in Fig. 1, the following observations can be made:

First, the reader first should observe that the classification accuracy of EDBCs could be improved by means of appropriately choosing the dimensionality of the subspaces. In particular, it should be observed that the error rates of EDBCs using the l_2 and l_1 metrics are almost the same.

Second, it should also be pointed out that, for Difficult Data, the error rates of Eig-DBC and PS-DBC (indicated with the solid and dot lines, respectively) generally decrease as the dimension of the subspaces increases; the slope of the former is slightly greater than that of the latter. However, the error rate of DBC-DR, which is indicated with the dash-dot lines, is nearly flat.

Third, it should be mentioned that, for Highleyman, the error rates of Eig-DBC-l_2 and l_1, marked with the $\bigcirc$ and $\square$ symbols, respectively, are completely the same, while, for Difficult Data, they are slightly different. In the case of *knnc*, the error rate of the former is marginally higher than that of the latter, but in the case of *ldc*, the situation is the opposite.

4.2 Experiment # 2 (UCI Data)

In order to investigate the behavior of these approaches, we repeated the above experiment with a few real-life datasets cited from the UCI repository [21].

The specific characteristics (# dimensions / # objects / # classes) of the UCI datasets are summarized in Table 1.

First, this was done with two traditional DBCs (i.e., the PS and DR-based methods), and with EDBC for two datasets cited from the UCI, namely, Laryngeal1 and Dermatology. Fig. 2 shows a comparison of the error rates obtained. Here, the details of the experiment are the same as in Fig. 1.

From the plots shown in Fig. 2, the results obtained are similar to those in Fig. 1. In particular, when choosing the dimensionality of the subspaces (and also the cardinality of the representation set) as the same as that of the original input-feature space, all of the error rates achieved with l_1-norm, marked with $\triangle$, $\triangleleft$, and $\square$ symbols in red, are *slightly* lower than those of l_2, marked with $\triangledown$, $\triangleright$, and $\bigcirc$ symbols in blue color. This means that l_1-norm works better than l_2-norm does.

Second, to measure the effectiveness of the l_2 and l_1 metrics for performing EDBC, as well as DBC, the experiment described above (of estimating error rates) was repeated with other UCI datasets [i.e., Diabetes (768/8/2), Heart (297/13/2), Liver (345/6/2), Sonar (208/60/2), and Wine(178/13/3)], where the three numbers in brackets represent the numbers of dimensions d, samples n, and classes c, respectively. Here, the dissimilarity was measured with the two metrics. Moreover, the dimensionality of the dissimilarity matrices was determined to be the same as that of the original input-feature space (i.e., $|P| = |T|$ and $q = d$). That is, the dimensions of the Eigen spaces (and the numbers of prototypes) for Diabetes, Heart, Liver, Sonar, and Wine are 8, 13, 6, 60, and 13, respectively. The other details of the experiment are the same as in Fig. 1.

Table 2 presents the classification error rates obtained with *knnc* and *ldc* for the five UCI datasets, showing the characteristics similar to the ones we obtained in Fig. 2. In the table, we show that almost all of the lowest error rates (marked with the * symbol) were achieved with the l_1 metric except with 'Liver'.

Although it is hard to quantitatively compare the various approaches using l_2 and l_1 distances, we merely counted the numbers of the error rates highlighted with the * marker, obtained with the five UCI datasets. From this count, it can be observed that Eig-DBC-l_1 performs slightly better (higher score) than the other DBCs in terms of the classification accuracy. From this consideration, a question arises: *Why does the Eig-DBC-l_1 approach not work in certain applications?* The theoretical investigation of the underlying reason for this remains to be done.

In review, it is not easy to crown one particular measuring method with superiority over the others in terms of solving the dissimilarity-measuring problem. However, in terms of classification accuracy, EDBC using l_1 metric (Eig-DBC-l_1) seems to be more useful for certain kinds of significant datasets than the other PS/DR based methods (e.g., PS-DBC-l_2, PS-DBC-l_1, DBC-DR-l_2, DBC-DR-l_1, and Eig-DBC-l_2).

5. Conclusions

In an effort to improve the classification performance of DBC, instead of utilizing the well-known PS-based and DR-based methods, we studied a distance measuring technique based on Eigen spaces of data. To achieve this improvement of DBC, we first computed eigenvectors and eigenvalues of the training dataset. Then, we performed DBC in the Eigen spaces spanned by the subset of principal eigenvectors, where the dissimilarity was measured with the family of l_p-norms, including the Manhattan (l_1) distance and the

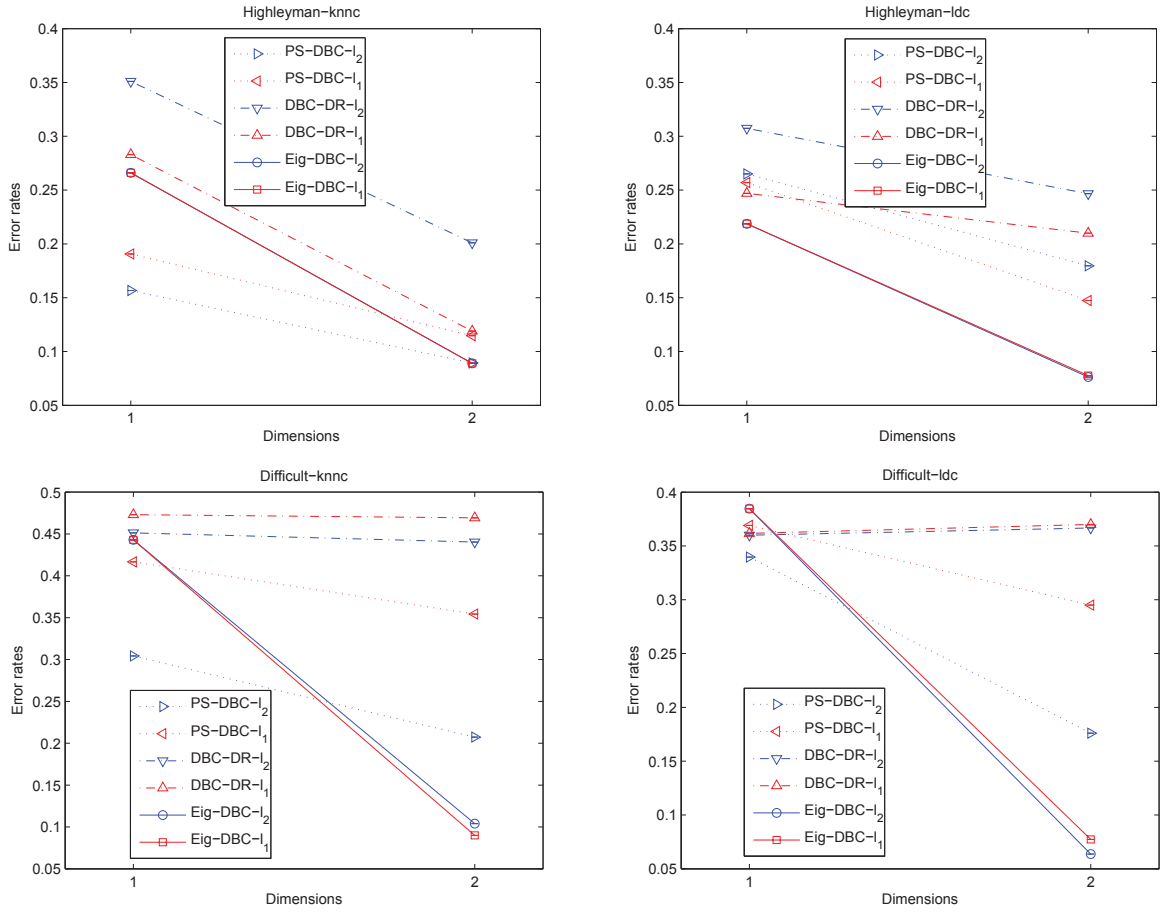


Fig. 1: Plots comparing the classification error rates estimated using *knnc* and *ldc* built in eigenspaces: (a) top and (b) bottom; (a) and (b) are obtained with Highleyman and Difficult Data using the l_2 and l_1 metrics.

Table 1: Characteristics of the UCI datasets used in the experiment.

datasets names	# of features	# of classes	# of objects	
			total #	object # per class (%)
Dermatology	33 (Categorical, Integer)	6	366	[30.60 16.67 19.67 13.39 14.21 5.46]
Diabetes	8 (Categorical, Integer)	2	768	[65.10 34.90]
Heart	6 (Categorical, Integer, Real)	2	297	[53.87 46.13]
Laryngeal1	16 (Integer, Real, etc.)	2	213	[38.03 61.97]
Liver	6 (Categorical, Integer, Real)	2	345	[42.03 57.97]
Sonar	60 (Integer, Real, etc.)	2	208	[46.63 53.37]
Wine	13 (Integer, Real)	3	178	[33.15 39.89 26.97]

Euclidean (l_2) distance. The proposed scheme was tested on two artificial datasets, and on some additional UCI datasets, and the results obtained were compared with those of other methods. Our experimental results demonstrate that the classification accuracy of DBC in Eigen spaces was improved when the dimensionality of the Eigen spaces had been appropriately chosen. In particular, the experimental results demonstrate that l_1-norm is justified in measuring the dissimilarity in the Eigen spaces. Although we have shown that the performance of DBC can be improved by employing the l_1-norm in Eigen spaces, many tasks remain unchallenged. One of them is to investigate further the hypothesis that improvement in the Eigen space DBC can be achieved only when using the Manhattan distance. Another area for future research is improvement in the classification efficiency by selecting an optimal, or nearly optimal, dimension of the

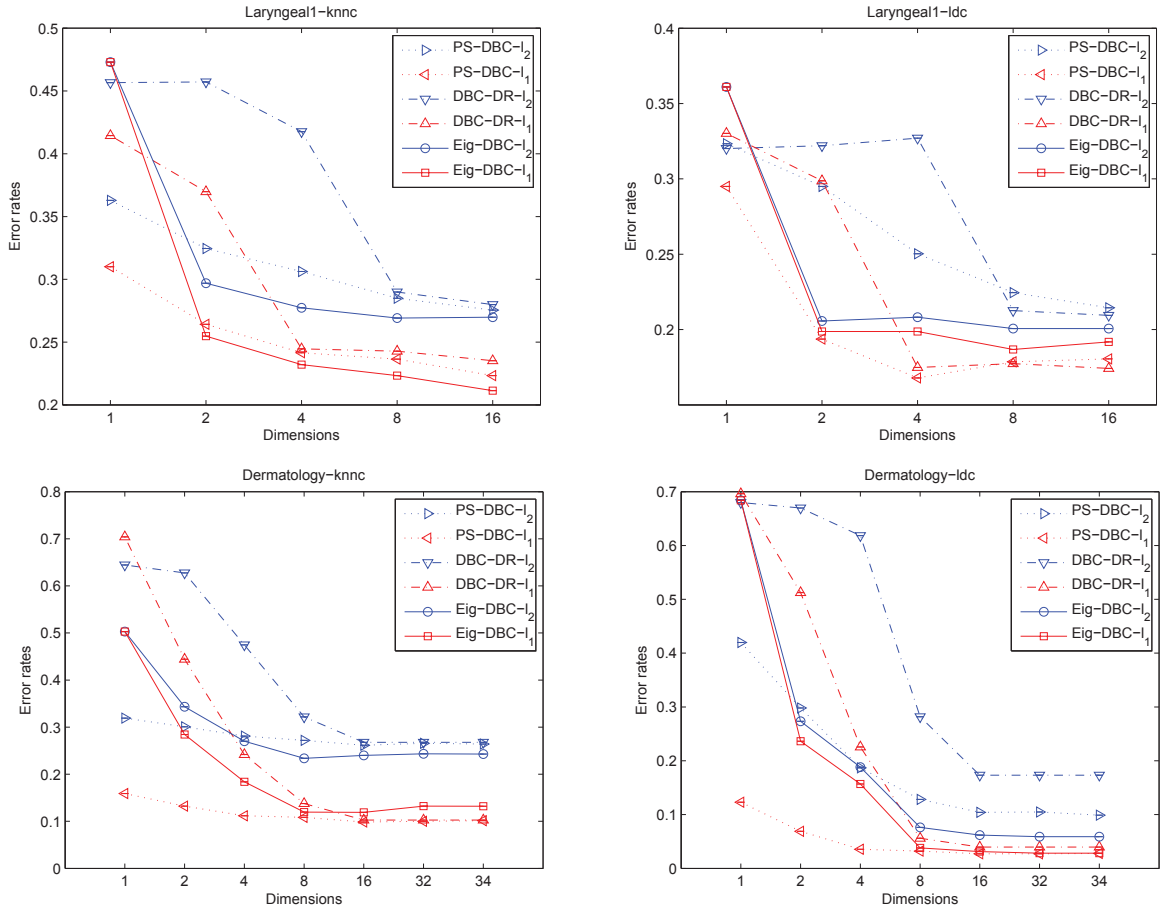


Fig. 2: Plots comparing the classification error rates estimated using *knnc* and *ldc* built in eigenspaces: (a) top and (b) bottom; (a) and (b) are obtained with Laryngeal1 and Dermatology using the l_2 and l_1 metrics.

Table 2: A numerical comparison of the mean error ($\pm$ std) rates(%) obtained with *knnc* and *ldc* for the five UCI datasets. For each row, the lowest one is highlighted with a * marker.

data sets	classifiers for DBC	PS-DBC		DBC-DR		Eig-DBC	
		l_2	l_1	l_2	l_1	l_2	l_1
Diabetes	*knnc*	34.38±2.42	33.25±2.66	34.72±2.93	33.07±2.40	34.20±3.24	*32.80±2.61
	ldc	25.82±2.72	23.80±2.67	25.99±2.85	23.70±3.05	23.33±2.72	*22.43±2.54
Heart	*knnc*	41.08±4.64	40.14±4.11	41.80±4.01	39.86±4.74	41.90±4.43	*38.87±4.19
	ldc	32.78±4.81	22.43±3.96	32.52±4.83	24.41±4.33	32.93±4.65	*21.40±3.66
Liver	*knnc*	43.22±4.11	42.02±5.30	44.42±4.56	42.36±4.40	43.53±4.44	*40.85±5.43
	ldc	33.45±4.04	32.60±3.43	36.51±4.27	36.63±4.72	*30.27±4.23	30.89±3.09
Sonar	*knnc*	20.13±4.69	19.67±4.32	20.59±5.49	*19.54±4.82	20.52±5.06	20.65±4.42
	ldc	17.25±3.69	16.60±3.84	17.32±3.82	*15.16±5.04	15.56±4.08	18.82±4.93
Wine	*knnc*	29.53±5.43	*23.26±5.86	29.61±6.14	25.12±6.62	29.30±5.91	25.50±7.00
	ldc	26.90±5.45	*16.59±5.91	27.13±5.67	23.02±6.78	27.52±5.08	17.75±6.27

Eigen spaces for the DBC. It is also not yet clear which kinds of significant datasets are more suitable for the scheme.

Acknowledgments

This work was supported by the National Research Foundation of Korea funded by the Korean Government (2012R1A1A2041661). The author is very grateful to Prof. Duin retired from the Delft University of Technology for the instructive discussions and his valuable comments.

References

[1] Pękalska, E. and Duin, R. P. W., *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*, Singapore: World Scientific Publishing, 2005.

[2] Duin, R. P. W., "Non-Euclidean problems in pattern recognition related to human expert knowledge," in *Proc. of the 12th International Conference on Enterprise Information Systems (ICEIS2010)*, Funchal, Madeira - Portugal, 2011, LNBIP-73, p. 15–28.

[3] Pękalska, E., Duin, R. P. W., and Paclík, P., "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition*, vol. 39, pp. 189–208, 2006.

[4] Kim, S. -W., "An empirical evaluation on dimensionality reduction schemes for dissimilarity-based classifications," *Pattern Recognition Letters*, vol. 32, no. 6, pp. 816-823, 2011.

[5] Riesen, K., Kilchherr, V., and Bunke, H., "Reducing the dimensionality of vector space embeddings of graphs," in *Proc. of the 5th International Conference on Machine Learning and Data Mining*, 2007, p. 563–573.

[6] Hu, Y. and Wang, Z., "A similarity measure based on Hausdorff distance for human face recognition," in *Proc. of the 2006 International Conference of Pattern Recognition (ICPR 2006)*, Hong Kong, 2006, ICPR-3, p. 1131–1134.

[7] Vivek, E.P. and Sudha, N., "Robust Hausdorff distance measure for face recognition," *Pattern Recognition*, vol. 40, pp. 431–442, 2007.

[8] Adini, Y., Moses, Y., and Ullman, S., "Face recognition: the problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 721–732, 1997.

[9] Yu, J., Amores, J., Sebe, N., Radeva, P., and Tian, Q., "Distance learning for similarity estimation," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 30, no. 3, pp. 451–462, 2008.

[10] Weinberger, K.Q. and Saul, L.K., "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[11] Duin, R. P. W., Bicego, M., Orozco-Alzate, M., Kim, S.-W., and Loog, M., "Metric learning in dissimilarity space for improved nearest neighbor performance," in *Proc. of the Joint IAPR International Workshop (S+SSPR2014)*, Joensuu, Finland, August 20-22, 2014, paper LNCS 8621, Published in: P. Franti, G. Brown, M. Loog, F. Escolano, M. Pelillo (Eds.) Springer, Heidelberg, p. 183-192.

[12] Oja, E., *Subspace Methods of Pattern Recognition*, England: Research Studies Press, 1983.

[13] Jolliffe, I. T., *Principle Component Analysis*, New York: Springer-Verlag, 2002.

[14] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J., "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[15] Ruiz-del-Solar, J. and Navarrete, P., "Eigenspace-based face recognition: a comparative study of different approaches," *IEEE Trans. Sys., Man, and Cybern.(C)*, vol. 35, no. 3, pp. 315–325, 2005.

[16] Chen, X., Flynn, P. J., and Bowyer, K. W., "PCA-based face recognition in infrared imagery: Baseline and comparative studies," in *Proc. of AMFG (the IEEE International Workshop on Analysis and Modeling of Faces and Gesture)*, Nice, France, 2003.

[17] Sadeghi, M., Samiei, M., and Kittler, J., "Fusion of PCA-based and LDA-based similarity measures for face verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 2010.

[18] Kim, S.-W. and Duin, R.P.W., "Dissimilarity-based classifications in eigenspaces," in *Proc. of the 16th Iberoamerican Congress on Pattern Recognition (CIARP2011)*, Pucón, Chile, 2011, paper LNCS-7042, p. 425–432.

[19] Kim, S.-W., "An empirical study on improving dissimilarity-based classifications using one-shot similarity measure," *Digital Signal Processing*, vol. 27, pp. 69–78, 2014.

[20] Duin, R. P. W., Juszczak, P., de Ridder, D., Paclík, P., Pękalska, E., and Tax, D. M. J. *PRTools 4: a Matlab Toolbox for Pattern Recognition*, Technical Report, Delft University of Technology, The Netherlands, 2004. [Online]. Available: http://prtools.org/

[21] Asuncion, A. and Newman, D. J.. (2007) UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

Automatic Vickers Microhardness Measurement based on Image Analysis

B. N. Coelho[2], A. Guarda[1], G. L. Faria[2], and D. Menotti[1]
[1]Computing Department, Federal University of Ouro Preto (UFOP), Ouro Preto, MG, Brazil
[2]Metalurgic and Material Engineering Department, UFOP, Ouro Preto, MG, Brazil

Abstract— *The Vickers Microhardness test consists of applying a force on the material surfaces and inducing a local plastic strain. By analyzing the dimentions of generated indentation in function of the applied load, the microhardness parameter can be determined. This test has been largely used for research development and quality control analyses due to its several applications. However, a problem associated with this technique is the influence of the human operator on the measures performed. Aiming to reduce, or even eliminate, the dispersion of results, this work proposes to design an image analysis-based method for automatic identification of sample surface printed and determination of its dimensions. The data used for its calibration is obtained from tests on samples of steel, cast iron, and aluminium and cooper alloys. The experimental results corroborates the method capability for performing the required measures with good precision and excellent reproducibility.*

Keywords: Vickers Microhardness, Pattern Recognition, Computer Vision, Automatic Microhardness Measurement

1. Introduction

The Vickers Microhardness test was initially proposed aiming to determine the hardness of metallic alloy phases or constituents applying the concept of minor scale penetration and of measuring the sample resistance to local plastic strain [1].

The ASTM E384 standard suggests that the Vickers indenter have to descend to the sample with an average rate between $15\mu m/s$ and $70\mu m/s$. The load application time must be between 10s and 15s, with an estimated tolerance around 2s [1], [2].

In test execution, when the indenter is removed from sample surface, it is highlighted a permanent mark on it, which was caused by the local plastic strain of the tested material. The microhardness value is defined as the rate between the applied load and the total area of pressed pyramid. The Vickers microhardness, studied in this paper, may be determined by

$$HV = 1.854.5 \times (F/d^2) \quad (1)$$

where F is the load (gf) and d is the average diagonal of the mark (μm) [1], [2], [3], [4].

A typical problem of this technique is the influence of several operational factors on result representativeness and reproducibility. Many authors studied some parameters and its influence on obtained results. The main studied parameters were: test machine calibration, vibration of test machine during the measurement execution, conservation conditions of ocular and objective lens, illumination conditions, inadequately load selection, and operator errors [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16].

Vander Voort [16], in 1989, studying the Vickers Microhardness (base of ASTM E384), have selected seven different samples and made an experimental plan using six different loads. The tests were executed in twenty four different laboratories. In each laboratory, only one operator did the measurements. In his work, he concluded that the diagonal measurement is the most influent error source (the smaller mark, the higher experimental error). He also concluded that the higher the applied load the smaller the experimental deviation. However, the average values obtained in each laboratory, using the same experimental parameters, were considerably different. Vander Voort explained that pointing the operator and machine calibration conditions as the main causes of these dispersions [16].

Image analysis and computater vision systems have been used aiming to the automation of several operational procedures. In many cases, these procedures, when executed manually, may have influence on final results. In the technical literature, there are several works with different methodologies to automatically measure microhardness [17], [18], [19], [20], [21]. In general, the image analysis systems starts with the indentation image digitalization, usually with cameras coupled to microscopes or to the own microhardness machine. In a next step, these images are pre-processed and any irregularities or defects are removed and the interest areas are highlighted and submitted to identification and calculus algorithms. In this work, the preliminary results of an implementation of our method with objective to measure Vickers microhardness automatically are presented.

2. Materials and Methods

2.1 Materials

We selected some samples of different kind of steels with the purpose of evaluating distinct materials in the Vickers

Fig. 1: Steps of an automatic microhardness measurement system

microhardness test and use them as initial data for building a data base for calibration of our implementation. The following materials were used: normalized eutectoid carbon steel, austenitic stainless steel LASER welded, ferritic stainless steel, and region of the molten zone (MZ) of ABNT[1] 1020 welded steel. Each material presented a peculiar behavior regarding both resistance and plastic strain promoted by the diamond indenter given the applied load. The images obtained from the micro-structure of these materials were also very different, allowing a broad calibration to our implementation. For the Vicker microhardness test, it was used a Pantec microhardness tester, in which loads from $10gf$ to $1000gf$ can be used. The Vickers microhardness values found and the image obtained by optical microscopy were organized in a data base for calibration.

The pattern recognition and image processing algorithms used were implemented using C++ language and Eclipse IDE, running Linux Ubuntu 11.04 operational system. The experiments were run on a Sony VAIO notebook, model VPCEB36GM – Intel Core i5, M460 2.53GHz, 4Gb RAM.

2.2 Structure and Description of the proposed method

Systems that use computer vision techniques, as in this case, may depend on several sequential stages in which the result of a previous step significantly influence the next stage, therefore each stage must be performed with quality, generating reliable results lest error propagation. Leta [22], in 2004, proposed Brinell and Vickers hardness automatic measurement systems with the implementation of several computer vision algorithms in four steps, as shown in Figure 1.

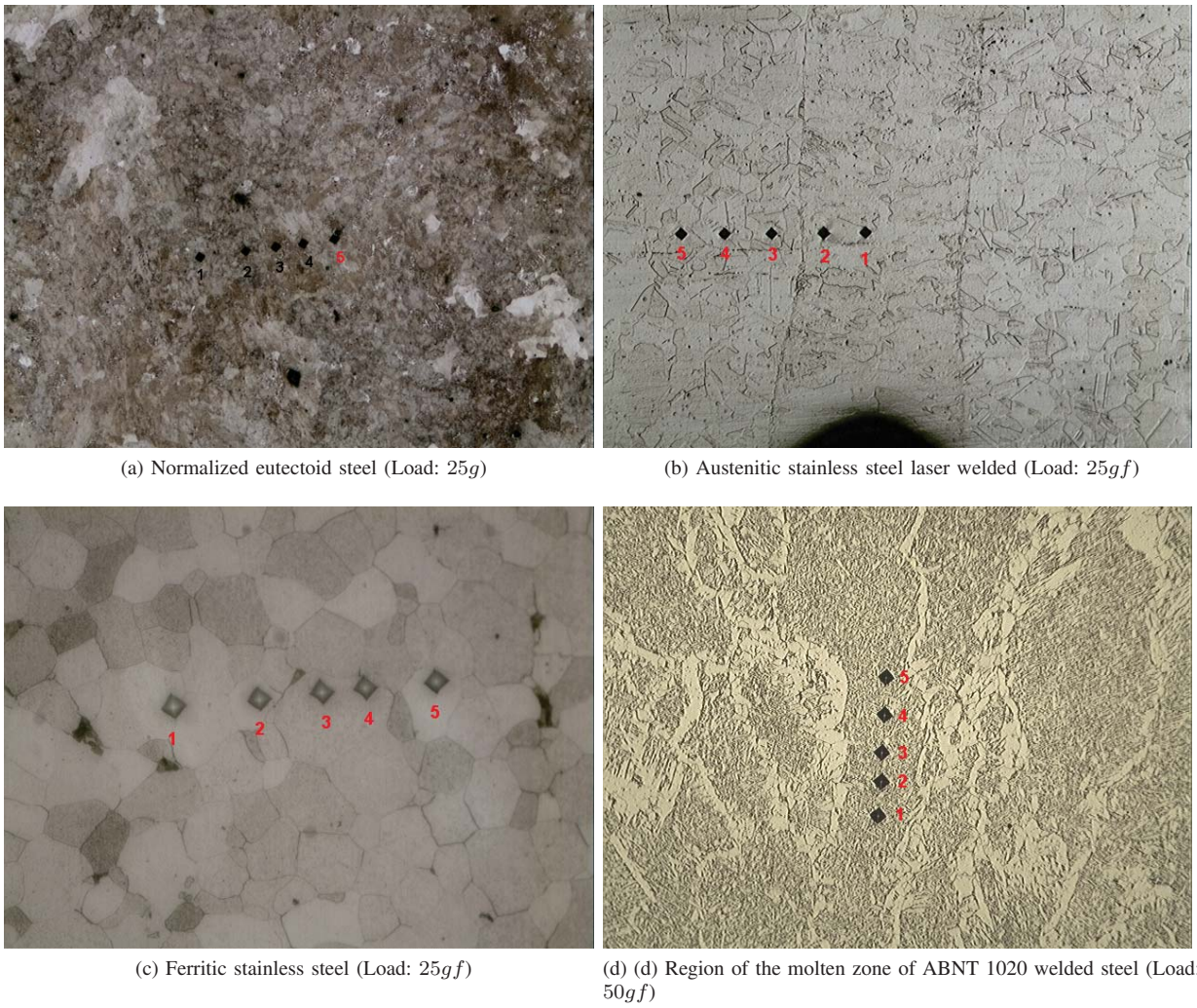
The proposed method was developed following these four steps, wherein the digitalization step is performed directly by the camera.

Before the other steps, the software starts requesting some basic information to operate effectively. This information depends solely on parameters used in microhardness tests, as well as the magnification used in the optical microscope. On the initial setup, we must provide the load value of hardness test (in grams), the exposure time of the indenter (in seconds), and the image magnification ($200\times$, $800\times$, etc.). The value of the load and time are needed to estimate the hardness value. The magnification value used on the image acquisition process is needed on the conversion of the distance in pixels to the indentation diagonals measures.

Then, in the second step, the software reads the images and begins the preprocessing stage in order to standardize the images due the wide diversity of materials that can be used. This is the most delicate stage, where it must be able to point out the indentations, isolating the regions which are not required for the measurement of microhardness, including the background image, which is formed by the material itself. This step is most critical in materials whose indentations are not very visible, or mingle with porosities, grain boundaries or inclusions. There are several preprocessing methods, each with their specific characteristics, adapting better to specific types of images. Among these methods are the application of filters and thresholds, the latter showing the best results for images in this work. The method involving thresholds divides the image histogram into two sections, ranging from 0 to 255. This causes the image to be transformed into black and white, which greatly facilitates the automatic recognition of certain geometries. In preliminary studies with various materials, the threshold that presented the best results was determined. The lower and upper limits that showed the best results were 40 and 100 respectively. This range of values proved the most suitable for the patterns of microhardness images. Within this range of values, are analyzed every possible threshold value, one by one, totalling 60 scans on the image. This preprocessing step is the most important because we should prepare the image in order to highlight the indentations, and suppress regions of little interest, as the background of the image.

In the third step, indentation recognition, it was used an object recognition algorithm, aiming at identifying only indentations, excluding other forms present in the figure. The image is decomposed in the three RGB components in the process of identifying and searching for polygons. Several scans are performed varying the threshold value, and only the first scan is accomplished differently, where is applied the Canny operator [23] and the dilation operator [24] with a 3×3 structuring element. The Canny operator is used to

[1]Brazilian Association for Technical Standards, in Portuguese - *Associação Brasileira de Normas Técnicas*.



(a) Normalized eutectoid steel (Load: $25g$)

(b) Austenitic stainless steel laser welded (Load: $25gf$)

(c) Ferritic stainless steel (Load: $25gf$)

(d) (d) Region of the molten zone of ABNT 1020 welded steel (Load: $50gf$)

Fig. 2: Vickers microhardness profile on. In all images, $200\times$ magnifier.

identify squares with gradient shading. The dilation operator eliminates potential gaps between the edges of the segments.

In each scan, all possible polygons that approach an indentation hardness are searched. The algorithm was initially calibrated with parameters to limit the maximum and minimum size of the polygons in order to minimize the number of possible indentations. For this purpose is used the average size of the perimeter and the area of âĂŃâĂŃindentations. Moreover, the polygons having the following properties are removed: the number of vertices is not four; non-convex; in $400\times$ magnification images obtained by optical microscopy, the area of âĂŃâĂŃa circle circumscribing the indentation must be between 15 and 200. In addition, the cosine of the angles are also evaluated, seeking to forms square like, accepting only small values âĂŃâĂŃfor the cosine.

For each scan involving a threshold, it can be found numerous objects that resemble the desired shape of the indentation. So, we need to select only those involving the prescribed parameters. The algorithm identifies, in all images, the regions with the greatest number of polygons found, and calculates the average position of each of the four vertices of these polygons (lateral edge of the indentation). In addition, it removes the overlaps between them. Thus, one obtains only one possible polygon for each region of the image representing each of the indentations.

Thus, the algorithm identifies each polygon, representing each indentation, and measures its diagonals, estimating the value HV microhardness of the material.

Finally, the algorithm generates an image superimposed on the original image by adding the edge contour of the indentation, showing the polygons found and the identification of each indentation. This identification is used to find the hardness value of that particular indentation in a table, which contains the information of the indentations (number,

HV value, dimension, XY matrix with the position of the vertices, etc.).

After the identification of the indentation, it is calculated the diagonals sizes. Obtaining the HV microhardness value of the material in the final stage is relatively simple because they are exact calculations using Equation 1. However, the result of this step depends entirely on the quality of the results obtained in the previous steps.

2.3 Experimental procedures

The Vickers microhardness test was performed according to ASTM E384 [1], applying the load during 10 seconds. It was used loads of $10gf$, $25gf$, and $50gf$, selected according to the kind of material to be analyzed. For each sample, it was performed five measurements. After the microhardness tests, images of micro-structures were acquired using optical microscopy using $200\times$, $400\times$, and $800\times$ magnifier highlighting the indentations, identified by numbers 1, 2, 3, 4, and 5 (Figures 2a, 2b, 2c, and 2d). 640×480 pixel images containing the indentations were acquired using a camera coupled to an optical microscope. For each one of these obtained images, the Vickers microhardness value was manually measured by a qualified and experienced professional, according to the standards.

The next step is to use the implementation of the proposed method for identifying of indentations and so the microhardness measurements calculations in an automatic way. The first user action to use the implementation is to setup the initial parameters, providing the load values used in the microhardness test, the magnification used for image acquisition besides the image to be processed. The load values is required to calculate the microhardness values, besides the magnification factor to image acquisition, to convert distance to pixels, when computing the diagonal indentations.

After initial parameters setup, the image preprocessing is computed and the microhardness values of indentations are found. One of the major advantages of an approach automated by computers is the processing speed that is this case takes only few milliseconds for indentation. In the experiments run here, we used images with one to five indentations. The user receives a table containing the HV microhardness values measured, plus a final draw by adding the edge contour of the indentation on the original image. Our implementation also allows some advanced configurations such as the modification of the bound of the indentation sizes and enable overlapping indentations such that better adjustments are made. Finally, the microhardness values measures for the traditional and manual method were compared to the values automatically measured by our implementation.

3. Results and Discussion

The implementation of the method developed is capable to work with distinct materials, presenting consistent results with Vickers microhardness values manually measured by an experienced human operator. The techniques and algorithms used in the implementation showed to be effective and robust in the hardness computation, and given the current development state of the implementation, the obtained results are very promising. Table 1 presents the loads (in gf) used for each sample, the Vickers microhardness values (HV) obtained manually and their respective values automatically measured by the implementation of the proposed method, for each one of the indentations.

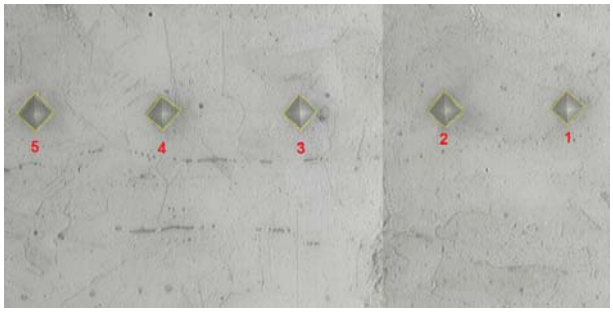
It is possible to note that some automatically measured Vickers microhardness values are presented as "-". These measurements were not performed due to several problems found by the implementation, which are described further. Figures 3a and 4a illustrates the best results obtained by the proposed method, while Figures 3b and 4b present comparison between the values automatically measured and the ones performed manually. For the austenitic stainless steel laser welded (see Figures 3a and 3b), the results are considered very promised, presenting values slightly greater than the manually performed measures, but following the same trend values. In Figures 4a and 4b, the results for the ferritic stainless steel are shown. It can be observed that the values are very close to the first three indentation measures, however more divergence values are presented in the last two measures. Both results are very promising and validate the proposed method, showing its potential for future developments.

It is possible to note that the value automatically measured in the fifth indentation of the ferritic stainless steel (Figura 4a) is quite greater than the value manually measured. It can be seen in this image that the edges of this fifth measurement, obtained from the automatic identification in yellow, are within the indentation, which results in a smaller average diagonal than the real value. A consequence of this poorly identification is the increase in the microhardness value measured. This is one of the problems found when using the method for automatic measurement and it should be correct in the step of indentation recognition, through the optimization of calibrated parameters or by using other techniques. Several defects were found in the indentation recognition and identification steps in the Vickers microhardness profiles, hindering the final value calculation. Some of these defects are presented in Figure 5a and Figure 6.

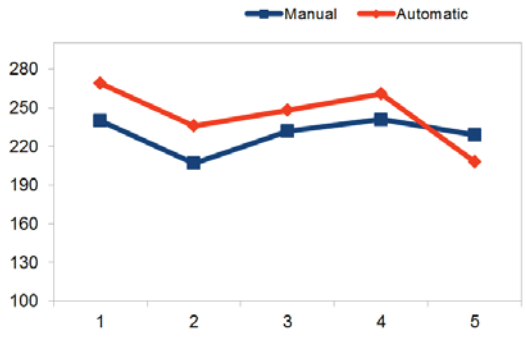
In Figure 5a, the indentation number one was no detected by the method because the presence of a dark color constituent (perlite) close to its edge. The presence of such noise made difficult the step of identifying indentations. By observing the indentations numbers 2, 3, and 5 in Figure 5a, irregularities can be seen between the edge detected and the real one, which causes divergence in the values measured

Table 1: Applied loads (gf) and results of manual and automatic Vickers microhardness (HV) measurements.

Materials	Load (gf)	Microhardness Vickers (HV) Manual	Microhardness Vickers (HV) Automatic	Materials	Load (gf)	Microhardness Vickers (HV) Manual	Microhardness Vickers (HV) Automatic
Austenitic stainless steel laser welded	25	1 - 240 2 - 207 3 - 232 4 - 241 5 - 229	1 - 269 2 - 236 3 - 248 4 - 261 5 - 208	Ferritic stainless steel	25	1 - 208 2 - 212 3 - 199 4 - 206 5 - 200	1 - 214 2 - 227 3 - 223 4 - 269 5 - 282
Normalized eutectoid steel	25	1 - 333 2 - 425 3 - 338 4 - 332 5 - 355	- 2 - 971 3 - 779 4 - 982 5 - 867	Region of the molten zone of ABNT 1020 welded steel	25	1 - 227 2 - 212 3 - 222 4 - 220 5 - 225	- 2 - 220 3 - 191 4 - 214 5 - 200

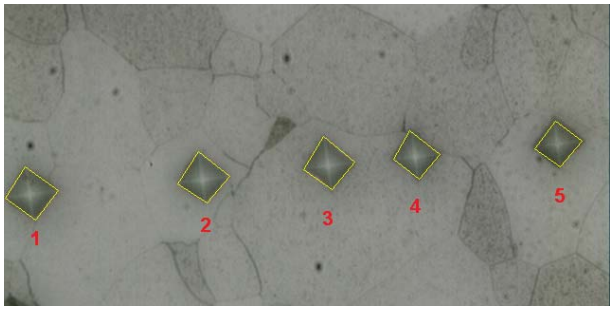


(a) Indentations automatically identified - 800× magnifier

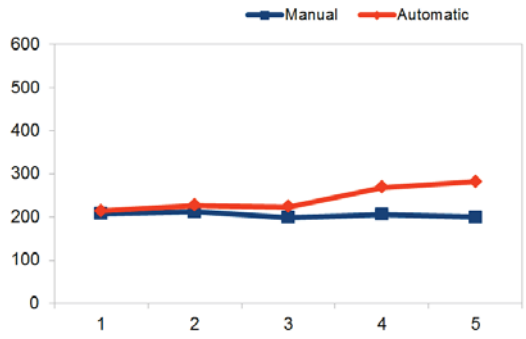


(b) Comparing Vickers microhardness values manually and automatically measured

Fig. 3: Experimental results for the austenitic stainless steel laser welded.

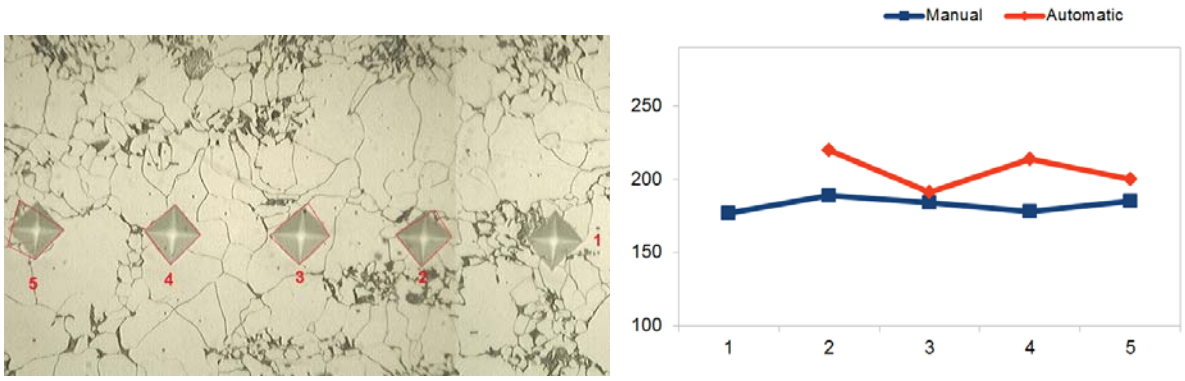


(a) Indentations automatically identified - 800× magnifier



(b) Comparing Vickers microhardness values manually and automatically measured

Fig. 4: Experimental results for the ferritic stainless steel.



(a) Defects on the identification of the indentations - 800× magnifier (b) Comparing Vickers microhardness values manually and automatically measured

Fig. 5: Experimental results for the MZ of ABNT 1020 welded steel.

Fig. 6: Defects of identification of indentation on eutectoid steel. 800× magnifier.

of microhardness. In Figure 6 (eutectoid steel), the values measured were about three times greater than the real values. It can be seen that only the center of each indentation was correctly identified, which caused a sharp divergence of values. Moreover, it was detected a false indentation (over the fifth one), with shape similar to a real indentation. It is important to report that the eutectoid steel, which is the material with the greatest hardness in this study, was the one with the greatest imprecision in the values of microhardness measures. This fact was previously mentioned by some researchers [1], [2], [3], [4], [5], [14], [15], [16]. This result was exacerbated because its micro-structure have very complex optical contrast mechanisms and this characteristic hinders the automatic identification of the indentations.

The irregularities at the edge of indentations contribute to errors in the identification process, causing divergence in the performed measures, because the different size of indentation edges, which, in turn, can vary to larger or smaller. This changes at the identification of edges can occur because of the preprocessing techniques used. Although they enhance the image making possible its analysis by enhancing the indentations, the process mainly affect the assessment of the perimeter of the indentation edges [22]. We believe that these problems can be minimize by using machine learning techniques.

4. Conclusions

The results obtained are very promising, and have corroborated the potential of the image analysis-based method, which is still in development. The use of the implementation for the automatic microhardness measurement provides an increase in speed of the process. Besides it also provides standardized and reproducible results rather than the manual measurement which is time consuming and affected by human factors.

The best results were obtained for the austenitic stainless steel laser welded and the ferritic stainless steel, where the measurements made by the method were similar to the values obtained manually. The worst case was for the eutectoid iron, in which no correct indentation was identified.

The implementation of the method was performed using free software, which facilitates its future development and extensions in partnership with other research institutions. For future works, the automation of the equipment physic phase is on focus, including the sample movement and positioning, the indentation task and the image digitalization.

Regarding the software obtained from the implementation of the method, it is still required to refinements before its conclusions. More specifically, to improve the calibration steps and to study new machine learning methods, besides the development of a friendly user interface.

5. Acknowledgements

The authors would like to thank UFOP, REDEMAT, FAPEMIG, CAPES, and CNPq for the financial support. The authors would like also to thank especially the Heat Treatment and Optical Microscopy Laboratory (LTM), as well the MSc. Paulo Sérgio Moreira for the collaborations in manual microhardness measurements.

References

[1] A. S. for Testing and M. (ASTM), "Astm e384 - standard test method for microindentantion hardness of material," p. 37, 2009.

[2] R. F. Campbell *et al.*, "A new design of microhardness tester and some factors affecting the diamond pyramid hardness number at light loads," *Transactions of the American Society for Metals*, vol. 40, no. 1, pp. 954–982, 1948.

[3] R. G. Kennedy and N. W. Marrotte, "The effect of vibration on microhardness testing," *Materials Research and Standards*, vol. 9, pp. 18–23, 1969.

[4] A. R. G. Brown and E. Ineson, "Experimental survey of low-load hardness testing instruments," *Journal of the Iron and Steel Inst.*, vol. 169, pp. 376–388, 1951.

[5] M. Factor and I. Roman, "Vickers microindentation of wc?12coating: Part 1: statistical analysis of microhardness data," *Surface and Coatings Technology*, vol. 132, no. 2–3, pp. 181–193, 2000.

[6] S. Igarashi, A. Bentur, and S. Mindess, "Microhardness testing of cementitious materials," *Advanced Cement Based Materials*, vol. 4, no. 2, pp. 48 – 57, 1996.

[7] D. Ye and Z. Wang, "An approach to investigate pre-nucleation fatigue damage of cyclically loaded metals using vickers microhardness tests," *International Journal of Fatigue*, vol. 23, no. 1, pp. 85 – 91, 2001.

[8] D. Ye, X. Tong, L. Yao, and X. Yin, "Fatigue hardening/softening behaviour investigated through vickers microhardness measurement during high-cycle fatigue," *Materials Chemistry and Physics*, vol. 56, no. 3, pp. 199–204, 1998.

[9] D. Ye, D. Wang, and P. An, "Characteristics of the change in the surface microhardness during high cycle fatigue damage," *Materials Chemistry and Physics*, vol. 44, no. 2, pp. 179–181, 1996.

[10] C. Zeng, W. Tian, W.-H. Liao, and L. Hua, "Study of laser cladding thermal damage: A quantified microhardness method," *Surface and Coatings Technology*, vol. 236, no. 0, pp. 309 – 314, 2013.

[11] C. A. Della Rovere, F. S. Santos, R. Silva, C. A. C. Souza, and S. E. Kuri, "Influence of long-term low-temperature aging on the microhardness and corrosion properties of duplex stainless steel," *Corrosion Science*, vol. 68, no. 0, pp. 84–90, 2013.

[12] P. Stella, I. Giovanetti, G. Masi, M. Leoni, and A. Molinari, "Microstructure and microhardness of heat-treated ti?6al?2sn?4zr?6mo alloy," *Journal of Alloys and Compounds*, vol. 567, no. 0, pp. 134–140, 2013.

[13] N. W. Thibault and H. L. Nyquist, "The measured knoop hardness of hard substances and factors affecting its determination," *Transactions of the American Society for Metals*, vol. 38, pp. 271–330, 1947.

[14] L. P. Tarasov and N. W. Thibault, "Determination of knoop hardness numbers independent of load," *Transactions of the American Society for Metals*, vol. 38, pp. 331–353, 1947.

[15] G. F. Vander Voort, "Results of an astm e04 round robin on the precision and bias of measurements of microindentation hardness. factors that affect the precision of mechanical tests," *ASTM STP*, vol. 1025, pp. 3–39, 1989.

[16] ——, "Operator errors in the measurement of microindentation hardness. accreditation practices for inspections, tests, and laboratories," *ASTM STP*, vol. 1057, pp. 47–77, 1989.

[17] P. R. Rebouças Filho, T. S. Cavalcante, V. H. C. Albuquerque, and J. M. R. S. Tavares, "Brinell and vickers hardness measurement using image processing and analysis techniques," *Journal of Testing and Evaluation*, vol. 38, no. 1, pp. 1–7, 2010.

[18] M. Hruz, J. Siroky, and D. Manas, "Robust image processing technique for knoop hardness measurement," in *17th Symposium IMEKO TC 4, 3rd Symposium IMEKO TC 19, and 15th IWADC Workshop*, 2010, pp. 146–151.

[19] A. Maier, G. Niederbrucker, S. Stenger, and A. Uhl, "Efficient focus assessment for a computer vision-based vickers hardness measurement system," *Journal of Electronic Imaging*, vol. 21, no. 2, pp. 1–14, 2012.

[20] L. Yao and C.-H. Fang, "A hardness measuring method based on hough fuzzy vertex detection algorithm," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 3, pp. 950–962, 2006.

[21] P. Yong, S. Yuekang, J. Yu, and Z. Shibo, "A new method for automatically measuring vickers hardness based on region-point detection algorithm," in *4th International Symposium on Precision Mechanical Measurements (SPIE)*, vol. 7130, 2008, pp. 1–6.

[22] F. R. Leta, V. B. Mendes, and J. C. S. D. Mello, "Medição de identações de dureza com algoritmos de visão computacional e técnicas de decisão com incertezas," *ENGEVISTA*, vol. 6, no. 23, pp. 15–35, 2004.

[23] J. A. Canny, "Computational approach for edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[24] E. Dougherty, "An introduction to morphological image processing," in *Tutorial Textes in Optical Engineering*. SPIE-International Society for Optical Engine, 1992, vol. TT9, pp. 1–162.

A Threshold-based Pattern Recognition to Filter Noise on an Embedded On-cell Touch Panel

Yen-Ting Chen*, Tseng-Yi Chen*, Heng-Yin Chen†, Wei-Kuan Shih*
*Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
†Display Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan

***Abstract*—[1] On-cell touch is a kind of a touchscreen technology which reduces the thickness of display. Based on the architecture of on-cell touchscreen technology, cover glass structures is integrated with touch sensors. Although a one-glass solution can successfully reduces the thickness of display, the sensitivity of touch detection is also decreased because of the integration of cover glass structures and touch sensors. Recently, most solutions increase the sensitivity of touch detection by built-in extra hardware devices such as controller and display-driver IC. However, the extra devices also results in higher cost of production. To resolve this issue, this study proposes an algorithm called Threshold-based Pattern Recognition (TPR), which filters touch sensor noises on a on-cell touch panel. For system demonstration, we integrate the proposed algorithm into Linux driver on the Friendly-ARM 210 embedded platform. As the demonstration shows, the proposed algorithm can correctly reflect the coordinates of the on-cell touch panel on the display.**

Keywords*—*Pattern recognition; on-cell touch; embedded system; sensor noise

I. Introduction

Generally, touchscreen technologies can be classified into three categories: One Glass Solution (OGS), On-Cell touch, and In-Cell touch. Because of the difficulties in in-cell touch-screen implementation, on-cell touch become the most popular technology used in display productions. In on-cell touch architecture, the projected capacitive (PCAP) touch sensor is directly attached to the glass front of the display, and the touch-screen controller is also fully integrated in the display module. Due to the architecture of on-cell touch technology, the sensitivity of touch sensors is decreased, as shown in Figure 1.

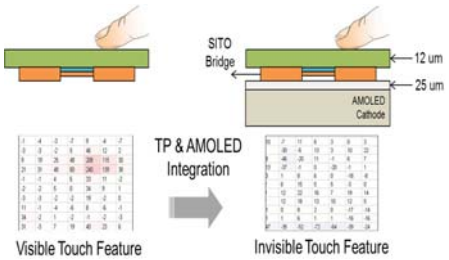


Fig. 1: A problem of on-cell touch.

To resolve the issue, many excellent prior works have been proposed for increasing the sensitivity of touch sensors on on-cell touch panel. For instances, Yang et al. [1] proposed a control IC, which adopts the differential sensing method to enhance the dynamic range of sensing voltage and to be robust to display noise; Kim et al. [2] presented a display-driver IC embedding a capacitive-touch-screen controller, and Kim et al. [3] showed ultrathin and highly sensitive input/output devices consisting of a capacitive touch sensor (Cap-TSP) integrated on thin-film-encapsulated active-matrix organic light-emitting diodes (OLEDs). Although the above excellent works have improved the sensitivity of on-cell touch panels, the cost of touchscreen panel also has been raised because of the extra IC/hardware controller. Reducing the cost of touchscreen panel, this study proposes a threshold-based filtering algorithm to filter the noises on on-cell touchscreen. The proposed algorithm is integrated into the on-cell touchscreen's driver in Linux. For system verification, we deploy the proposed algorithm in the Android system, which runs on a embedded platform called Friendly ARM Tiny-210.

The remainder of the paper are as organized as follows. Section II describes system architecture and detail implementation of threshold-based pattern recognition algorithm. In section III, the verification result of the proposed algorithm is presented. Sections IV concludes the paper and lists our future work.

II. Implementation of Threshold-based Algorithm

A. System Architecture

The implementation of the proposed algorithm is based on a standard Android system. So, the system architecture is a standard development flow on Android system. First of all, we develop an I^2C driver in Linux kernel. To verify the correctness of the I^2C driver, we also develop a application, which can report the touched coordinates on the experimental display. Additionally, in order to connect the application and driver, a Java native interface (JNI) has been developed. In the JNI library, the raw data from I^2C driver can be passed to the developed application, and then the developed application can process the information from I^2C driver to show on the screen. Figure 2 shows the system architecture of the I^2C driver implementation.

In this system architecture, we implement the proposed algorithm in the I^2C driver because all raw data are received in I^2C driver at first stage. After receiving the raw data from the on-cell touchscreen in the I^2C driver, the procedure of the threshold-based algorithm will be launched to process the raw data for delivering the correctness position to the developed application. The details of the proposed algorithm is presented in the next subsection.

[1] This paper is being submitted as a poster.

Fig. 2: The system architecture of implementation.

B. Threshold-based Algorithm

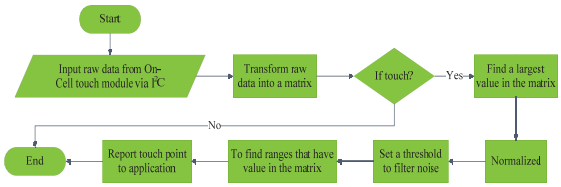


Fig. 3: The flow chart of the proposed algorithm.

For the brevity, we use a flow chart to explain the procedure of the threshold-based algorithm. Figure 3 shows the process of the threshold-based algorithm. First of all, the proposed algorithm receives the raw data form on-cell touchscreen via the developed I^2C driver, and then the algorithm transform the raw data into a matrix. To detect a touch behavior, the proposed algorithm will real-time detect the changes in the matrix. If the values in the matrix has been changed, the algorithm launches a touched event. In the touch event, the proposed algorithm finds a largest value in the matrix and sets the largest number as upper bound for next process. When the algorithm finds the largest value in the matrix, the algorithm uses the value to normalize the rest of values in the matrix. Then, a threshold will be set in the normalized matrix for filtering the noise on the on-cell touchscreen. After that, the proposed algorithm applies a pattern recognition algorithm to find the boundary of the touched area on the on-cell touchscreen. Finally, the proposed algorithm will report the touched point to the developed application.

III. Demonstrations

A. Experimental Environment

To verify the correctness of the proposed algorithm, we implemented the proposed algorithm on a Friendly ARM Tiny-210 embedded platform. The experimental platform ran Android system. The experimental platform has a 1 GHz Samsung S5PV210 ARM Cortex-A8 processor, 512 MB RAM, 1 GB NAND Flash, and 256 Byte EEPROM. In the experimental platform, we integrated the proposed algorithm in the I^2C driver and developed an application for verification.

B. Demonstration Results

Figure 4 shows the demonstration result of double-points touch on the experimental on-cell touch panel. Figure 5 shows the demonstration result of five-points touch. In Figure 4 and 5, the application shows the correctness coordinates on display when the user touch the on-cell touch panel.

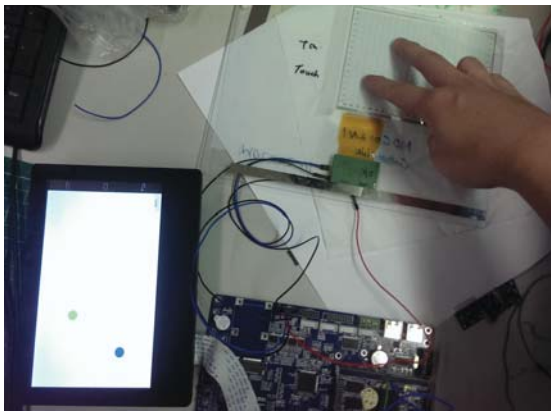


Fig. 4: The demonstration of double-points touch.

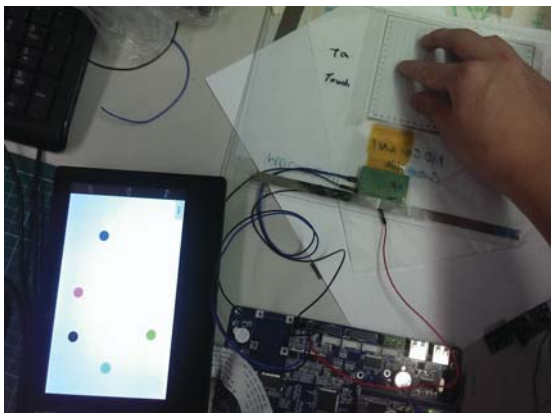


Fig. 5: The demonstration of five-points touch.

IV. Conclusion and Future Work

In this paper, we propose a threshold-based algorithm to filter the noise on an on-cell touch panel. In the demonstration, the results show that the proposed algorithm can correctly map the touched points on the on-cell touch panel to a display. In the future, we will integrated our algorithm into different hardware platforms and operating systems.

References

[1] I.S. Yang and O.K. Kwon, "A touch controller using differential sensing method for on-cell capacitive touch screen panel systems," in *IEEE Transactions on Consumer Electronics*, vol. 57, issue 3, pp. 1027-1032, August 2011.

[2] H.r. Kim, Y.K Choi, S.H. Byun, S.W. Kim, K.H. Choi, H.Y. Ahn, J.K. Park, D.Y. Lee, Z.Y. Wu, H.D. Kwon, Y.Y. Choi, C.J. Lee, H.H. Cho, J.S. Yu, and M. Lee, "A mobile-display-driver IC embedding a capacitive-touch-screen controller system," in *2010 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 114-115, San Francisco, CA, USA 7-11 February 2010.

[3] S. Kim, W. Choi, W. Rim, Y. Chun, H. Shim, H. Kwon, J. Kim, I. Kee, S. Kim, S. Lee, and J. Park, "A Highly Sensitive Capacitive Touch Sensor Integrated on a Thin-Film-Encapsulated Active-Matrix OLED for Ultrathin Displays," in *IEEE Transactions on Consumer Electronics*, vol. 58, issue 10, pp. 3609-3615, October 2011.

Classification Using Angle and Radius of Feature Vector

Z. Iscan
Centre for Cognition and Decision Making, National Research University Higher School of Economics, Russian Federation

Abstract - *In this paper, use of angle and radius information for feature space classification is proposed. The performance of the classification using either angle or the radius was evaluated on two different feature spaces for three and four-class classification problems. The results were compared with the well-known K-Nearest Neighbor (K-NN) and Naïve Bayes (NB) algorithms in terms of the ability to classify the feature space and classification time. Results show that angle and radius-based classification could generate better classification performances, especially when there are few training vectors available. Moreover, proposed methods were computationally more efficient than K-NN and NB algorithms. However, optimum combination of angle and radius-based classification is needed for developing a general classifier which will perform well in classification of different feature patterns*.

Keywords: angle; radius; classification; k-nearest neighbor; naïve bayes.

1 Introduction

Improvements in digital technology increased the amount and size of images acquired in different fields [1]. Images are composed of different meaningful patterns and pattern classification is a way of understanding and categorizing the different meaningful content embedded in the images. Therefore, image processing and pattern recognition are very closely related disciplines [2].

Although human beings have very advanced cognition capabilities [3], huge amount of data should be mined in order to reach more concrete and abstract information. Therefore, automation of this process is needed and different algorithms have been proposed in literature to classify different patterns in images [4].

Pattern classification algorithms use features that represent the patterns' discriminative information. If the features are good enough to separate different classes, the classification algorithm does not need to be complex. However in practice, features of different classes overlap [5]. In this case, classification algorithm should decide on how to determine the boundary of separation between different classes [6]. In this context, the classifier's ability to partition the feature space becomes important. This partition is depending on the kind of distance metric (e.g. Euclidean, Manhattan, ellipsoid) [7] and the location of initial nodes (i.e. reference points).

This study was inspired by spider web, where the edges of web can be defined by finite set of angles and radii. Using these sets, web area can be mapped. Therefore, angle and radius based separations of the feature space were proposed and the classification performances of these approaches were compared with a widely used instance-based algorithm (K-Nearest Neighbor (K-NN)) [8] and a probabilistic algorithm (Naïve Bayes [9]).

In section 2, the feature space and the methods used for classification are explained. Results are given in section 3. The discussion of the results and the conclusions are presented in section 4 and section 5, respectively.

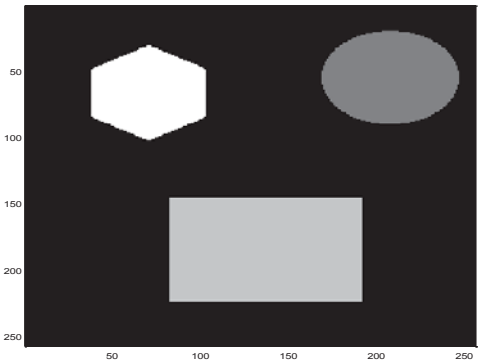
2 Methods

2.1 Dataset

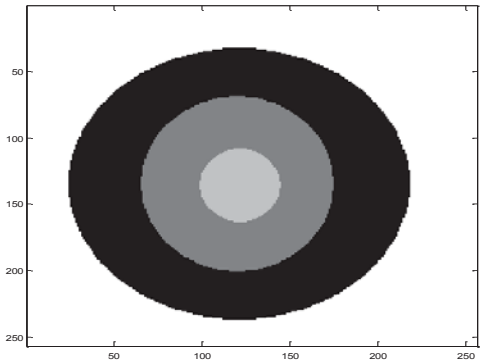
In order to present the proposed methods, two gray-level images with three and four-class problems were generated in 256 × 256 pixel sizes. In Fig. 1, these two images correspond to feature spaces are presented. In Fig. 1a, black regions do not belong to any class. In this case, there are three classes. However, in Fig. 1b, all colors (including white) indicate a different class. Therefore, there are four classes. For each generated image, two sets of training feature vectors (10 from each class, 100 from each class) were generated by selection of x and y coordinates of the image (See the colorful points of different classes in Figs. 3a-6a). Therefore, training feature vectors are composed of two features which correspond to the x and y coordinates of the image.

Before the calculation of the angle or radius of a given training vector (i.e. point on the image), the average of all training vectors (TR) were calculated to be the center (C_x, C_y) of the spider web (denoted by cyan cross on Figs. 3a-6a) as below.

$$C_x = \frac{\sum_{i=1}^{N} TR_x(i)}{N}, \quad C_y = \frac{\sum_{i=1}^{N} TR_y(i)}{N} \qquad (1)$$



(a) Three class problem



(b) Four class problem

Figure 1. Images for three (a) and four-class (b) problems.

In (1), TR_x and TR_y are the x and y coordinates of the given training vectors. *N* is the total number of the training vectors. After finding the center of the training vectors, depending on the image, angle or radius-based approach is followed. In Fig. 2, the angle (α) and the radius (r) calculations are depicted. Radius is the Euclidean distance between of the center of the spider web and the related training point as given in (2).

$$r = \sqrt{(y_1 - C_y)^2 + (x_1 - C_x)^2} \tag{2}$$

The angle can be calculated with the inverse trigonometric function, arctangent.

$$\alpha = \arctan\left(\frac{y_1 - C_y}{x_1 - C_x}\right) \tag{3}$$

2.2 Angle-based classification

In the angle-based classification, first of all, the angles of all training vectors are calculated using the equation 3. Afterwards, the angle intervals are extracted using the algorithm below, which sorts the angles from the minimum to the maximum and extracts intervals for the specific classes.

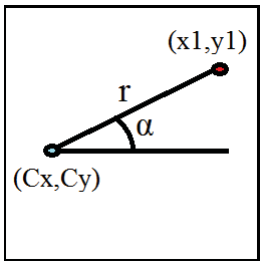


Figure 2. Angle (α) and radius (r) of the point (x1, y1).

Algorithm for extracting intervals

Sort the training vectors according to the angle/radius (min to max value)
Define the initial interval that corresponds to the minimum angle/radius and define its class
***For** all sorted vectors*
***If** current and next vectors belong to the same class:*
- *Set the current interval's max as the next vector's angle/radius*

else
- *Update the current interval's max to be the average of the current interval's max and the next vector's angle/radius*
- *Create a new interval*
- *Set the min of the new interval to be the max of current interval*
- *Set the max of the new interval to be the min of the new interval*
- *Assign the next vector's class to the new interval's class*

After extracting the intervals using the algorithm, there should be a correction for the last angle < 360˚ and the first angle > 0˚ due to the circular intersection of the angles. This correction is given below.

Correction Algorithm for the last angle<360˚ and the first angle>0˚

***If** the last vector before 360˚ and the first vector after 0˚ belong to the same class:*
- *Extend the last interval's max to 359.99˚*
- *Make a new interval (min=0˚, max=first angle) with same class*
- *Add the new interval to the top of the sorted intervals*

else
- *Calculate the circular midpoint of the last vector before 360˚ and the first vector after 0˚*
- ***If** the midpoint angle is > 0˚*
  - *Update the first interval's min to be the midpoint angle.*
  - *Extend the last interval's max to 359.99˚*
  - *Make a new interval (min=0˚,max=midpoint angle) with same class of the last interval*
  - *Add the new interval to the top of the sorted intervals*
  - ***else***
  - *Update the last interval's max to be the midpoint angle*
  - *Make a new interval (min=midpoint angle, max=359.99˚) with same class of the first interval*
  - *Add the new interval to the bottom of the sorted intervals*
  - *Make a new interval (min=0˚, max=first angle) with same class of the first interval*
- *Add the new interval to the top of the sorted intervals*

Once the interval extraction is accomplished, classes are defined for each interval. Therefore, a test vector's class can be determined by calculation of its angle and finding the corresponding interval's class.

2.3 Radius-based classification

In the radius-based classification, first of all, the radii of all training vectors are calculated. Afterwards, the radii intervals are extracted using the same algorithm given in section 2.2 (without correction part). Similarly, a test vector's class can be found by calculation of its radius and finding the corresponding interval's class.

2.4 K-Nearest Neighbor algorithm

K-Nearest Neighbor is one of the most commonly used algorithms in pattern classification [10]. Although it is not a computationally efficient algorithm, its simple coding and high performance are desirable in many applications. There is no training procedure for this instance-based algorithm. A new vector's class is determined by checking the nearest training vectors' class. K is the number of the closest training vectors. In this paper, K=1 which means that test vector's class will be the closest training vector's class.

2.5 Naïve Bayes algorithm

Naïve Bayes (NB) is also one of the most commonly used algorithms in pattern classification [10]. It is based on the Bayes Theorem, which links the posterior probability with the prior probability and likelihood. NB assumes that the features are independent [9] and it assigns the test sample to the most probable class.

3 Results

All classifications were performed on Matlab Software (MATLAB and Statistics Toolbox Release 2009b, The MathWorks, Inc., Natick, Massachusetts, United States).

In Figs. 3-4, angle (b), K-NN (c), and NB (d) based classification results are given for three class case when the number of training vectors (Ntr) were 30 (10 from each class) and 300 (100 from each class) respectively. Similarly, in Figs. 5-6, radius (b), K-NN (c), and NB (d) based classification results are given for four-class case when the number of training vectors were 40 (10 from each class) and 400 (100 from each class) respectively.

In Table I, percent differences between the angle vs. K-NN-based and angle vs. NB-based classification results are given for the three-class case when Ntr=30 and Ntr=300 respectively (See Figs. 3-4).

TABLE I. PERCENT DIFFERENCE BETWEEN ANGLE-BASED AND K-NN / NB-BASED CLASSIFICATION RESULTS.

Ntr	Comparison	Percent (%) difference
30	K-NN vs angle	1.05
	NB vs angle	9.25
300	K-NN vs angle	1.29
	NB vs angle	3.07

In Table II, percent difference between the radius vs. K-NN-based and radius vs. NB-based results are given for the four-class case when Ntr=40 and Ntr=400 respectively (See Figures 5-6).

TABLE II. PERCENT DIFFERENCE BETWEEN RADIUS-BASED AND K-NN / NB-BASED CLASSIFICATION RESULTS.

Ntr	Comparison	Percent (%) difference
40	K-NN vs radius	17.85
	NB vs radius	53.50
400	K-NN vs radius	13.15
	NB vs radius	29.54

In Table III, classification times (in terms of training and test time) for three-class and four-class images are given.

TABLE III. CLASSIFICATION TIMES FOR THREE-CLASS AND FOUR-CLASS IMAGES.

Classes	Ntr	Classifier	Train (s)	Test (s)
3	30	Angle-based classifier	0.06	0.25
		K-Nearest neighbor	-	10.53
		Naïve Bayes	0.38	368.88
	300	Angle-based classifier	0.08	0.23
		K-Nearest neighbor	-	57.09
		Naïve Bayes	0.66	373.05
4	40	Radius-based classifier	0.06	0.44
		K-Nearest neighbor	-	12.18
		Naïve Bayes	0.38	504.67
	400	Radius-based classifier	0.06	0.84
		K-Nearest neighbor	-	74.36
		Naïve Bayes	0.40	508.60

4 Discussion

Here, angle-based three-class and radius-based four class results were discussed separately.

4.1 Classification on the Three-class Images

In Figs. 3-4, angle (b) and K-NN-based (c) approaches generated very similar results. In Table I, the difference between angle-based and K-NN based classification of feature space was very small and did not change very much depending on the number of training vectors. This can be visually validated in Figs. 3-4. However, angle (b) and NB-based (d) classification results were apparently different from each other. It was observed that NB results were sensitive to the number of training vectors and they became better when Ntr was bigger. It should be noted that the classification problem given here is suitable for classification by angle.

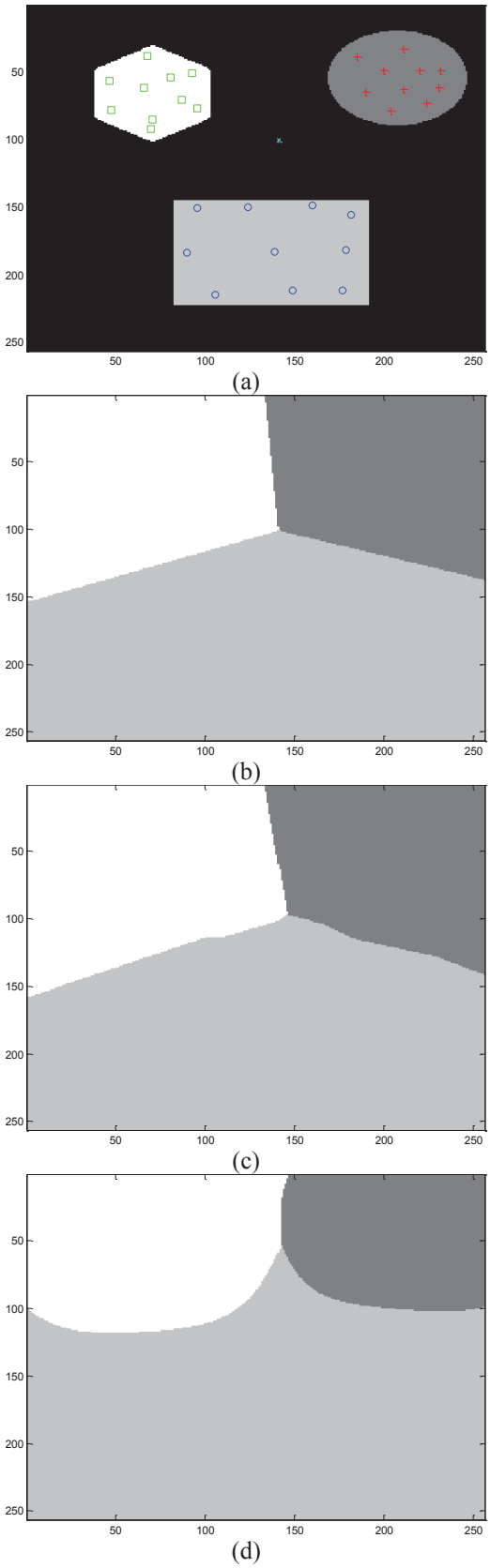


Figure 3. a) Three classes b) Angle-based, c) K-NN-based d) Naïve Bayes-based classification, Ntr=30.

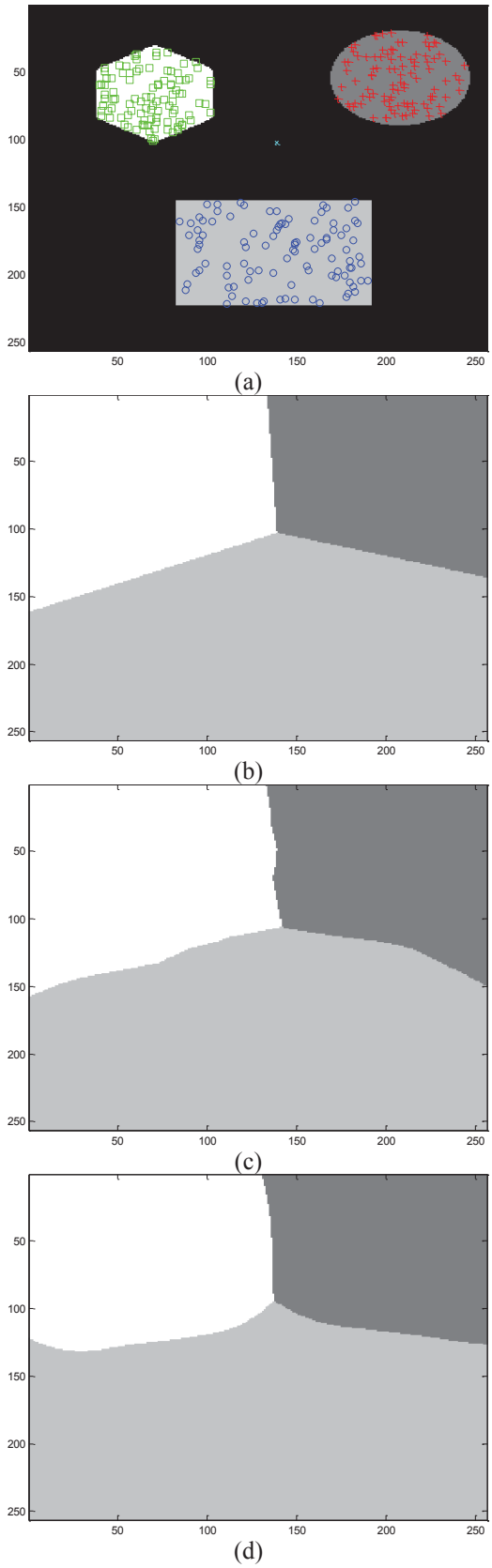


Figure 4. a) Three classes b) Angle-based, c) K-NN-based d) Naïve Bayes-based classification, Ntr=300.

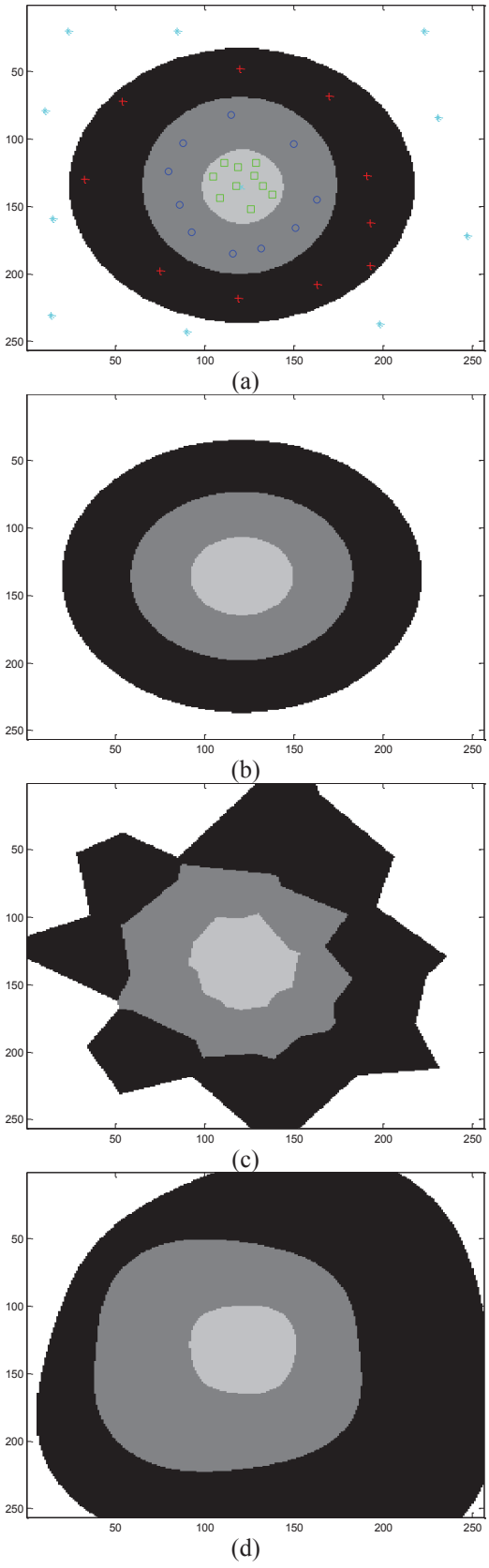


Figure 5. a) Four classes b) Radius-based, c) K-NN-based d) Naïve Bayes-based classification, Ntr=40.

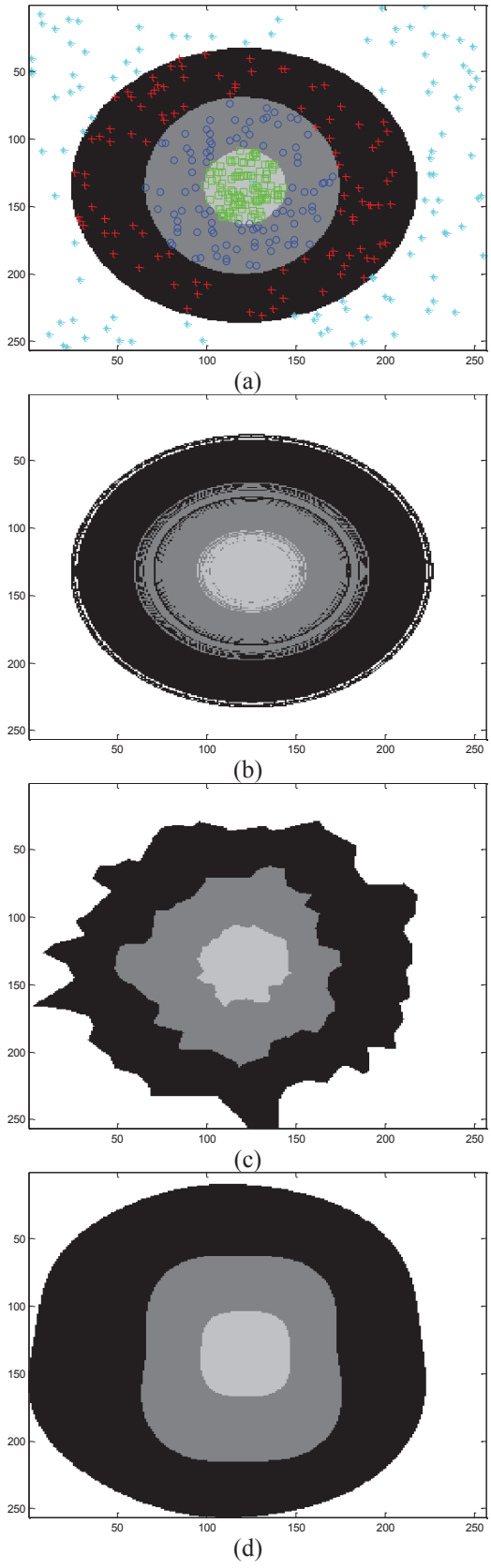


Figure 6. a) Four classes b) Radius-based, c) K-NN-based d) Naïve Bayes-based classification, Ntr=400.

4.2 Classification on the Four-class Images

In Figs. 5-6, the differences between radius-based and K-NN-based classification of feature space were easily noticeable. In Fig. 5, it was observed that K-NN-based classification with few training vectors (Ntr=40) was not able to capture the circular distribution of the classes. However, radius-based classification was very successful in capturing this distribution. NB algorithm was also able to capture the circular distribution of the classes when Ntr=40 although the results were far from the original image. When the number of the training vectors was higher (Fig. 6, Ntr=400), the difference between all three classification approaches decreased. In this case, K-NN and NB-based classification results were improved whereas the radius-based classification results worsened due to the intersection of radii of training vectors' classes. However, this problem of radius-based approach can be handled using filters. Here, again it should be noted that the classification problem given here is suitable for classification by radius.

4.3 Classification Time

In Table III, the huge differences in classification times between both angle-based vs. K-NN/NB and radius-based vs. K-NN/NB classifications are worth mentioning. In angle and radius-based approaches, the classification times for the whole images were < 1s and they did not suffer from the number of training vectors. NB classifications did not suffer from the number of training vectors either. However, NB classification times were the longest. K-NN based classifications were very susceptible to the number of training vectors. Nevertheless, it should be noted that there are methods to improve the efficiency of K-NN processing [11].

4.4 Further improvements

Individual high performances of angle and radius-based approaches are promising for combining these two approaches. However, in order to avoid overfitting problem, this should be performed in an optimum way.

In both of the proposed approaches, the center of the spider web was calculated as the average of all training points. This approach will be sub-optimal when the classes are not equally distributed. Therefore, in the selection of center, the number of the training vectors from each class and class variances can be incorporated.

These approaches were originally proposed for two-dimensional feature vectors but, they can easily be applied to three-dimensional problems. In this case, for the radius calculation, there will be another component for the z-axis. However, in the angle-based approach, there will be two angles: One for the altitude and the other for the azimuth.

5 Conclusions

Angle and radius-based classification approaches were fast and did not require complex calculations. Comparisons with K-NN and NB algorithm on two feature spaces with different classes show the potential of the proposed approaches. Here, angle and radius-based approaches were presented separately. In this case, high performance is achieved for special feature distributions, which are suitable for angle or radius-based classification. In the future, optimum combination of angle and radius-based classification should be examined in order to reach a general classification method. Otherwise, these individual approaches will generate high accuracies in only a very limited type of classification problems and they will not be able to classify different feature space distributions accurately. Furthermore, comparison of the proposed methods with the other classification methods is needed for a more complete assessment of these methods.

6 Acknowledgment

The article was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

7 References

[1] Y. Rui, T. S. Huang and S-F. Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues, Journal of Visual Communication and Image Representation, 10(1): 39–62, 1999.

[2] K-S. Fu and A. Rosenfeld. Pattern Recognition and Image Processing, IEEE Transactions on Computers, C-25(12): 1336-1346, 1976.

[3] E. Herrmann, J. Call, M. V. Hernàndez-Lloreda, B. Hare and M. Tomasello. Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis, Science, 317(5843): 1360-1366, 2007.

[4] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 28(5): 823-870, 2007.

[5] K. A. Heller and Z. Ghahramani. A Nonparametric Bayesian Approach to Modeling Overlapping Clusters, In Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07), proceedings, 2: 187-194, San Juan, Puerto Rico, March, 2007.

[6] W. Tang, K. Z. Mao, L. O. Mak and G. W. Ng. Classification for Overlapping Classes Using Optimized Overlapping Region Detection and Soft Decision, In 13th

Conference on Information Fusion (FUSION), proceedings, pp. 1-8, Edinburgh, United Kingdom, July, 2010.

[7] C. Böhm, S. Berchtold and D. A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases, ACM Computing Surveys – CSUR, 33(3): 322-373, 2001.

[8] E. Fix and J. L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report no. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[9] T. M. Mitchell. Machine Learning, McGraw-Hill, 1997.

[10] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z-H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg. Top 10 algorithms in data mining, Knowledge and Information Systems, 14(1): 1–37, 2008.

[11] C. Yu, B. C. Ooi, K-L. Tan and H. V. Jagadish. Indexing the Distance: An Efficient Method to KNN Processing, In 27th International Conference on Very Large Data Bases, (VLDB'01), proceedings, 421-430, Rome, Italy, September, 2001.

SESSION

IMAGING SCIENCE ALGORITHMS FOR 2D AND 3D + KNOWLEDGE AND INFORMATION ENGINEERING AND EXTRACTION

Chair(s)

TBA

Text Line Extraction Using Seam Carving

Christopher Stoll*, Yingcai Xiao, and Zhong-Hui Duan
Department of Computer Science, The University of Akron, Akron, Ohio, USA
* cas2@zips.uakron.edu

Abstract - *This paper describes a novel technique for information extraction based upon seam carving. Modifications needed to adapt the seam carving process to the new problem domain are explained. Then, two output methods, direct area detection and information masking, are described. Finally, potential modifications to the technique, which could make it suitable for use in other domains such as general computer vision or bioinformatics, are discussed.*

Keywords: text line extraction, seam carving

1. Introduction

While researching the possible application of the eigenface technique [1, 2] towards optical character recognition it was discovered that the technique only worked well when the candidate characters, the characters to be recognized, were scaled and centered exactly as the characters which were used for training. Unlike faces, characters do not share fixed points of reference that can be used to scale and rotate them into a standard form; a robust text line extraction method is required. Existing text extraction approaches, such as X-Y Cut, run-length smearing, and whitespace analysis, did not produce output suitable for the proposed new recognition approach. It is proposed that to achieve the desired results of extracting text from scanned documents, a seam carving approach be used.

With an iterative seam carving approach it will be possible to segment the image, split out the lines of text, and split out the characters from the detected lines. The early iterations identify the text blocks, top and bottom of the characters are identified during the horizontal line splitting step, and the left and right of the characters are identified during the vertical character splitting step. Given the precise boundaries of the candidate characters, they can be projected onto a vector with a preset number of dimensions, and that vector can be used for image recognition. If this approach where to be viable it could further eliminate the need for specific deskewing, connected component analysis, and other preprocessing algorithms. The orientation of the relevant seams could be determined either at the line level or at the character level.

This paper will begin by providing a brief explanation of seam carving, then a novel approach to text line extraction based upon seam carving will be described. Modifications to the seam carving technique, which are necessary to apply seam caving to text line recognition, will be discussed. Then, two variations on the approach, direct area detection and information masking, will be described. Direct area detection attempts to directly locate the boundaries of text lines, while information masking provides a mask under which information is located. Finally, future work and possible alternative applications will be explored.

2. Proposed Approach

The seam caving approach to text extraction has a lot in common with whitespace analysis. The main goal of the seam carving approach described in this paper is to use the whitespace in a document to identify where the interesting information lies; rather than looking for rectangular covers which are joined into large seams, the seam carving approach attempts to find whitespace seams directly.

Consider a blank document, or a white image. With a properly constructed seam tracing function, one that attempts to maintain straight lines, all horizontal seams would go straight across the image. If a single letter or word is added to the center of the document, as seams are examined, from the top down or from left to right, they will begin to deviate as they get closer to the letter (Figure 1). The deviation should reach its maximum around the center of the word, then switch direction, and begin to deviate less.

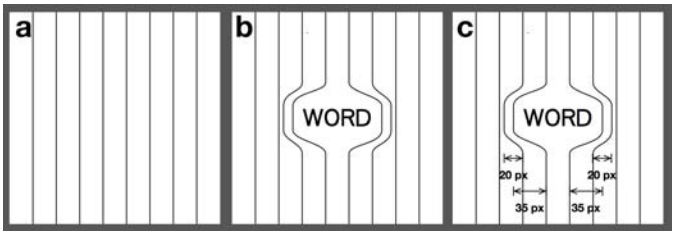


Figure 1: (a) seams in a blank document (b) seams in a document which contains a word (c) seam deviation amounts in a document which contains a word

Seen in this way, just as physical matter in space causes distortions in space-time, information in a document causes distortions in the document space. With the seam carving approach to text segmentation the distortions made by the information are used to identify the boundaries of information within the document. The seam carving approach just needs slight modification to better support this new approach.

2.1. Seam Carving

Seam Carving was originally described by Avidan and Shamir as a technique for content aware image resizing. [3] Their insight was that it could be possible to resize images through the removal of unimportant information from the image rather than though cropping or scaling. The use of

seams, which travel a jagged path completely through the image, allow each row or column to remain the same width as the less important pixels are removed.

2.1.1. Preprocessing Functions

The original seam carving paper does not mention explicit preprocessing steps, and they may not be appropriate for image retargeting, but are needed when the approach is applied to text extraction. First, color images are converted to greyscale since color information is less valuable for the purposes of text extraction.

In addition to converting color images to greyscale, different contrast enhancement techniques were experimented with. Otsu binarization was used, but it did not consistently yield the best results for this application. In many situations passing each pixel through a simple cosine function yielded the best results. In common photo editing applications this is known as curve adjustment, here a cosine function was used for the shape of the curve.

2.1.2. Energy Functions

Since the goal of seam carving is to remove pixels which will not be noticed, Avidan and Shamir considered the images' "energies." Each pixel in an image will have an energy value which essentially represents how similar it is to the pixels around it. There are various methods for calculating energies, and since text is normally contrasted against its background almost any method should work. Here we will use the same simple gradient magnitude function used by Avidan and Shamir.

2.1.3. Seam Traversal

Given the results of an energy function, seams must next be calculated for the image. For image retargeting the image can be treated as a graph and seams can be removed iteratively, but Avidan and Shamir chose to take a dynamic programming approach.

For vertical seams the dynamic programming matrix is filled in a top-to-bottom, left-to-right manner; a seam value is calculated for each pixel by adding the minimum of the left three pixels in the current pixel's 8-connected neighborhood (left above, left, left below) to the current pixel's energy value.

For image retargeting, when analyzing vertical seams, the right column of pixels can be checked for minimums once the dynamic programming matrix is filled. The minimum values represent the starting points for seams which can be removed. From these points backtracking is used to remove pixels.

2.1.4. Seam Traversal Modifications

To apply seam carving towards text extraction the seam traversal procedure is modified. First, with text images, seam values create "shadows" which can obscure information following large features (Figure 2b). This effect could be due to the dynamic programming approach's tendency to propagate errors; to check that, a greedy approach was considered.

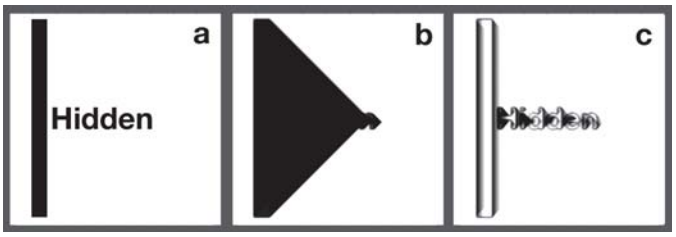


Figure 2: Example of seam "shadows." Original (a), seam values (b), seam values with decrement (c)

The use of a greedy approach, where just the next minimum is taken rather than adding it to the present value, results in a seam value map which is nearly identical to the energy map. The greedy approach does eliminate shadows, however the final results are much worse (Figure 3b). It turns out that the shadows are an important part of this approach. Like a spearhead the shadows guide seams around interesting portions of the image; without the chamfered edges, seams would be more likely to go straight through perpendicular edges rather than flowing around them.

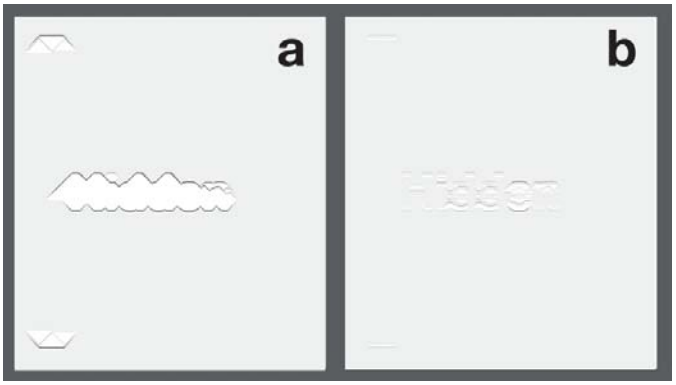


Figure 3: All seams traced (seam traversal darkens image, so white areas have no seam traversal), dynamic programming approach with decrement (a) and greedy approach (b)

Since the greedy approach did not yield the desired results, the dynamic programming approach will be modified to optimize the effect of the shadows. The dynamic programming matrix is filled as described by Avidan and Shamir, except that each pixel's resulting seam value is reduced by some value k if it is greater than zero (Equation 1). This is a key modification, without this decrement seams would never travel between text lines due to how seam values are carried across the image. A k value of 1 was experimentally determined to yield satisfactory results (Figure 2c). Based upon what was learned form taking the greedy approach, it is likely that the optimal k value is image dependent. Further research needs to be done to identify the optimal k value.

$$M(i, j) = e(i, j) + \min(M(i-1, j-1), M(i-1, j), M(i-1, j+1)) - k \quad (1)$$

2.1.5. Text Extraction Steps

Additional, unique steps are required to use seam carving for text line extraction. Rather than examining the last column or row for minimum values, each pixel in the terminal row or column is examined. In fact, finding the seams of minimum value does not generally work with text images. For most images containing text the last row or column will contain a large number of minimum values since the background of most documents is some uniform color. So, backtracking must be performed for each of the pixels in the terminal row or column. Note that since we are not removing any lines, energy values will not be recomputed between seam traversals.

The process of backtracking must also be modified to support the new problem domain. For image retargeting the actual seam path matters little, whereas for text line extraction it is expected to proceed in a straight line since text is normally written in lines. So, in addition to considering the values of the next step when backtracking, the overall deviation from straight is also considered.

0	0	0	0	0	0	0	0	0	0
0	0	0	8	16	16	8	0	0	0
0	0	8	32	32	32	32	8	0	0
0	0	32	16	0	0	16	32	0	0
0	0	32	16	0	0	16	32	0	0
0	0	32	32	32	32	32	32	0	0
0	0	32	8	0	0	8	32	0	0
0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	0
0	0	0	8	16	16	8	0	0	0
0	0	8	32	32	32	32	8	0	0
0	0	32	16	0	0	16	32	0	0
0	0	32	16	0	0	16	32	0	0
0	0	32	32	32	32	32	32	0	0
0	0	32	8	0	0	8	32	0	0
0	0	0	0	0	0	0	0	0	0

Figure 4: Possible (light grey) and probable (dark grey) seam paths are shown for a standard seam carving approach (left). Top and bottom seam paths (dark grey) are shown for backtracking with added deviation constraints (right)

If a seam has deviated from vertical or horizontal then the algorithm will force the seam back in that direction. When there are more than one possible backtracking move with the same value, the algorithm has discretion to choose which path to take. For the purposes of this text extraction approach, the algorithm must take the center path whenever possible, unless there is a net deviation. When there is a deviation the implementation must take the choice that minimizes the deviation. This results in seams which minimize the deviation and minimize the energy of the path.

2.2. Information Extraction

From here there are two distinct approaches to identifying the distortions in the image seams. Direct area detection uses heuristics to identify information as originally proposed, whereas information masking crates an image mask based upon where no seams travel.

2.2.1. Direct Area Detection

The first way to identify distortions in the seams is to check for seams which have the most net deviation from straight, as originally discussed. As a horizontal seam traverses across the image each column's deviation from the starting row is recorded. The sum of these deviations is the net deviation. When a seam's net deviation changes sign it represents the bottom of an area, and the seam prior to it represents the top of an area. Long straight runs at the beginning and end are trimmed off.

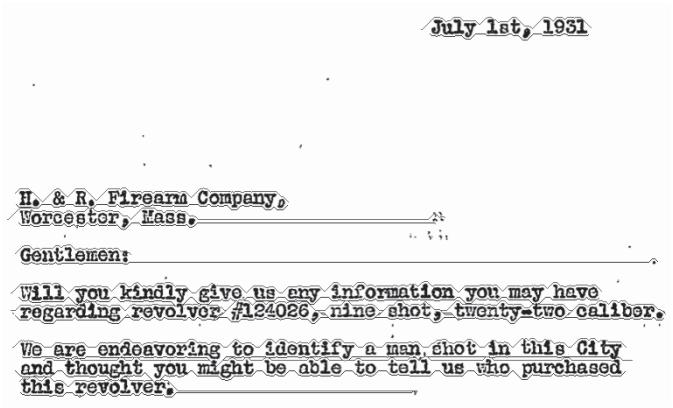


Figure 5: Direct Area Detection

In practice, seams defining the beginning and ending boundaries (upper and lower for horizontal seams or left and right for vertical seams) are rarely consecutive. If a beginning seam has not been found (without a corresponding ending seam) then local maximum net deviations are sought. Once a local maximum is found, a local minimum net deviation is sought. This process repeats as the entire image is processed.

2.2.2. Information Masking

Another approach to identify distortions in the seams is to look for pixels where seams overlap. As seams divert around information they will tend to go through the pixels just outside the upper and lower boundary. If a matrix is created to keep track of how many seams pass through each pixel, then the matrix cells nearest to information will have higher values (black areas in Figure 6) and matrix cells where information resides will have zero values (white areas in Figure 6).

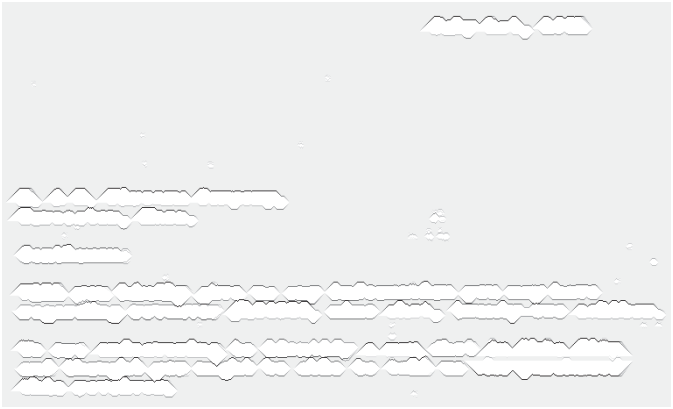


Figure 6: Information Masking

2.2.3. Method Comparison

Direct area detection and information masking both work reasonably well when the document is simple, but in-

formation masking automatically works much better on more complicated document layouts. When there are three columns of irregularly laid out words the difference in character size between the columns prevents lines from being properly identified. In the cases where lines are identified correctly, they are actually in different columns and should not be treated as single lines. Another issue with direct area detection is that words near the edge are sometimes not properly detected. The problem appears to be that the leading and trailing lines are trimmed prematurely due to the lack of seam run length. Finally, direct area detection, without further enhancements, begins to fail faster on documents which are skewed; this is especially problematic since handling skew is normally a strength of the seam carving approach.

Information masking is superior to direct area detection when the target text is skewed. However, when text skew reaches a certain point even information masking, under the current implementation, begins to degrade in performance. Currently, direct area detection clearly begins to break down at two degrees of skew, and stops working completely by the time the text is skewed by eight degrees. Information masking, since it does not rely upon heuristics, continues to work at eight degrees of skew, but the seam markers at the beginning and ending of lines becomes stretched.

3. Discussion

In order to be a viable solution the shortcomings mentioned in the above sections should be overcome, otherwise there is less compulsion to use this technique over the established techniques. The problem of skew is minor; it has been shown, through the information masking output, that the raw seam carving technique can identify text that is skewed. It is only the heuristics for direct area detection that need to be improved. And, if all else fails, a dedicated deskewing algorithm can be implemented as a preprocessing step.

The problem of handling more complicated text layouts, however, is more significant. Again, the raw seam carving technique, as shown through the information masking approach, can identify text regardless of its position, but keeping the text logically grouped is, as it always has been, more problematic.

3.1.1. Improved Skew Handling

In order to improve the handling of skewed text the approach could be modified so that it does not attempt to force seams to travel at precisely 90 or 180 degrees. Currently the algorithm encourages seams to end in the same row or column in which they started. A better method would be to identify deviation trends; if the seam is steadily deviating more, then it could be assumed that the information that is displaying it is skewed. Rather than attempting to force seams to travel at exactly 90 degrees or 180 degrees the algorithm could attempt to force the seam to travel along the trend angle.

Since direct area detection performs so poorly at detecting starting and ending seams when the text is skewed, that portion should also be improved. Currently it simply evaluates the magnitude of the net deviation, once it changes sign the top and bottom seams are presumed to have been found, but this is not necessarily true with skewed text. Checking for a sign change in magnitude should account for the overall angle of seams within a certain range.

In all cases the heuristics used should be improved through more thorough data analysis, which in turn requires improved parameterization of the algorithm. The values used for heuristics are the result of manually examining the algorithm's performance over a very small data set. The use of a larger and more varied set of example images should yield better heuristic values.

3.1.2. Handling Complicated Layouts

There are various possible techniques to overcome the problems of locating text within columns and other more complicated layouts. Most of these possible techniques are based upon methods described in other text extraction approaches.

The first option is to run the seam carving algorithm both vertically and horizontally, then using a heuristic to determine which direction is more appropriate to split in first. One possible heuristic is total resistance. The idea is that, as each of the seams are explored, when they are forced to deviate from a straight path they are experiencing resistance. For direct area detection the net deviation for each seam is calculated, so it would be trivial to sum these quantities. The sum net deviations for horizontal and vertical seam traversal should then be compared. The direction with less resistance, or a lower sum net deviation, is the direction which should be explored first. This technique helps when identifying an information mask, but it is still very crude. It tends to favor moving horizontally (for left-to-right or right-to-left texts) and does not help improve direct area detection.

Another approach to higher level segmentation would be to take a kD-tree or X-Y cut type of approach. Instead of initially tracing all the seams, the minimum seams will be found, as is done with typical seam carving. The difference comes in how "minimum seam" is defined. Rather than checking the terminal value of the seam this approach would need to find the seam with the minimum net deviation. Since there are likely to be multiple seams with the same minimum net deviation, this approach would need to find the center seam of the largest grouping of minimum net deviation seams. The identified seam would be used to split the image and the process would be ran again on each of the halves. This was not implemented as a part of this project, but is a goal of future research.

3.2. Extracting Finer Details

Setting aside this approach's current limitations, more of its benefits will be described. Given a non-skewed document which contains simple content with a manhattan layout, it is possible to iteratively run the same seam carving process on identified sub areas. If the first pass of the algorithm identi-

fied sentences (Figure 5), then the next pass of the algorithm will identify phrases, words, or letters (Figure 7). The technique is simply used against areas identified in the first step, but in a perpendicular direction. This process can be repeated until the base case or halting condition, which is not yet fully defined, is reached.

Figure 7: Seams identified in previously identified areas of interest. The first line (left) and second line (right) identified in the document from Figure 5

3.3. Asymptotic Analysis

The first step in the seam carving process is to read in an image and create an appropriate data structure. The example program performs brightness, contrast, and energy calculations as distinct steps, but this is done to allow for flexibility during testing, a production program would combine these into a single step. The Difference of Gaussians could be performed in one pass given the appropriate data structure. So, assuming that the image has a width of n and a height of m, then this process will take O(*nm*).

The next step is to fill the seam matrix, which requires iterating over the image again, taking O(*nm*). Finally, all the seams must be traversed. Considering horizontal seam evaluation, each of the n starting pixels is considered as a starting point, and the seam must cross the entire height m of the image, so this process takes another O(*nm*). The runtime performance for a single pass is 3 * O(*nm*), and 5 * O(*nm*) when both horizontal and vertical seams are evaluated. So, this approach performs in O(*nm*) time, where n is image width and m is image height. This conclusion is supported by empirical analysis (Figure 8).

run-times

size	seconds
320 x 320	0.104
640 x 640	0.364
960 x 960	0.668
1280 x 1280	1.37
1920 x 1920	2.744
2560 x 2560	5.126

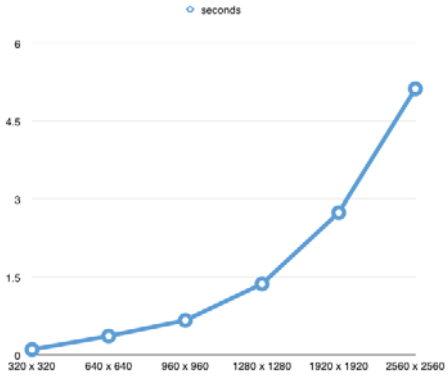


Figure 8: Actual run-time performance data; both directions, average of 20 runs

It should be noted that although vertical and horizontal seam analysis both perform in essentially O(n^2) time, real world performance can be drastically different due to algorithm implementation and hardware limitations. If the image data structure uses row major order, then cutting horizontal seams can take twice as long to accomplish on some hardware due to how memory is cached and moved to the processor.

3.4. Related Work

Seam carving has previously been used in the area of handwriting recognition, however this approach takes a slightly different perspective on the problem. Saabini and El-Sana describe a method for applying seam carving to handwritten text line extraction [11]. When they surveyed the research on text line extraction for handwriting recognition they also found that most of the techniques relied upon some sort of connected component analysis, so they devised a novel technique based upon seam carving.

The algorithm described by Saabini and El-Sana calculates the image's or document's energy in such a way (using a signed distance transformation) that the selected seam paths went through the lines of handwritten text rather than attempting to identify the whitespace. Once the text line's center line is found it expands vertically to identify the full height of the line.

One major advantage of their approach is that small components, such as the dot above an "i" can be included in the row based upon what the algorithm learns about row heights. The new technique described in this paper can be susceptible to leaving out such small components, especially when processing images which have not first been subdivided. In practice, for single lines of text, the technique described here automatically includes small pieces of information due to how the heuristics are set find maximum deviation.

Another benefit of Saabini and El-Sana's technique is that since it is tracking lines of text, it can handle multi-skew and lines that actually touch each other. For the technique described in this paper, even if a moving average was used to determine line skew, it will still have difficulty handling multi-skew. This technique can also cope with lines that touch at least as well as the method described by Saabini and El-Sana.

Asi, Saabini and El-Sana subsequently demonstrated that an enhancement to Saabini and El-Sana's approach could identify the precise boundaries between text lines [19]. Their method's ability to do this is entirely dependent upon it ability to find the text lines' medial path, which in turn relies upon single columns of handwritten information. Whereas their work is likely to perform superiorly in identifying lines of free-form handwritten text, the method described in this paper is likely to excel in typed or printed documents. The method described here is intended to run recursively, so there is some expectations that the information it finds will be subsequently subdivided, whereas their approach is intended to identify information using a single pass.

3.5. Future Work

The existing implementation of seam carving for text line extraction performs as well as other notable text extraction methods. Like them it has trouble with skewed text and more complicated layouts, traditional problems in this area of

research. Fortunately, there are some techniques to possibly overcome these shortcomings. The first improvements should be in the area of handling skewed text. The technique should consider some sort of moving average of the seam angles, but more research will need to be done in this area.

Also, experiments should be conducted where different energy maps are used. The use of signed distance transforms (SDT), as Saabini and El-Sana, should be evaluated[11]. This approach will be looking for local maximum (the whitespace), rather than local minimum. Also, the approach of using elliptical convolution kernels which align to the direction of travel for the image energy calculation, as described by Zhang and Tan, should be explored [19]. Zhang and Tan also described a method for limiting the distance which energy should be propagated which should be incorporated into this approach.

Another area of future work is the implementation of a kD-tree or X-Y cut approach to higher level document segmentation. This will help define the structure of the document and improve actual text extraction. Since a dynamic programming approach has been taken and due to the nature of seams within text documents, it should be possible to reuse seam data as the document is subdivided, thus reducing computation time. A check may be required to ensure that the border pixels of the sub-area all contain zero weights. Also, a more precisely defined base condition would need to be defined to limit the number of subdivision attempts.

An alternative approach to higher level document segmentation using seam carving would be to use the idea of "zooming," or otherwise considering the level of detail (LOD). This idea is based upon how people process information within their view; people can get an overview of what they see or they can understand fine details within it, but those are generally two distinct steps. To simulate this process the candidate image could have its size reduced and contents blurred to decrease the definition of finer details. Then, when seam carving is ran it will not find sentences, for example, but rather it would find paragraphs or columns.

The approach of using LOD could bring multiple benefits. First, since the seam caving approach runs in polynomial time, a reduction in image size would greatly in crease runtime performance. However, the dynamic programming matrix would have to be recreated at each zoom level which could negate any realized gains. If this modification was combined with the kD-tree modification, it would make identification of higher level divisions easier. The idea of "zooming out" also presents opportunities for early recognition or at least setting a context for subsequent recognition steps. It may be possible to run the low resolution areas of interest through an algorithm which quickly detects basic shapes, or one that uses textons [12] to detect textures. If this approach were viable it would have implications outside of the realm of OCR; it could be used in the field of bioinformatics or for general computer vision.

Existing implementations of this approach have been experimentally applied to bioinformatics. There are techniques being developed to identify carcinomas based upon examining differences of protein profiles; presently the differentiation of the protein profiles is done by manual inspection [16]. Enhanced version of the seam carving approach may allow for algorithmic differentiation of protein profiles.

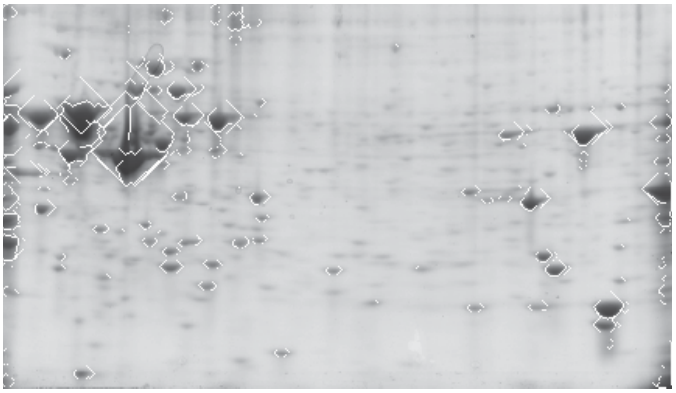


Figure 9: Experimental application toward protein differentiation; white lines around dark areas show boundaries found running this approach both vertically and horizontally

Successful application to bioinformatics may further lead to applications toward general computer vision. Consider an image which contains a stop sign. If the algorithm identifies the sign as being interesting in one of the initial iterations, and if the octagonal shape could be detected, then it may not even be necessary to continue processing the area — if a person sees a stop sign out of the corner of their eye they do not need to read the word "STOP" to understand the meaning of it. For most driving adults, seeing a red octagon is enough to understand that the symbol means stop. Also, if "STOP" is being recognized but some of the letters are illegible, the fact that the word being recognized is within an octagon would induce the algorithm to respond that the word is "STOP."

3. Conclusions

A promising new approach to text extraction has been described based upon seam carving. The seam carving approach brings many of the same benefits of white space analysis techniques; it operates largely parameter free, it does not make assumptions about document layout and it works regardless of text direction or page orientation. And though it also has many of the same limitations that whitespace analysis, there are some promising methods unique to this approach which can help overcome those shortcomings and allow for it to be used outside of the original problem domain.

Applying the seam carving approach to text line extraction reinforced that the original technique was designed for images and not text documents. When seam values are calculated, the original technique causes "shadows" to be cast which can block out smaller pieces of information. The shadows represent errors which, unlike images, easily get propagated through the mostly blank space of a text document. The shadows were reduced by adding decay to the seam value function.

The approach described by Avidan and Shamir also relies upon the ability to find minimal terminal seam values from which to backtrack. For text documents the terminal value of all seams is likely to be zero. This is due to the amount of whitespace, space without information, in text documents. Images have a higher information density and are not normally padded with headspace. Modifying the technique so that every seam terminus is examined is necessary to apply seam carving to text line extraction.

The original seam carving technique further allowed for indeterminate seam paths (determined by the implementation). For resizing images it is probably beneficial to have some randomness in the seam path, but that works against text line extraction. Text normally flows in straight lines, and it makes sense to consider that when looking to extract text lines. To apply seam carving to text extraction the concept of net deviation was introduced; in addition to minimizing the seam path cost the net deviation of the seam is also minimized.

These minor innovations have enabled the seam carving algorithm to be applied to task of text line extraction; it builds upon existing techniques, yet represents a truly novel approach. The potential benefits of perfecting this approach could have impacts beyond optical character recognition, and is thus worthy of more research.

4. References

1. L. Sirovich and M. Kirby, “Low-dimensional procedure for the characterization of human faces,” Journal of the Optical Society of America A, Volume 4, Issue 3, pp. 519-524, 1987.

2. M. Turk and A. Pentland, “Face recognition using eigenfaces,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591, 1991.

3. S. Avidan, and A. Shamir, "Seam carving for content-aware image resizing,” ACM Transactions on graphics (TOG), Volume 26, Number 3, 2007.

4. K. Y. Wong, R. G. Casey, and F. M. Wahl, “Document Analysis System,” IBM Journal of Research and Development, Volume 26, Issue 6, pp. 647-656, 1982.

5. G. Nagy and S. Seth, “Hierarchical representation of optically scanned documents,” International Conference on Pattern Recognition - ICPR, 1984.

6. L. O’Gorman, “The Document Spectrum for Page Layout Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 15 Issue 11, pp. 1162-1173, 1993.

7. H. S. Baird, “Background Structure In Document Images,” In Advances in Structural and Syntactic Pattern Recognition, pp. 17-34, 1992.

8. T. M. Breuel, “Two Geometric Algorithms for Layout Analysis,” Document Analysis Systems, 2002.

9. T. M. Breuel, “Robust least square baseline finding using a branch and bound
algorithm,” Document Recognition and Retrieval, SPIE, pp. 20-27, 2002.

10. K. Kise, A. Sato, and M. Iwata, “Segmentation of Page Images Using the Area Voronoi Diagram,” Computer Vision and Image Understanding, Volume 70, Number 3, pp. 370-382, 1998.

11. R. Saabni and J. El-Sana, “Language-Independent Text Lines Extraction Using Seam Carving,” 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 563-568, 2011.

12. B. Julesz, “Textons, the Elements of Texture Perception, and their Interactions,” Nature, Volume 290, pp. 91–97, 1981.

13. J. Ha, R. M. Haralick, and I. T. Phillips, “Recursive X-Y Cut using Bounding Boxes of Connected Components,” Third International Conference on Document Analysis and Recognition, pp. 952-955, 1995.

14. L. O’Gorman, “Image and document processing techniques for the RightPages electronic library system,” International Conference on Pattern Recognition, pp. 260-263, 1992.

15. G. Nagy, J. Kanai, M. Krishnamoorthy, M. Thomas, and M. Viswanathan, ‘‘Two Complementary Techniques for Digitized Document Analysis” ACM Conference on Document Processing Systems, 1988.

16. T. Shi, F. Dong, L. S. Liou, Z. H. Duan, A. C. Novick, J. A. DiDonato, ‘’Differential Protein Profiling in Renal-Cell Carcinoma” Molecular Carcinogenesis, Volume 40, pp. 47-61, 2004.

17. Medical Article Records Groundtruth (MARG), National Library of Medicine, http://marg.nlm.nih.gov, 2005.

18. S. Firilli, F. Leuzzi, F. Rotella, F. Esposito, ‘’A Run Length Smoothing-Based Algorithm for Non-Manhattan Document Segmentation,” Italy, 2012.

19. A. Asi, R. Saabni and J. El-Sana, “Text Line Segmentation for Gray Scale Historical Document Images,” International Workshop on Historical Document Imaging and Processing (HIP), 2011.

20. X. Zhang and C. L. Tan, “Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving”

Global Optimal and Minimal Solutions to K-means Cluster Analysis

Ruming Li [1], Xiu-Qing Li [2], and Guixue Wang [3*]
[1,3]Key Laboratory of Biorheological Science and Technology (Chongqing University), Ministry of Education; Bioengineering College of Chongqing University, Chongqing, 400044, China
[2]Molecular Genetics Laboratory, Potato Research Centre, Agriculture and Agri-Food Canada, Fredericton, New Brunswick, E3B 4Z7 Canada

Abstract - *The K-means clustering method has been widely used in nonhierarchical cluster analysis of multi-dimensional data sets. Chronic problems with the K-means clustering algorithms since the 1960s have been the local minima of clustering results, computationally expensive iterations, and suboptimal solutions with large-size data sets. Their usabilities are controversial and risky. Without using traditional heuristics, we announced our milestone solutions to K-means cluster analysis in a novel global partitioning model. They were developed to overcome all of these problems and make the K-means algorithm truly optimal. These solutions innovated traditional algorithms and introduced new methods for rationally partitioning a data set from the globality and integrity of data structure and object relationships. The theories and techniques involved were experimentally proven by all successful implementations. The globally optimal results confirmed that the breakthrough was achieved in rapidly classifying any type/size of data sets into any number of disjoint clusters with a global minimum of total error sum of squares (TESS).*

Keywords: cluster analysis, K-means clustering, global optimal partitioning, clustering error and quality, global optimization and minimization, global minimum TESS.

1 Introduction

The *K*-means clustering method is a major nonhierarchical or partitional classification technique that has been most commonly used in information processing and exploratory data analysis such as statistics, informatics, phylogenetics, gene clustering, data mining, pattern or trend recognition, image segmentation, and machine learning. Particularly *K*-means clustering is more efficient in processing of a massive data than hierarchical clustering that is computationally expensive on a very large data set. *K*-means clustering is used to divide a set of objects (aka, entities, cases, observations, data points, samples, or items) into *K* subsets or clusters (aka, classes, groups or partitions). The separated clusters are disjoint and the members in a cluster are sufficiently close or similar to each other and sufficiently distant or dissimilar to non-members in other clusters [1-3]. The number *K* of such clusters can be either known a priori or pre-set by the algorithm or pre-defined by users. The clustering error is measured by a total error sum of squares (TESS) in statistics and the criterion of optimal clusterings takes the least TESS that expresses the minimized clustering error. Suppose that x is an arbitrary data point in the m-dimensional data space and let l be the number of clusters and n be the number of objects (data points) in a cluster, the least objective function TESS is defined as:

$$\text{Minimum} \sum_{k=1}^{l} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{kij} - \bar{x}_{kj} \right)^2$$

where $\bar{x}_{kj}$ is the j-th component of a mean vector or centroid for the k-th cluster. When the least TESS is satisfied, it translates into the accurate (most reasonable) cluster membership that is achieved for each of clusters. That is, all members within a cluster are closest or most similar to each other and most distant or dissimilar to non-members otherwise.

Traditionally, the partitional clustering method operates on the various algorithms that are unexceptionally based upon the *K*-center model or centrotype in each cluster. In the *K*-means model, the centrotype is the arithmetic mean vector or cluster centroid (barycenter). The primary *K*-means procedure starts from initial *K* seeds or cluster centers and then uses an iterative refinement heuristic with centroids. Unfortunately this algorithm terminates with a local convergence and does not definitely find the globally optimal cluster configuration corresponding to the objective function minimum. The main reasons are that it operates on the center-based model and is inevitably sensitive to the initial cluster centers aside from its local assignment of data points. The algorithm can be run multiple times to reduce these effects but there is no guarantee that it should converge to a global minimum even if a stopping criterion is met. What is worse, this would bring another issue that such algorithmic iterations cause a heavy computational load.

The *K*-means algorithm has been being improved since the year 1967 [4-6]. The significant progress was made in the early 21st century with a couple of global *K*-means algorithms being a new paradigm [7-10]. These modified algorithms introduced or adopted the typical incremental approach to clustering that dynamically adds one cluster center at a time. Although these algorithms still could not get rid of the center-based model in which issues associated with cluster centers and iterations

remained, they provided better heuristics that were approximating global optimal solutions [11-13].

For the larger number of clusters to partition and a huge-size data set, all the heuristic algorithms based on the *K*-center model would have the undesirable performance for an optimal clustering solution. Especially, the overwhelming demand for such a desired solution quality through an improved *K*-means algorithm is that a globally minimized TESS must be first of all satisfied.

Then what is the perfect way of partitioning a given data set into disjoint clusters that realizes a global optimization and ends up with the highest quality of clusterings? What is the ultimately minimum TESS reached from such *K*-means clusterings no matter what size a data set has? Our methodology and algorithm devised took care of these issues and worked around all these problems from the perspectives of data integrity, global partitioning, implementation efficiency and outcome robustness. They changed conventional thinking and abandoned the *K*-center model [14]. The breakthrough was obtained in that it is a multidisciplinary solution instead of the pure mathematical development. This particularly involves application of computational intelligence, informatics, combinatorics, logic, operations research, and statistics to the algorithmic or programmatic solutions. In this study, all underlying concepts and theories were addressed and experimentally proven. The objective function TESS that the algorithm should minimize was implemented by our software ParCluster (Partition Cluster for short). For convenience of study, the above notation used is effective thereafter. Also, out of consideration for space and simplicity, we have to use small-sized yet poorly-clusterable sample data for demonstration purposes, but their principles are universally extensible.

2 Methods

2.1 Data clustering error definitions

In *K*-means clustering, the relative importance or individual weight of the clustering error for a single member in a cluster can be quantified and expressed as an error sum of squares (SESS) in statistics. It is the summation of squared differences between each variable and its mean in the centroid for an *m*-dimensional data point; that is,

$$\text{SESS} = \sum_{j=1}^{m} \left(x_j - \bar{x}_j \right)^2$$

where the square-root of SESS is the Euclidean distance. The greater SESS a single member has, the farther it is away from its centroid and also the more distant or dissimilar to other members in a cluster (i.e., the weaker membership). Likewise, for all members in a cluster, the aggregated SESSs are termed as CESS with respect to the clustering error resulted from the entire cluster. That is,

$$\text{CESS} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} - \bar{x}_j \right)^2 .$$

In the special case of a singleton cluster that contains one member only, CESS equals zero.

2.2 Data extraction techniques and criteria

There are two techniques, one is SESS-based and the other is CESS-based, and three criteria to extract data and assign it to another cluster. The first data extraction criterion is defined as: If a member object in a cluster has a larger SESS than it does when placed in another cluster, it should be grabbed from its cluster and allocated to that cluster. The second data extraction criterion is defined as: Compute the difference between the SESS of each member object in a cluster and the resulting SESS when it is placed in another cluster. If these differences are positive (i.e., if the first criterion is met), choose the largest one and its corresponding object should be grabbed and allocated as such. The third data extraction criterion is defined as: If a member object in a cluster is placed in another cluster, the sum of resulting CESSs for the two current clusters is smaller than it is for the two previous clusters. Then that member object should be grabbed and allocated as such.

2.3 Data partitioning theory

For a given objects (data points) in the *m*-dimensional data space or Euclidean space, they have a data structure consisting of integral parts of all members. For this reason, all of their data points cannot be individually, separately or locally treated and manipulated in terms of global optimization. That is, each partitioning of data points must be performed under a global setting where each point dynamically coordinates with any other points [15], which collectively contributes to minimiza- tion of the total clustering error. Since the classic *K*-center model starts with *K*-partitions and fails to provide a data-dependent setting, it is trapped in a local setting from the very beginning and never bails out. Instead we introduce a novel, unbiased, data-driven global partitioning model for *K*-means clustering.

In our new model, each partitioning of data points is performed with all other points being taken into account. It is a straightforward classification process without using an iterative refinement heuristic with centroids. What it does is starting from a bipartition with a minimized TESS. Then further partition one of clusters, which has the largest SESS, into a tripartition and so forth until a specified number of clusters are partitioned. For each cluster partitioned throughout the clustering, its CESS is guaranteed to be a minimum. Thereby, each level of clusterings is guaranteed to have a minimized TESS, so is the TESS when a final level of clusterings is done. The rationale is that theoretically those clusters having established, firm, or close memberships are not necessarily re-partitioned and only the cluster having a global maximum SESS needs to do so. It is termed "partitional cluster". All partitions and assignments of data points follow a global maximal difference law throughout. This law requires that a maximum "ESS" due to the maximal difference always take precedence over others. However, when a cluster having the second-to-maximum SESS has the larger CESS than does the one having the maximum SESS, both need to be minimized in TESS. This protects the result from being

non-minimized when their SESSs differ so little. Specially, if there are clusters having tie maximum SESSs, all of them are subject to TESS minimization. In addition, one or more data points from other clusters may be dynamically extracted to join the new cluster (reshuffled) and orchestrated to lead to a global optimization. As the number of clusters increases, more of the closest members (patterns) get isolated (recognized) from clusters. Then the clustering process becomes increasingly simpler and easier, no matter what size a data set may have [11]. The general data flow directions and order in the partitioning process were illustrated in Fig. 1.

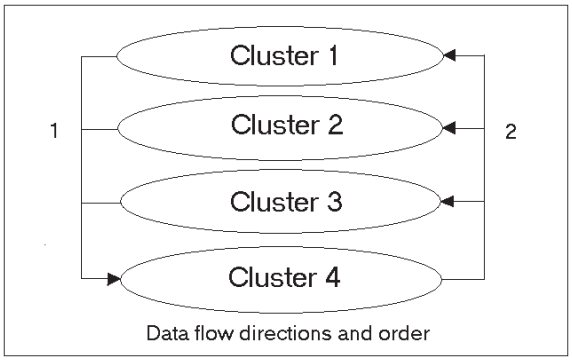


Fig. 1. A general view of the partitioning process from the existing Clusters 1, 2, and 3 to the newly-generated Cluster 4. The data flow among them starts from 1 and ends up with 2.

The principal procedure and algorithms are outlined as belows.

1. Bipartition all data points into two clusters starting with a maximum SESS.
2. Partition one of clusters into two sub-clusters with a global maximum SESS.
3. Re-partition that cluster as SESS-based and CESS-weighted tripartitions.
4. Re-partition that cluster as tri-partitions based on global object relationships.
5. Re-partition that cluster as tri-partitions based on the dichotomous evaluations.
6. Post-partition all such-obtained clusterings, each being globally optimized.
7. Keep track of all TESSs yielded and take a minimum as the convergent end.
8. Repeat Steps 2-7 for a specified number of clusters until they are partitioned.

Step 1 produces the primitive level of bipartition. Step 2 is the partition using a global maximum SESS from a partitional cluster as a seed. Step 3 refers to the partition using a global maximum SESS+CESS as the criterion of a partitional cluster; this step is skipped if Step 2 uses the same partitional cluster. Step 4 refers to the partition using deviators from a partitional cluster that have relationships with objects in other clusters. Step 5 refers to the partition taking a maximum SESS from a maximum SESS+CESS partitional cluster and taking a maximum SESS unbiasedly from any other clusters. This two-way value taking is called dichotomous evaluation. Step 6 refers to the partition using a global TESS minimizer after all the above tasks are done, which is a way to produce a guaranteed global minimum. Step 7 records all TESSs of optimized clusterings and takes a minimum as the optimum reached for that level. Step 8 iterates the above procedure till the given partitions are made.

The idea behind the theory is to follow the global maximal difference law and to gain an equilibrium at which each cluster gets the least CESS and forms a stable membership from an initial level of clusterings. Then break the equilibrium, regain it, break it again, and so on until the objective function TESS in a final level of *K*-means clusterings is minimized. The bottom line is that a hard-to-reach global optimum can be achieved by a guaranteed optimization from a primitive level of clusterings up to the last one. That is, break the harder big problem into the smaller one that is easier to optimize. Each level of globally optimized results makes the last global minimum reached.

2.4 The context of pivotal-point theorem

For a given *m*-dimensional data set, assuming that it is partitioned into multiple clusters, there may be some data points that cannot be allocated normally from cluster to cluster based on the data extraction criteria. This is caused by the properties of a couple of data points having very close or similar component (i.e., variable) values as well as complementary values. The extreme case is that they have all the equal component values across *m*-dimensional variables. That is, some data points may be identical or duplicate, especially in low dimensional data sets. Complementary values are those component values that have small-for-large values for one variable and large-for-small values for another variable. They are exemplified as follows:

	Variable 1	Variable 2
Data point 1:	3.0	4.0
Data point 2:	4.0	3.0
Mean vector:	3.5	3.5

where Point 1 is smaller than Point 2 for Variable 1 and larger than Point 2 for Variable 2 in a *2*-dimensional data set and hence the net effect is a decrease in difference between the two points, as their contributions to the mean vector have the same value (3.5). This will make such complementary values behave like close-component data points. All these properties make it difficult to partition such data points into a common cluster while they are separate or into different clusters while they have tight bonds within a cluster. It should be noted that these properties are also partially responsible for early clustering terminations against stopping criteria and locally converged minima. To solve the problems with data points being of such properties, the solution is to follow the pivotal-point theorem.

Pivotal-point theorem: The data points possessing all close, equal or complementary components in *K*-means clustering are defined as and behave like pivotal points. By playing a pivotal

role, they can preclude data points from being further partitioned based on the data extraction criteria, or terminating data point allocation among clusters hinges on them. When one (or more) of these pivotal points is (are) assumed (forced) to be placed in a target cluster, the further partitioning of data points may be allowed to proceed. This can finally lead to a global optimization of *K*-means clustering and its TESS minimization if a given condition is met. The condition met requires that, based on the second data extraction criterion, a data point with the largest negative difference, instead of the largest positive one, be chosen as the pivotal point to extract. After pivotal point(s) is (are) extracted and assumed to be placed in a target cluster, *K*-means clustering will tend to global optimization and TESS minimization if and only if either of the following scenarios applies:

1) For one pivotal point, it will cause an increase in TESS first but its extraction and placement may induce other point(s) to become extractable, which finally, collectively results in the less overall TESS.

2) For multiple pivotal points, their successive extractions and placements in a target cluster yield the smaller and smaller absolute value of the negative difference until it turns to positive one. Also this may induce other points to be extractable and results in a net decrease in TESS.

There is a constraint that, when such an absolute value becomes larger, the extraction of pivotal points stops. This makes the algorithm non-greedy but enable the global optimization and minimization. When pivotal points become extractable, we say that further partitioning data points is allowed to proceed; that is also to say, a balance due to local optimality is broken or local "convergency".is passed or skipped. This is the most significant contribution of the theorem to realizing a global optimality for the *K*-means solution. Particularly, to handle equal-component pivotal points (identical objects), they are grouped, grabbed, and moved as one unit across clusters to ensure inseparability, integrity, and efficiency throughout the clustering.

Again, the concept underlying the theorem is following the gain-and-break law of the aforementioned cluster membership equilibrium. Here the equilibrium is the early stopped partitioning process or local or near-optimal "convergence" of TESS.

2.5 Global maximal SESS method

If a member object has the global maximal SESS in a cluster, it contributes the most to the total clustering error (TESS) and also results in the maximal CESS for that cluster. Thus, if that incompatible object is eliminated from that cluster, it must be the right candidate member to generate a new cluster for a bipartition. That is, take that object as an initial seed and use it to start partitioning of data points.

2.6 Global object relationship method

In addition to the above core techniques used in *K*-means clustering, another key technique is required to continuously isolate the closest members from a cluster. Generally, the members joined in a cluster are arranged in the order of inherent relationship (closeness or similarity) [12,14]. The member(s) last joined in a cluster may have the weakest membership but have a global object relationship, which is (are) deviator(s) from that cluster. It is (they are) the initiator(s) for generating a new cluster.

For a tripartition or a higher level of clusterings, use the global object relationship method or the CESS-weighted global maximal SESS method. For a guaranteed global minimal TESS, the deviators are calculated from their memberships or exhaustively searched within a partitional cluster.

3 Results and discussion

The clustering criterion to judge if an ultimately minimal TESS is "converged" from our *K*-means clustering adopts the outcome of executing a brute-force algorithm that computes the TESSs for all possible combinations of data points for a given number of clusters. That is, if one of all the possible TESSs computed is a minimum, it is taken as a reference standard or benchmark for testing if our minimal TESS is the ultimate minimum for a global optimal solution. Moreover, all the real or simulation data sets used for the study were noisier, tougher, and more challenging than general data sets, as they contained hard-to-split pivotal points. This is especially true for a whole- number and/or extremely low dimensional data set. With such ill-clusterable data, they were nevertheless computationally harder to cluster by the K-means algorithms that suffered from apparent errors despite their small sizes as shown in this study. In other words, our successful acquisition of the maximum reduction of the clustering error is effected by all the devised computational laws and global minimizer theories. Through them, all data clustering should result in global minima, which are not necessarily data size-related.

Table 1. A *four*-dimensional (4 variables) real data set of 17 real-number objects (data points) that contains pivotal and close membership points* and has a moderately low dimensionality.

Objects	Variable 1	Variable 2	Variable 3	Variable 4
lau	0.38	626.5	601.3	605.3
ccu	0.18	654.0	647.1	641.8
bhu	0.07	677.2	676.5	670.5
ing	0.09	639.9	640.3	636.0
com	0.19	614.7	617.3	606.2
smm	0.12	670.2	666.0	659.3
bur	0.20	651.1	645.2	643.4
gln	0.41	645.4	645.8	644.8
pvu	0.07	683.5	682.9	674.3
sgu	0.39	648.6	647.8	643.1
abc	0.21	650.4	650.8	643.9
pas	0.24	637.0	636.9	626.5
lan	0.09	641.1	628.8	629.4
plm	0.12	638.0	627.7	628.6
tor	0.11	661.4	659.0	651.8
dow	0.22	646.4	646.2	647.0
lbu	0.33	634.1	632.0	627.8

*The values for Variables 2, 3 and 4 are in close proximity across both objects and variables, which are hard to split from among their data points as they have a poor clusterability for such a data set.

Table 2. A demonstration of the global optimal K-means clustering technique following the pivotal-point theorem and using the *4*-dimensional real data set (*size = 17*) in Table 1 and the global maximum SESS method in our natural, data-driven global partition model.

Bipartition	The first level of clusterings	Data point partitioning	To proceed[¶]	TESS
Cluster 1 = Cluster 2 =	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17 9 ← It has the largest SESS and is a solely unbiased seed.	Point 9 is grabbed and used to build Cluster 2.	Allowed (4165.18)	13043
Cluster 1 = Cluster 2 =	1,2,4,5,6,7,8,10,11,12,13,14,15,16,17 9,3	Point 3 is grabbed and allocated to Cluster 2.	Allowed (3413.59)	9630
Cluster 1 = Cluster 2 =	1,2,4,5,7,8,10,11,12,13,14,15,16,17 9,3,6	Point 6 is grabbed and allocated to Cluster 2.	Allowed (1784.07)	7846
Cluster 1 = Cluster 2 =	1,**2**,4,5,7,8,**10**,**11**,12,13,14,16,17 9,3,6,**15**	Point 15 is grabbed and allocated to Cluster 2.	Stopped (665.09)	7181*
Cluster 1 = Cluster 2 =	1,2,4,5,7,8,10,12,13,14,16,17 9,3,6,15,**11**	Point 11 is assumed to be placed in Cluster 2.	Assumed (-580 .78)	7761
Cluster 1 = Cluster 2 =	1,4,5,7,8,10,12,13,14,16,17 9,3,6,15,**11**,**2**	Point 2 is assumed to be placed in Cluster 2.	Assumed (-280.35)	8042
Cluster 1 = Cluster 2 =	1,4,5,7,8,12,13,14,16,17 9,3,6,15,**11**,**2**,**10**	Point 10 is assumed to be placed in Cluster 2.	Assumed (-139.77)	8181
Cluster 1 = Cluster 2 =	1,4,5,7,8,12,13,14,17 9,3,6,15,11,2,10,16	Point 16 is grabbed and allocated to Cluster 2.	Allowed (131.21[$])	8050
Cluster 1 = Cluster 2 =	1,4,5,8,12,13,14,17 9,3,6,15,11,2,10,16,7	Point 7 is grabbed and allocated to Cluster 2.	Allowed (387.62)	7663
Cluster 1 = Cluster 2 =	1,4,5,12,13,14,17 9,3,6,15,11,2,10,16,7,8 ←The last joined member	Point 8 is grabbed and allocated to Cluster 2.	Stopped (497.45)[§]	7165#

*The first reached TESS (7181) is a local minimum when a regular stopping criterion is early met.
#The second reached TESS (7165) is the global minimum when the pivotal-point theorem applies.
$The smaller and smaller absolute value of the negative difference until turn to positive one (131.21).
§The values in the brackets are the amounts by which the clustering error or TESS is reduced.
¶To proceed with data point partitioning according to the data extraction criteria on the CESS basis.

Table 3. A demonstration of the global optimal K-means clustering technique following the pivotal-point theorem and using the *4*-dimensional real data set (*size = 17*) in Table 1 and the global object relationship method as well as taking the data extraction criteria to proceed with data point partitioning on the CESS basis by which all three clusters are ingeniously, properly and robustly generated.

Tripartition	The second level of clusterings	Data point partitioning	To proceed	TESS
Cluster 1 = Cluster 2 = Cluster 3 =	1,4,5,12,13,14,17 9,3,6,15,11,2,10,16,7 8 ← It is a deviator, and a natural, data-driven initiator.	Point 8 is grabbed and used to generate Cluster 3.	Allowed	6775
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15,11,2,10,16,7 8,4	Point 4 is grabbed and allocated to Cluster 3.	Allowed	6359
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15,11,2,10,7 8,4,16	Point 16 is grabbed and allocated to Cluster 3.	Allowed	5993
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15,11,2,7 8,4,16,10	Point 10 is grabbed and allocated to Cluster 3.	Allowed	5535
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15,11,2 8,4,16,10,7	Point 7 is grabbed and allocated to Cluster 3.	Allowed	4921
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15,11 8,4,16,10,7,2	Point 2 is grabbed and allocated to Cluster 3.	Allowed	4192
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6,15 8,4,16,10,7,2,11	Point 11 is grabbed and allocated to Cluster 3.	Allowed	3162
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,12,13,14,17 9,3,6 8,4,16,10,7,2,11,15	Point 15 is grabbed and allocated to Cluster 3.	Stopped	2959*
Cluster 1 = Cluster 2 = Cluster 3 =	1,5,14 9,3,6 8,4,16,10,7,2,11,15,**12**,**13**,**17**	Points 12, 13, and 17 are assumed to be successively placed in Cluster 3.[$]	Assumed	3417
Cluster 1 = Cluster 2 = Cluster 3 =	1,5 9,3,6 8,4,16,10,7,2,11,15,12,13,17,14	Point 14 is grabbed and allocated to Cluster 3.	Allowed	3048

Cluster 1 =	1,5			
Cluster 2 =	9,3,6,15	Point 15 is grabbed and allocated back to Cluster 2.	Stopped	2858[#]
Cluster 3 =	8,4,16,10,7,2,11,12,13,17,14			

[*]The first reached TESS (2959) is a local minimum when the regular stopping criterion is early met.
[#]The second reached TESS (2858) (same result as the brute-force) is the global minimum when the pivotal-point theorem applies.
[$]Like 11,2,10 in Table 2, they are assumed to be consecutively placed in Cluster 3 without showcasing the process again to save space.

Table 4. A larger-size *2*-dimensional (V) sample data set of 50 whole-number objects (O) (data points) that contains some hard-to-partition pivotal points and has an extremely low dimensionality, and the SESS of each object was computed.

O#	V1	V2	SESS	O#	V1	V2	SESS
1	3	6	0.08	*26*	4	7	2.08
2	4	9	10.88	*27*	5	0	36.88
3	1	6	4.88	*28*	4	4	3.88
4	0	5	10.88	*29*	2	9	11.68
5	2	7	2.88	*30*	0	8	15.08
6	3	8	4.88	*31*	1	2	19.28
7	7	9	24.68	*32*	2	3	9.28
8	5	3	11.08	*33*	3	4	3.28
9	6	4	11.08	*34*	4	5	1.28
10	2	8	6.28	*35*	5	6	3.28
11	3	7	1.48	*36*	0	1	33.28
12	3	5	0.68	*37*	1	5	5.48
13	6	6	7.88	*38*	2	4	4.68
14	4	5	1.28	*39*	3	3	7.88
15	3	9	10.28	*40*	4	2	15.08
16	1	4	8.08	*41*	3	6	0.08
17	2	3	9.28	*42*	4	6	0.68
18	0	6	10.28	*43*	2	5	2.08
19	5	9	13.48	*44*	2	7	2.88
20	3	8	4.88	*45*	4	9	10.88
21	1	7	6.28	***46***	6	0	**41.48**
22	3	9	10.28	*47*	7	10	32.08
23	2	6	1.48	*48*	4	9	10.88
24	5	6	3.28	*49*	6	8	12.68
25	3	4	3.28	*50*	5	8	8.08

Max O# = 46 Max SESS = 41.48 TESS = 473.99

***Proof of pivotal-point theorem*:** A real data set with the typical characteristics of pivotal points was given in Table 1 and the pivotal-point theorem was experimentally verified and justified in Table 2. From there, the data points 11, 2, and 10 were identified as pivotal points (bold ones) that were responsible for early meeting a stopping criterion and locally reached minimum (7181). The reason is that there are close relationships between the data points of 15 and 11, 11 and 2, and 2 and 10; when any couple of such points gets separated, the two clusters to which they belong reach an equilibrium or "dead point". That is, both the points make both the clusters less differential such that their abilities to grab a point from the opposite are balanced. When this equilibrium is broken by assuming a pivotal point to be placed in a target cluster in a given condition, it leads to "convergence" at the global minimum (7165) (same result as the brute-force). Therefore, the highest quality of clusterings is attained with the correct cluster membership {1,4,5,12,13,14,17} and {9,3,6,15,11,2,10, 16,7,8} rather than with {1,2,4,5,7,8,10, 11,12,13,14,16,17} and {9,3,6,15}.

The central idea behind the theorem is making the "less differential" clusters differential. Note that the workings from this proof are universally extensible and applicable to big data as well with no exceptions. This theorem also turns out to be one of our most important findings and data partitioning theories that make a global optimal *K*-means solution possible.

Table 3 exhibited the stepwise process of the second level of clusterings as the number of clusters to be partitioned rose to 3. The partitioning processes of the higher levels of clusterings are analogous to this. When a data set size grows, it works the same without any less optimization and minimization as one may think. This is determined by the globally interactive property and integrity of data structure and object relationships (Tables 4 and 5). Take a notice that the initial cluster memberships used for the tripartition was inherited from the result from the bipartition in Table 2. It should be pointed out that not all the partitioning processes have to exploit the pivotal-point theorem; it is employed wherever necessary. As shown in Table 3, two groups (patterns) of the closest members {9,3,6,15} and {1,5} got isolated (recognized) from clusters, which would make the next-level clustering process much simpler and easier.

Table 5. A demonstration of the properness and robustness of a member object with the global maximum SESS when used as an initial seed to start partitioning of data points as compared to any other objects with the smaller SESS than it.

O#	SESS	TESS	$TESS^2$	O#	SESS	TESS	$TESS^2$
46	**41.48**	474	**431.67**	*7*	24.68	474	448.82
27	36.88	474	436.37	*31*	19.28	474	454.33
36	33.28	474	440.04	*30*	15.08	474	458.61
47	32.08	474	441.27	*40*	15.08	474	458.61

Note: $TESS^2$ is the TESS resulted from exclusion of an object from the initial cluster {1,2,3, • • •,50}, which is inversely proportional to its parental TESS.

As shown in Table 4, the maximum SESS (41.48) of the initial cluster was computable in terms of statistical error that revealed the identity of a member object responsible for the maximum error contribution. As a result of a collection of data points and their integral memberships, this maximum error component is always identifiable whatever a data set size would be. And it is always effective, proper, and robust for that to be used for the maximum reduction of the clustering error when eliminated. In Table 5, only Object 46 with the global maximum SESS (41.48) resulted in the maximum reduction of TESS brought down from 474 to 431.67 when taken off. This was effected by following the global maximal difference law everywhere. Here we only take the maximal difference (41.48) from among all the SESSs, which always brings in TESS minimization whose functionality is unrelated to a data size.

Table 6. A demonstration of the properness and robustness of either the global maximal SESS object or deviators with the global object relationship when used to start partitioning of data points ($CESS^2$ is the reduced CESS).

K	Partitional Cluster	Partition Seeds	CESS	$CESS^2$
2	Initial cluster	46	473.99	431.67
3	Cluster 2	46 or 36,18,4	138.35	80.69
4	Cluster 3	46 or 36,31	80.69	43.09
5	Cluster 3	36 or 14,4,28	48.19	34.27
---	------	------	------	------

Table 7. A *2*-dimensional sample data set of 20 whole-number objects (O) (data points) that contains some hard-to-partition pivotal points* and has an extremely low dimensionality.

O#	V1	V2	O#	V1	V2	O#	V1	V2	O#	V1	V2
1	3	6	*6*	**3**	**8**	*11*	3	7	*16*	1	4
2	4	9	*7*	7	9	*12*	**3**	**5**	*17*	2	3
3	1	6	*8*	**5**	**3**	*13*	6	6	*18*	0	6
4	0	5	*9*	6	4	*14*	4	5	*19*	5	9
5	2	7	*10*	2	8	*15*	3	9	*20*	**3**	**8**

*The bold values are either identical or complementary components and some of the other data points are very close or similar to each other, thus being tougher than real data sets in the partitioning.

Table 8. A comparison of the TESSs reached from the *K*-means clustering of the *2*-dimensional sample data set (*size = 20*) in Table 7 using the brute-force benchmark ($BFB^{\$}$), our global optimal parti- tioning (GOP) and the software $SPSS^{\#}$ solutions.

K	BFB	GOP	SPSS	*K*	BFB	GOP	SPSS
2	86.20	86.20	86.53	*11*	5.00	5.00	6.24
3	50.62	50.62	52.55	*12*	4.00	4.00	4.34
4	33.23	33.23	33.23	*13*	3.16	3.16	3.50
5	24.66	24.66	26.13	*14*	2.50	2.50	3.00
6	17.83	17.83	24.40	*15*	2.00	2.00	2.16
7	13.40	13.40	17.13	*16*	1.50	1.50	1.50
8	9.06	9.06	9.06	*17*	1.00	1.00	1.00
9	7.16	7.16	8.07	*18*	0.50	0.50	0.50
10	6.00	6.00	7.24	*19**	0.00	0.00	0.00

*The data set is supposed to be partitioned into at most 19 clusters because of the two identical objects (data points) that should go in one cluster due to their zero difference.

#Taken from the statistic software SPSS output (using Sum of Error Mean Squares × d.f. and membership information). SPSS implements the typically iterative *K*-means clustering algorithm.

$BFB is used as the compelling evidence of the global minima only, but not as a practical solution, as getting it is prohibitively time-consuming (by the day, week or longer run time, depending on the number of clusters and data size).

Likewise, the deviators are computable for all higher levels of clusterings (K>2). They are available in a global, natural, unbiased, and data-driven setting without artificial operations. That is, they are produced due to the context where they do not belong to any existing clusters, thereby being eliminated and automatically classified as a new sub-cluster. Table 6 gave the 5 top levels of clusterings and their CESSs had maximal amounts of reduction. When all of these workings are fulfilled, a correct new cluster is set up and each level of clusterings has been optimized.

Table 9. A 3-dimensional real-world data set (*size = 16*)*

O#	V1	V2	V3	O#	V1	V2	V3
1	50	50	9	*9*	40	40	5
2	28	9	4	*10*	50	50	9
3	17	15	3	*11*	50	50	5
4	25	40	5	*12*	50	50	9
5	28	40	2	*13*	40	40	9
6	50	50	1	*14*	40	32	17
7	50	40	9	*15*	50	50	9
8	50	40	9	*16*	50	50	1

*The data set contains multiple hard-to-partition repeated data points.

Table 10. A comparison of the TESSs reached from the *K*-means clustering of the *3*-dimensional real-world data set (*size = 16*) in Table 9 using the brute-force benchmark (BFB), our global optimal partitioning (GOP) and the SPSS* solutions.

K	BFB	GOP	SPSS	*K*	BFB	GOP	SPSS
2	1759.33	1759.33	1790.85	*7*	103.85	103.85	103.86
3	743.71	743.71	959.33	*8*	27.66	27.66	29.80
4	422.85	422.85	460.45	*9*	17.00	17.00	18.67
5	286.85	286.85	381.43	*10*	8.00	8.00	10.67
6	182.85	182.85	272.12	*11*	0.00	0.00	0.00

*Taken from the statistical software SPSS output TESS (using Sum of Error Mean Squares × d.f.).

Table 11. A comparison of the TESSs reached from the *K*-means clustering of the *2*-dimensional sample data set (*size = 50*) in Table 4 using the brute-force benchmark (BFB), our global optimal parti- tioning (GOP) and the software SPSS.

K	B&G*	SPSS	*K*	B&G	SPSS	*K*	B&G	SPSS
2	258.644	264.912	*15*	21.249	24.290	*28*	6.499	7.172
3	181.576	181.576	*16*	18.649	20.502	*29*	5.833	6.657
4	140.777	150.512	*17*	16.983	19.338	*30*	5.166	5.680
5	104.386	116.280	*18*	15.483	18.336	*31*	4.500	5.149
6	80.477	81.840	*19*	14.249	17.329	*32*	4.000	4.662
7	67.478	74.261	*20*	13.083	15.900	*33*	3.500	3.995
8	56.477	63.714	*21*	12.083	14.500	*34*	3.000	3.328
9	47.482	51.332	*22*	11.083	11.592	*35*	2.500	2.670
10	40.938	46.520	*23*	10.083	11.718	*36*	2.000	2.156
11	35.714	41.886	*24*	9.250	11.440	*37*	1.500	1.664
12	30.599	35.074	*25*	8.499	9.825	*38*	1.000	1.176
13	27.199	28.897	*26*	7.833	8.160	*39*	0.500	0.500
14	24.166	25.848	*27*	7.166	7.498	*40*	0.000	0.000

*Since BFB equals GOP, we use B&G to stand for their shared values.

The computational results from all the numerical experiments were correspondingly equal between the brute-force and our algorithm. The sample data sets and their globally reached or "converged" results were demonstrated in Tables 7, 8, 9, 10, and 11, respectively.

Table 12. A comparison of the globally optimal values reached by the *K*-means clustering based on the global optimal partitioning (GOP) technology and those obtained from the best-known global or near global solutions* (listed in ascending order of magnitude for the number *K* of clusters partitioned).

Iris Plant			Heart Disease		
K	Known optima	GOP *K*-means	*K*	Known optima	GOP *K*-means
2	152.348	152.3479517603579	***2***	**598900**	598939.9625573959
3	78.851	78.85144142614601	***5***	327970	327542.51402046264
4	57.228	57.228473214285714	***10***	202220	200302.90233807705
5	46.446	46.44618205128205	***15***	147710	146988.31833135852
6	39.040	39.03998724608724	***20***	117780	116877.55985613653
7	34.298	34.298229665071766	***25***	102130	98512.82346652425
8	29.989	29.988943950786055	***30***	88795	85564.07854828662
9	27.786	27.786092417308087	***40***	68645	66402.38570451768
10	25.834	25.834054819972504	***50***	55894	54092.01516688864
Liver Disorders			**Ionosphere**		
K	Known optima	GOP *K*-means	*K*	Known optima	GOP *K*-means
2	423980	423980.8837969465	***2***	2419.4	2419.364807189691
5	218260	218255.95536164998	***5***	1891.5	1889.7165311526685
10	127680	127416.55683351347	***10***	1569.4	1550.0535842095453
15	97474	96756.18094954843	***15***	1401.4	1355.1803607937275
20	81820	80044.73383914215	***20***	1271.4	1213.6066254685106
25	70419	67845.66696431948	***25***	1148.6	1095.1795601937533
30	61143	58967.039226129564	***30***	1046.9	990.6881120696817
40	47832	46582.47583897826	***40***	856.58	815.3967562367158
50	39581	37530.131872894075	***50***	702.58	671.9685317355181

*They are the known theoretical values derived from the cluster function [8,9]. For Iris plant data set, its best-known optima are also the optimal values derived from the cluster function by the known global minimizer [8,9].

For the capability of our global optimal partitioning (GOP) technology that can be extended (generalized) to the scenarios of big data, four well-known benchmark data sets from the UCI machine learning repository (http://archive.ics.uci.edu/ml/machine-learning-databases/) were used to verify the power of the GOP-based *K*-means solutions. In Table 12, Iris data set has 150 instances (objects) of 4 attributes (variables) each. Heart disease data set has 297 instances of the first 13 attributes each. Liver disorders data set has 345 instances of the first 6 attributes each. Ionosphere data set has 351 instances of the first 34 attributes each. Note that Heart disease data set was obtained by removing those instances with a missing attribute from Cleveland raw data. All the above GOP-based optimal values were output from our software ParCluster v.2.0 and were much lower than or equal to the known optima. There is one slight discrepancy in the bold known optima of Heart disease data set as compared to ours. This is beyond explanation since these known optima are theoretical values and may not be the exact ones. To our best knowledge, all reported clustering errors from the multi-start *K*-means, global *K*-means, fast global *K*-means, modified global *K*-means, and efficient global *K*-means algorithms are greater than or far from even the above known optima. Practically, the clustering results from the GOP *K*-means should be treated as real, exact global optima against the theoretical criterion.

The optimized and TESS-minimized clustering result is of great importance in that its correct cluster memberships formed provide accurate information whereas the local- or near-optimal ones give an artifact or reflect wrong information. A cluster membership is very sensitive to clustering errors, which will make a quite difference in combination of objects. This is critical as the misrepresented object relationships could lead to serious consequences and hence is risky.

In particular, this global optimization and minimization would make genes accurately clustered in gene clustering from next-generation sequencing data. This also requires that all the alignment data with base or amino acid sequences be transformed into a similarity data for each of clusterings on which *K*-means clustering is based. A paradigm of its internal data structure is given in Table 13.

Table 13. An alignment data set with genes (G) and bases (B)

O#	B	B	B	B	B	B	B	B	B	B	B	B	---
G1	A	T	G	T	A	C	A	A	A	T	C	A	---
G2	A	T	G	A	A	C	T	G	C	A	G	C	---
G3	A	T	G	A	T	T	A	T	C	A	A	T	---
---	---	---	---	---	---	---	---	---	---	---	---	---	---

Some of the clustering results in the numerical experiments are illustrated in Fig. 2, Fig. 3, and Fig. 4, which are output from our software ParCluster v.2.0.

4 Conclusions and future work

Our global partitioning model underlying the global optimization and exact minimization algorithm provides a milestone approach to true solutions for *K*-means clustering. All computational results in terms of TESS from our algorithmic breakthrough turned out to be the global optima, namely the global minima. We managed to test that there must not be the smaller TESS than what is called minimum not only by exhaustive search but also by the integrated global minimizer theories (separate publication). Our approach proved to produce the lowest value as compared with any other algorithms or software and this difference tended to be greater as the number of clusters and data size became larger (as shown in Tables 8, 10, 11 and 12). Our GOP algorithm also ends up with the unique result or a 100% reproducible cluster membership no matter how many times it operates. These global minimum solutions make it realistic for the *K*-means clustering technique to be reliably and robustly applicable to information processing and data analysis. This gives the confidence that the highest clustering quality or accurate (the most reasonable) cluster membership is achieved for each of clusters. It means using this GOP technology will

make *K*-means clustering results free of risk and no longer controversial. Besides the least TESS that indicates the minimized clustering error, our algorithm has no size issue of a huge data set although it takes more run time with the complexity $O(kn^2m)$. And it also has no initial cluster centers and computational expenses resulted from the iterative two-step refinements in traditional and heuristic solutions. With our proven and mature techniques, the human dream comes true that one is able to classify any type/size of data sets into any number of disjoint clusters with a global minimum TESS. Preferably with the rationale behind GOP, data clustering with the arbitrary number *K* of clusters is capable of producing all the clustering results from *2* up to *K* for a one-time computation. Especially using its resumed clustering feature can inherit previous results and continue with partitioning to the next number of clusters *K* without having to start over from *K* = *2*. It also works well for clusters of arbitrary shape (e.g., non-convex type) and any size, and never gets empty clusters; noisy data and outliers can be isolated from the clustering as soon as possible, thereby eliminating any influence on the final clustering quality [16].

These desired properties/functionalities of nonhierarchical or partitional clustering technique have been expected and pursued since the year 1957 or 1967 or the earliest 1955. The ground-breaking approach and its innovative algorithm we developed put an end to this period of history lacking a global optimal and minimal solution to *K*-means cluster analysis. The Java-based software ParCluster v.2.0 built on the GOP *K*- means technology is available upon request (rli@alumni.lsu.edu) or from some web sites. Please refer to the supplementary material for more details and the pseudocode–based algorithm will be separately published for space reasons.

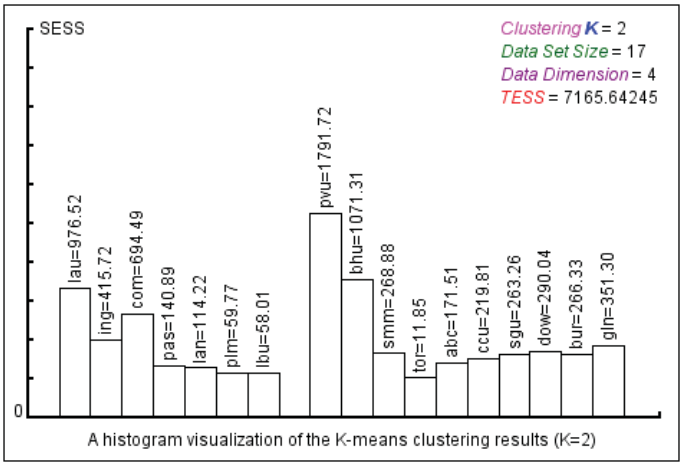


Fig. 2. A graphic view of the clustering result of bipartition (*K*=2) in Table 1 where the highest column of histogram represents the object (pvu) having the global maximum SESS in Cluster 2 (right). The object having the second-to-maximum SESS is bhu. Both obviously contribute the most error to the CESS of Cluster 2 and also to the TESS of all Clusters, making Cluster 2 a "partitional cluster" for the next-level clustering. Furthermore, those members having the similar heights of columns formed and showed closer or compatible relationships.

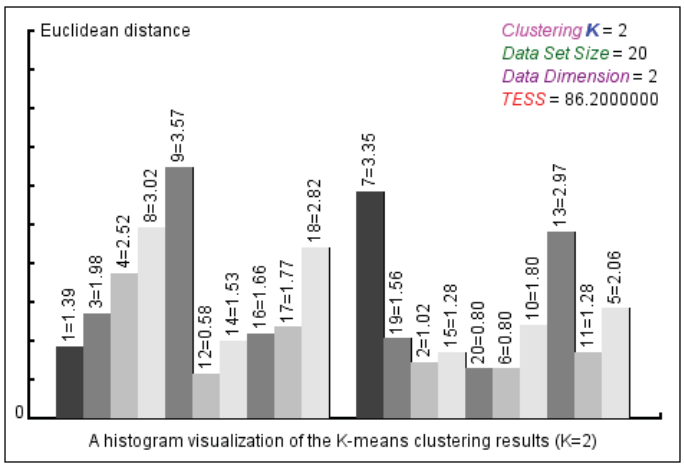


Fig. 3. The Euclidean distances about *K* cluster means were computed and displayed to indicate the logical distance of each object (data point) in the Euclidean space from Table 7. A full and shiftable recognition of the patterns of cluster memberships is enabled, no matter how huge a graph might be. Here Object 9 in Cluster 1 (left) is responsible for the global maximum SESS and is most distant from its cluster center.

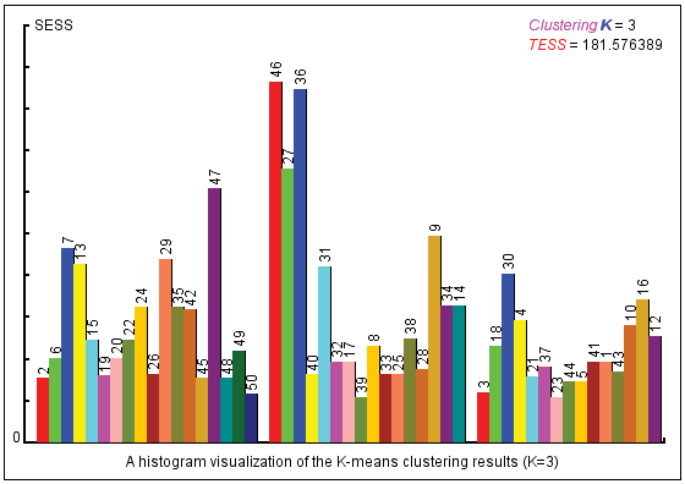


Fig. 4. A full view of tripartition of the *2*-dimensional data set (*size* = *50*) in Table 4, which is partially visible by shifting (moving) the histogram or zoomable. The red column is the first member object for each cluster. The objects 46 and 36 in Cluster 2 (middle) stand out as the globally first and second maximum SESSs respectively, apparently making their cluster a "partitional cluster" for the next-level clustering (*K*=4).

From the scientific viewpoint, it is harder to achieve a global optimality without the cost of computational complexity. Nevertheless, further improving its time complexity (for cost efficiency) and space complexity (for less intermediate data storage) merits consideration if the prerequisite of a resultant global minimum is met.

The next clustering problems to be resolved are what the optimal number of clusters would be and what all the possible different combinations (memberships) of data points would be for a given number of clusters under the same global minimum TESS. This would reveal more information about all of their potential memberships, relationships, and patterns in the identical condition.

5 Supplementary materials

The reader is referred to the on-line Supplementary materials for more details about the GOP *K*-means technology, technical appendices, additional demonstrations, and sample optimal and minimal solutions.

6 Authors' information

Bio of Ruming Li
A visiting professor at Chongqing University, China, received his BS and MS in China and two MSs and PhD in USA in applied statistics and quantitative genetics. Conducting research in bioinformatics, computational biology, genomics, proteomics, phylogenetics, applied mathematics, data structure, (Big) data science and analytics, data / information mining, pattern recognition, numerical analysis, algorithmic solutions and analytics, logic analytics, operations research, computer programming solution, and scientific software development.

Bio of Xiu-Qing Li
The research scientist of molecular genetics at Agriculture and Agri-Food Canada, received his B.S. in China and Doctorate d'Etat in France. Carrying out research in genome biology, bioinformatics, genotyping, molecular breeding, and taxonomic /homologous classification of DNA fingerprints.

Bio of Guixue Wang
The professor and Dean at Bioengineering College of Chongqing University, China, received his B.S., M.S., and doctorate in China. Engaging in teaching and research in biotechnology, bioinformatics, quantitative genetics, cellular and molecular bioengineering, cardiovascular biomechanics and biomaterials, biorheological science, and mechano-developmental biology.

7 Acknowledgements

This research was supported in part by Projects of China (2009ZX08009-109B) and National 111 Project (B06023).

8 References

[1] Lloyd, S. P. Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory **28**, 128–137, 1957, 1982.

[2] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. (Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Univ. of California Press), **1**, 281-297, 1967.

[3] Hartigan, J. A. Clustering Algorithms. (Wiley, New York, ed. 1), 1975.

[4] Jain, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. **31** (8) 651-666, 2010.

[5] Hamerly, G. & Elkan, C. Alternatives to the k-means algorithm that find better clusterings. (Proceedings of the 11th international conference on Information and knowledge management), 600-607, 2002.

[6] Kanungo, T. , et al. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transac- tions on Pattern Analysis and Machine Intelligence, **24**, 881-892, 2002.

[7] Likas, A., Vlassis, N. & Verbeek, J. J. The global k-means clustering algorithm. Pattern Recognition **36**, 451-461, 2003.

[8] Hansen, P., Ngai, E., Cheung, B. K. & Mladenovic, N. Analysis of global k-means, an incremental heuristic for minimum sum-of-squares clustering. J. of Classification **22**, 287-310, 2005.

[9] Bagirov, A. M. Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition **41**, 3192-3199, 2008.

[10] Daniel Aloise, Pierre Hansen, Leo Liberti An improved column generation algorithm for minimum sum-of-squares clustering. Mathematical Programming **131**(1-2) 195-220, 2012.

[11] Bakar, Z. A., Deris, M. M. & A. C. Alhadi, Performance analysis of partitional and incremental clustering. (SNATI 2005), ISBN: 979-756-061-6, 2005.

[12] Wilkin, G. A. & Huang, X. K-means clustering algorithms: Implementation and comparison. (2nd IMSCCS), 133-136, 2007.

[13] Chakraborty, S. & Nagwani, N. K. Analysis and study of incremental K-means clustering algorithm. (HPAGC, CCIS), **169**, 338-341, 2011.

[14] Kumar, A., Sinha, R., Bhattacherjee, V., Verma, D. S. & Singh, S. Modeling using K-means clustering algorithm. (1st Int'l Conf. on Recent Advances in Information Technology), 554-558, 2012.

[15] Charikar, M., Chekuri, C., Feder, T. & Motwani, R. Incremental clustering and dynamic information retrieval. SIAM J. Comput. **33**, 1417-1440, 2004.

[16] Mimaroglu, S. and Erdil, E. Obtaining better quality final clustering by merging a collection of clusterings. Bioinformatics **26** (20) 2645-2646, 2010.

Fuzzy Knowledge-based Image Annotation Refinement

M. Ivašić-Kos[1], M. Pobar[1] and S. Ribarić[2]
[1]Department of Informatics, University of Rijeka, R. Matejčić 2, 51000 Rijeka, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia

Abstract - *Automatic image annotation methods automatically assign labels to images in order to facilitate tasks such as image retrieval. Incorrect labels may negatively influence the search results so image annotation should be as accurate as possible.*

Labels pertaining to objects or to whole scenes are commonly used for image annotation, and precision is especially important in case when scene labels are inferred from objects, as errors in the object labels may propagate to the scene level. To improve the annotation precision, the idea is to infer which labels are incorrect using the context of other labels and the knowledge about objects and their relations. This procedure is here referred to as annotation refinement. The proposed approach used in this paper includes a fuzzy knowledge base and uses the fuzzy inference algorithms to detect and discard automatically obtained object labels that do not fit the context of other detected objects.

Keywords: automatic image annotation, annotation refinement, fuzzy knowledge representation scheme, fuzzy inference

1 Introduction

Automatic image annotation methods automatically assign labels from a predefined vocabulary to an unlabeled image to facilitate image search and retrieval. The goal is to bridge the so-called semantic gap [1] between the available features that can be extracted from the raw image data, such as color, texture, structure, etc. and the appropriate labels that can be used for retrieval.

When searching for images, object or scene labels are typically used, so the vocabularies of automatic image annotation systems contain labels that correspond to objects like *skyscraper*, *fox*, *flowers*, *airplane*, etc. or to whole scenes like *airfield*, *beach* or more general *outdoors*.

In general, a scene may be very complex and be composed of many different objects, so it should also be annotated with many object labels. The automatic recognition of objects depends on the method used for automatic annotation and the features extracted from images. Scene labels can be inferred from object labels if it is assumed that scenes are compositions of objects, as in [2]. Common background objects like *sky*, *grass*, etc., may appear in a number of different scenes while some objects are typical for only one scene, e.g. *train* for scene *railway*. Correctly detected typical objects greatly help inferring the corresponding scenes, however if an object is annotated with a label of a different typical object, it is very likely that the wrong scene label will be inferred and image will be incorrectly annotated. In scenes without a typical object (e.g. *Mountain*) even background objects could play an important role for image interpretation and scene inference. Image annotation on object level should therefore be as precise as possible regardless of whether the scene has a typical object or not so that misclassified objects don't negatively influence the image interpretation.

The image annotation task is closely related to the problem of classification or the problem of representing the correlations between images and labels, so the most of the automatic image annotation approaches proposed so far belong to the field of machine learning. The methods based on classification, like [3], classify images or image segments into predefined classes based on low-level features extracted from images. Probabilistic methods such as those based on the translation model [4] or on latent semantic analysis [5] learn relevance models to represent the correlations between images and labels. A recent survey of research made in that field can be found in [6, 7].

Graph-based image analysis algorithms have been proposed recently for exploring the correlations between annotated labels [8-14]. To detect and discard the irrelevant labels, a WordNet based semantic similarity is used in [14]. A graphical model is used in [11] for fusing visual content represented by a nearest spanning chain and label correlation by WordNet. A group of authors has examined several different approaches for annotation refinement: Markov chains in [10], conditional random fields (CRF) in [8], normalized Google distance (NGD) in [14] and re-ranking the annotations using the random walk with restarts algorithm in [9].

Incorporating knowledge into automatic image annotation procedure proved as a promising approach for improvement of annotation efficiency. Such an approach was proposed in [15] where a knowledge base and inference algorithms are incorporated into automatic image annotation system for multi-level image annotation.

In this paper we propose a pipeline for automatic image annotation with a knowledge based annotation refinement step, presented in Section 2. The refinement step uses a novel procedure for annotation refinement using fuzzy knowledge representation scheme and fuzzy inference algorithms, described in Section 3. The procedure is described with

examples from the outdoor image domain. Concluding remarks are given in Section 4.

2 Automatic annotation pipeline

The proposed pipeline for automatic image annotation system is shown in Fig. 1. The main stages are image segmentation, feature extraction, segment classification, aggregation of segment labels for image annotation and annotation refinement. It is assumed that an unlabeled image is input to the automatic annotation system. The outputs of the system are labels that are refined using our proposed algorithm.

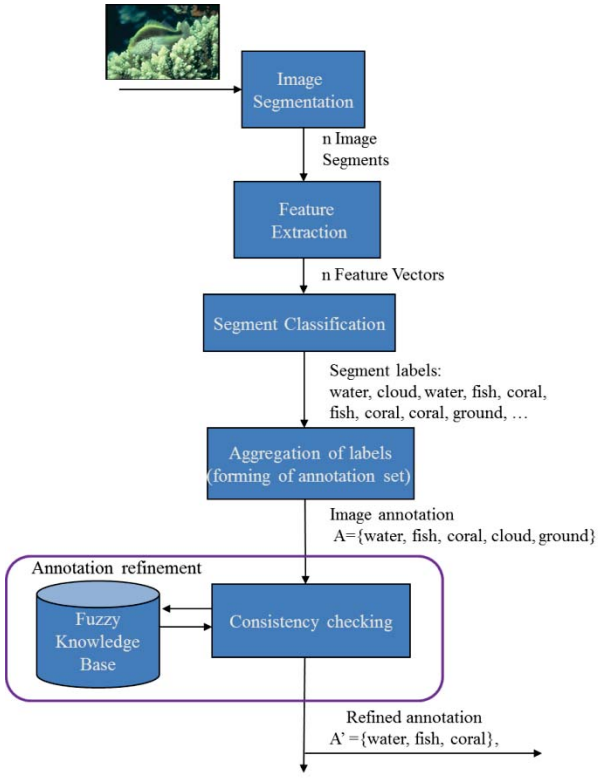


Figure 1. Automatic image annotation pipeline with annotation refinement

The image is first automatically segmented using the n-cuts algorithm [16] and low-level features such as color, texture, etc. are extracted from each segment. Each segment is represented with a feature vector and then classified into one of the predefined labels $d_i \epsilon D$. It is likely that one label will appear more than once because of actual multiple appearance of an object on the scene or because an object can be split into more than one segment. From the obtained segment labels the annotation set $A(e)$ is formed for an image e. Next, for each label d in the set A, a value γ(d) is set depending on the used classifier. If the classifier gives a confidence value for each label, then γ(d) is set to that value. Otherwise, the value γ(d) is set to 1. For the unlabeled image e in Fig. 2 the annotation set $A(e) = \{cloud, water, fish, coral, ground\}$ is formed from the segment labels *water, cloud, water, fish, coral, fish, coral, coral, ground, ….* Note that the labels *ground* and *cloud* are a result of misclassification because they are not present in the image.

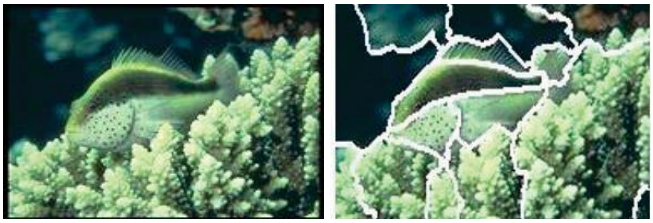


Figure 2. Example of an original unlabeled image and its segmented image

Most automatic annotation systems stop at this stage, but because some labels may be the result of misclassification, an additional annotation refinement stage is proposed here. The goal of this stage is to automatically detect and remove the misclassified labels. The procedure for annotation refinement is based on consistency checking and relies on the fuzzy knowledge base. In the fuzzy knowledge base the facts about objects in the domain of interest and their relations are defined. The facts from the knowledge base are used to select a subset A' of the initial annotation set A that satisfies the defined consistency rule. The set A' is used as the final annotation set and the rest of labels are discarded.

3 Annotation refinement

The aim is to annotate an image as precisely as possible, so additional knowledge should be exploited to detect and discard the misclassified labels. Incorrectly detected objects, both typical for a scene or background, can lead to misclassification of the corresponding scenes. Measures of precision and recall are used to evaluate image annotation. Precision of an annotation set *A(e)* for an image e is here defined as $Precision = \frac{r}{r+w}$, where r is the number of correct labels and the sum r+w is the number of all labels in the set. Recall is defined as $recall = \frac{r}{n}$, where n is the number of labels in the reference annotation set *R(e)* for the image e. Misclassification occurs commonly since different objects can have very similar values of low-level features extracted from images and therefore the classification models cannot be learned well. To reduce the influence of misclassification, the additional knowledge about the domain represented in the knowledge base may be used to check the consistency of labels with respect to the context of the image and the relations among objects.

3.1 A Fuzzy Knowledge-representation Scheme for Annotation Refinement

The knowledge about the objects that may appear in images is usually incomplete and uncertain, so a suitable knowledge-representation scheme is required. Here, a knowledge-representation scheme based on Fuzzy Petri Nets (KRFPN [17]), that supports inference from fuzzy knowledge is

modified and referred to as KRFPNr. The scheme is used to represent the facts in the knowledge base used for annotation refinement. The knowledge base includes pseudo-spatial, spatial and attribute relations between objects in outdoor scenes learned from the data in the training set.

The *KRFPNr* scheme is formally defined as the 12-tuple:

$$KRFPNr = (P, T, I, O, M, \Omega, f, c, \alpha, \beta, D, \Sigma), \qquad (1)$$

elements of which are described below.

$P = \{p_1, p_2, \ldots, p_n\}, n \in N$ is a set of places and $T = \{t_1, t_2, \ldots, t_m\}, m \in N$ is a set of transitions.

Each transition is associated with at least one input and output place by the input and output functions, $I: T \to P(P)\backslash\emptyset$ and $O: T \to P(P)\backslash\emptyset$.

A place may be marked with one or more tokens from the set $M = \{m_1, m_2, \ldots, m_l\}, l \geq 0$. The tokens are used to define execution of a Fuzzy Petri Net (FPN). Their distribution within places for each execution step *w=0,1,...* is given as $\Omega_w(p) \in P(M)$, where $P(M)$ is a power set of M. A place p is marked in step w if $\Omega_w(p) \neq \emptyset$. In our case, in the initial marking each place can have at most one token. A place that contains one or more tokens is called a marked place and it is important for execution of the transition.

To each place from the set P, a concept d from the set D is assigned by the bijective function $\alpha: P \to D$. The concepts from the set D may be object labels used for image annotation, e.g. *sky*, *whale*, *lion* or properties of objects like their color (*blue, orange,...*), position (*middle, left,…*), etc.

To each transition from the set T, a relationship r from the set Σ is assigned by the function $\beta: T \to \Sigma$. The set Σ contains all the relationships defined between concepts in the scheme, e.g. *occurs_with*, *is_near*, *consists_of*.

The uncertainty and confidence related to the concepts and the relationships between them are expressed by the values of the association functions $f(t_i), t_i \in T$, and $c(m_i), m_i \in M$. The degree of truth of the relationship mapped to a transition is given by the transition value $f: T \to [0, 1]$. The degree of truth of the concept mapped to the marked place is represented by the token value $c: M \to [0, 1]$. The transition values and the initial token values are defined according to the used training dataset.

3.1.1 Modelling the degree of truth of relations

The transition value represents the degree of truth of the relationship associated to the transition. In the KRFPNr scheme, three types of relationships are defined and included in the set Σ: the pseudo-spatial relationship, the spatial and the attribute relationship.

The $occurs\_with \in \Sigma$ relationship is a pseudo-spatial relationship that describes a mutual occurrence of two objects in the same scene. The reliability of the *occurs_with* relation between objects d_i and d_j, $d_i, d_j \in D$, is denoted as $d_j\ occurs\_with\ d_i$ and it is calculated based on the conditional probability of occurrence of object d_j, when object d_i is present on the scene. It represents the reliability that the object d_j appears on the scene along with the object d_i. It is assumed that the appearance of objects are independent events, i.e. that the appearance of one object in the scene does not depend on the occurrence of the other. Using the data in the training set, the reliability f_r of the relationship $d_j\ occurs\_with\ d_i$ is computed as:

$$f_r(d_j\ occurs\_with\ d_i) = P(d_j|d_i) = \frac{P(d_i \cap d_j)}{P(d_i)} \cong \frac{f_{ij} + mp_j}{f_i + m}, \quad i \neq j. \qquad (2)$$

The probabilities $P(d_j \cap d_i)$ and $P(d_i)$ are estimated from the data set using empirical relative frequencies of object occurrence. The joint probability $P(d_j \cap d_i)$ is estimated as the relative frequency f_{ij} of mutual occurrence of both objects d_j and d_i on the scenes in the training set, while the prior probability $P(d_i)$ is estimated as the empirical relative frequency f_i of occurrence of object d_i in the training set. An m – estimate is implemented [18] so an estimate of reliability can be obtained in cases when there are no examples of mutual occurrence in the training data set. The size of a virtual set of samples m is s chosen experimentally, and p_j is the estimated probability that a sample is the object d_j.

The truth value of the relationship $d_j\ occurs\_with\ d_i$ is generally not equal to the truth value of relationship $d_i\ occurs\_with\ d_j$ For example, the probability of appearance of *sky* is higher when *airplane* exists on the scene then the appearance of *airplane* if *sky* is on the scene, because *airplane* in most cases is in the *sky* but *sky* can often be without an *airplane* and occur with a large number of different objects like *trees*, *lion*, *train*, ...

The $occurs\_with$ relationship is used to check the consistency of segment classification. The truth value of the relation $d_i\ occurs\_with\ d_j$ is set to the transition value of the transition between places that correspond to dj and di and to which *occurs_with* relation is assigned:

$$f(t_j) = f_r(d_j\ occurs_{with} d_i), t_j : \beta(t_j) = occurs\_with, d_i \in I(t_j), d_j \in O(t_j) \qquad (3)$$

Spatial relationships such as *at_the_top* and *next_to* specify in our scheme the position or relative position of an object in the scene, and are used for consistency checking. Other types of spatial relationships such as topological, distance and internal relations [19], or any new concept or relation can easily be added to the scheme.

The truth values of spatial relations are computed using empirical relative frequencies of objects and their spatial positions with m-estimate, in a similar fashion as for computation of reliability of *occurs_with* relations. The transition value assigned to a spatial relation is set to the truth value of that relation.

3.1.2 Fuzzy inheritance

The fuzzy inheritance algorithm is used in the process of consistency checking for automatic annotation refinement.

Input:
Concept $d_i \in D$, depth of inheritance k.
Facts in the knowledge base
Output:
Properties of the concept d_i and the properties of concepts that lie at higher levels of hierarchical structure of concepts.
Steps:
For the given concept $d_i \in D$, find the corresponding place $p_k \in P$ using the function α^{-1}: $\alpha^{-1}(d_i) = p_k$.
Define the initial distribution of tokens $\Omega_0(p_k) = \{m_1\}, \Omega_0(p_j) = \emptyset, j \neq k$ and set $c(m_1) = \gamma(d_i)$.
For the distribution Ω_0 construct k levels of the inheritance tree.
Collect the leaf nodes.

Figure 3. The fuzzy inheritance algorithm

All the steps of the inheritance algorithm for the KRFPN are given in detail in [17], and are valid for the KRFPNr as well.The algorithm is based on the inheritance set of the KRFPNr, which is a concept derived from the reachability set of the ordinary Petri nets. The reachability set is defined as the smallest set of all reachable distributions of tokens, starting from an initial distribution and recursively applying the firing of enabled transitions to obtain the immediately reachable distribution of tokens [20]. The enabled transitions are fired in discrete steps in which new token distributions and token values are determined. After the transition has fired at a step w, a new token value $c(m_{x+1})$ is obtained at the output place as $c_j f(t_j)$ where $c_j = max_x\, c(m_x): m_x \in \Omega_w(p_i), p_i \in I(t_j)$ and $f(t_j)$ is the value of the transition t_j. In other words, the token with the maximum value at the input place of the transition together with the transition value determine the value of the token in the output place.

A transition t_j is enabled when every input place of the transition is marked: $|\Omega_w(p_i)| \geq 1, \forall p_i \in I(t_j)$. A place is marked if it contains one or more tokens. When a transition t_j fires, tokens simultaneously "move" from all the transition's input places $p_i \in I(t_j)$ to the output places $p_l \in O(t_j)$.

The inheritance set is represented with a fuzzy inheritance tree, while the reachability set is represented by a reachability tree. The main difference between the reachability set and the inheritance set of the *KRFPNr* are related to the semantic interpretation of places. Namely, to stop further firing of transitions at the places that represent the end of the hierarchical structure among concepts, the corresponding nodes have to be frozen. A node is frozen if it is an output node of a transition associated with attribute and spatial relationships. The generation of the inheritance trees may stop on frozen (F) nodes, on a predefined level k (k-terminal nodes, k-T), in which case a k-level inheritance tree is generated, or on terminal (T) nodes. A terminal node is a node with no enabled transitions.

The root nodes ${\pi_0}^i, i = 1,2, \ldots$ of the inheritance trees are formed according to the initially marked places and their corresponding truth values. The nodes of inheritance trees have the form $\pi_{\kappa i}(p_j, c(m_l))\ j = 1, 2, \ldots, p,\ \ l = 1, 2, \ldots, r,$ $0 \leq r \leq |M|$, where the first component specifies the place $p_j$where the token m_l is located and the second one is the token value, $c(m_l)$, κ represents the level of the tree, and i is the node index at that level.

The steps of the fuzzy inheritance algorithm used for consistency checking are shown in Fig 3.:

3.2 Consistency checking

The consistency of labels in set *A* obtained after classification of image segments is checked using the facts from the knowledge base. Among the labels in the annotation set *A*, there can be correct labels but also some labels that are a result of misclassification. We can assume that correct labels are consistent among themselves with respect to the *occurs_with* relationship defined in the knowledge base. If there is more than one misclassified label, it is possible that by chance they are also consistent among themselves. Thus, there may be more than one subset of mutually consistent labels in the annotation set *A* from which only one will be selected for annotation. In order to detect the misclassified labels, the proposed consistency checking algorithm identifies all subsets of mutually consistent labels in the annotation set. Taking into account both the size of such subsets and the confidence values *γ(d)* of each object *d*, the subset *A'* of labels with maximum RV value (Fig. 4) is selected for annotation. The remaining labels in the set *A\A'* are considered misclassified and are discarded.

The consistency checking procedure is formally defined in the Fig. 4

Input:
Initial annotation set A
Steps:
for each label d in A:
generate inheritance tree $\pi(d)$
add d and leaf nodes to a set $R(d)$
define a power set $P(A)$
initialization $A' := \emptyset,\ r := 0$
for each $S \in P(A)$:
compute $V := \bigcap_{d \in S} R(d)$
if $S \subseteq V$
compute $RV(S) := \sqrt{\frac{1}{|S|}\sum_{d \in S}(\gamma(d) + \varepsilon)^2}, \varepsilon <<$
if $RV(S) > r$
set $r := RV(S)$
set $A' := S$
Output:
The set A' of all consistent labels from A

Figure 4. Consistency checking procedure.

First, for each label d in the annotation set A an inheritance tree $\pi(d)$ is generated using the fuzzy-inheritance algorithm and the facts in the knowledge base. The inheritance trees are used to verify whether an $occurs\_with$ relationship is defined between that object and other obtained object labels. The leaf nodes of the tree $\pi(d)$correspond to labels that may occur together with the object d. For each object label d in the annotation set A, a set $R(d)$of possible objects that may occur in an image together with that object is defined using the leaf nodes of the inheritance tree $\pi(d)$.

In the next step, all subsets of A, i.e. elements of the power set $P(A)$that are consistent with respect to the *occurs_with* relationships are identified. A set $S \in P(A)$of labels is consistent if each label in the set has the *occurs_with* relation defined between it and all other labels in the set. This is equivalent to the condition that a set S is consistent if it is a subset of the intersection of sets of possible labels *R(d)* for each label d in S, $S \subseteq V, V = \bigcap_{d \in S} R(d)$.

For each consistent subset of labels, a RV value is computed that considers the size of the set and confidence values $\gamma(d)$ of each label in the set S. The set with the highest RV value is selected as the consistent set A' and used for annotation.

For instance let the image *e* on the Fig. 2. be considered for consistency checking. Then for the object *ground*, the appropriate place in the knowledge-representation scheme is determined by the function $\alpha^{-1}(ground) = p_{13},\ ground \in C$, and a token m_1 is placed in place p_{13}. According to the initially marked place, the initial token distribution is created $\Omega_0(p_{13}) = \{m_1\}, \Omega_0(p_j) = \emptyset, j \neq 13$. The corresponding root node of the inheritance tree is $\pi_0(p_{13}, \{\gamma(ground)\})$. The inheritance tree is formed by firing the enabled transitions (whose firing creates new nodes) until the condition for stopping the algorithm is satisfied or the desired depth *k* of the inheritance tree is reached. Figure 5 shows the 1-level inheritance tree on the *KRFPNr* scheme for the object *ground*.

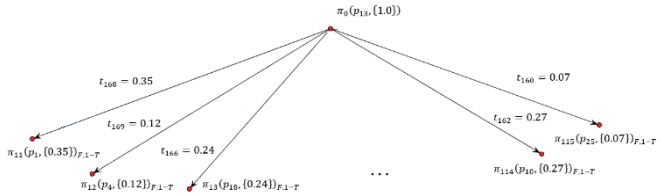


Figure 5. 1-level inheritance tree for the object *ground* detected as possible intruder in automatic annotation of an image *e*.

The truth value of the *occurs_with* relation between the root node and all other objects is determined by the token value in leaf nodes (the nodes in which the algorithm stops). The arcs of the inheritance tree are labeled with transitions $t_j \in T$ and values (t_j). For example, the arc t_{395} corresponds to the relation *occurs_with* between the nodes π_0 and π_{12}. The node π_0 corresponds to the object $ground$ at the place p_{13} and the node π_{12} to the object *water* a the place p_{25}. The truth value of the relation $f_r(ground\ occurs\_with\ water)$ is 0.07.

The inheritance tree has stopped on output nodes of the transitions, marked as frozen leaf nodes (F). Because the depth of the inheritance tree was 1, the leaf nodes are also marked as 1-terminal (1-T). Leaf nodes of inheritance tree include all objects from the knowledge base that are in relation *occurs_with* with object *ground*. The leaf nodes include places that function α maps to objects such as $\alpha(p_1) = sky, \alpha(p_4) = lion, \alpha(p_{10}) = grass,\ \ldots\ \alpha(p_{18}) = rock, \alpha(p_{25}) = water$ that make the codomain of relation *occurs_with* for a given object *ground*. These objects form the set R(*ground*), R(*ground*) = {*sky, lion, grass, ..., rock, water*}.

The inheritance trees are also formed for all other objects in the set $A(e) = \{water, fish, coral, cloud, ground\}$. For example, in Fig. 6 is the inheritance tree with the object *coral* as the root node and with all the objects from the knowledge base that are in relation *occurs_with* with the object *coral* in leaf nodes. After applying the function α on the places at the leaf nodes, $\alpha(p_{25}) = water, \alpha(p_{26}) = fish, \alpha(p_{17}) = sand, \alpha(p_{18}) = rock$ the codomain of relation *occurs_with* for a given object *coral is* formed as the set R(*coral*)={water, fish, sand, rock}.

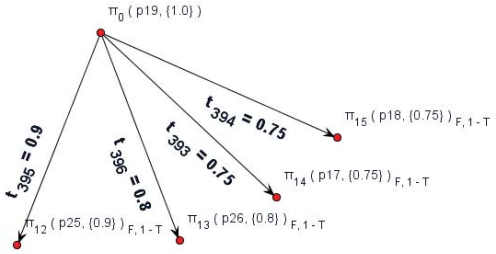


Figure 6. Inheritance tree for the object *coral.*

The intersection of R(*coral*) and R(*ground*) contains neither ground or coral so they cannot appear together. By running the rest of the algorithm for each subset of the power set $P(A)$, the sets {}, {*water*}, {*fish*},{*coral*},{*cloud*},...,{*water*, *fish*}, … , $\{water, fish, coral\}$ are identified as consistent. The subset with the greatest RV value is chosen as the consistent set G=$\{water, fish, coral\}$.

If the chosen set A' contains correctly classified labels, the average precision of the image annotation can be increased, Fig. 7a,b. However, if the majority of labels are misclassified, a set of wrong but mutually consistent labels might be chosen. In that case the annotation refinement can discard the correct labels and both recall and precision falls, Fig. 7c. In the Fig. 7, incorrect labels are printed bold.

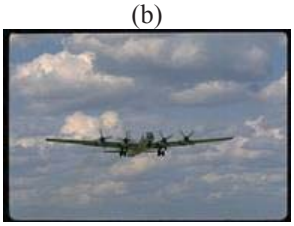
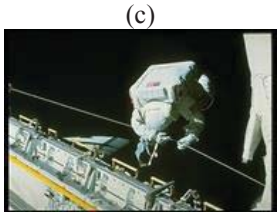
	Image example *e*	(a) 	(b) 	(c) 
Reference annotation	R(e)	*sky, wolf, trees, grass*	*airplane, sky, cloud*	*shuttle, astronaut*
Automatic annotation	$A(e)$	***coral***, *sky, wolf, trees, grass*	*airplane,* ***dolphin***, *sky, cloud*	*shuttle,* ***train, building***
	$Precision\ Pr = \frac{r}{r+w}$, $Recall\ Re = \frac{r}{n}$	$Pr = \frac{4}{4+1} = 0.8, Re = \frac{4}{4} = 1$	$Pr = \frac{3}{3+1} = 0.75, Re = \frac{3}{3} = 1$	$Pr = \frac{1}{1+2} = 0.33, Re = \frac{1}{2} = 0.5$
Annotation after refinement	$A'(e)$	*sky, wolf, trees, grass*	*airplane, sky, cloud*	***train, building***
	$Precision\ Pr = \frac{r}{r+w}$, $Recall\ Re = \frac{r}{n}$	$Pr = \frac{4}{4+0} = 1, Re = \frac{4}{4} = 1$	$Pr = \frac{3}{3+0} = 1, Re = \frac{3}{3} = 1$	$Pr = \frac{0}{0+2} = 0, Re = \frac{0}{2} = 0$

Figure 7. Positive (a), (b) and negative (c) example of annotation refinement.

4 Conclusion

In this paper, a knowledge based procedure is proposed for refinement of labels obtained with automatic image annotation. Knowledge about objects and their relationships in the domain of interest is exploited to find consistent sets of labels among the results of automatic annotation and to discard the inconsistent labels. To represent the knowledge about objects and their relationships, a fuzzy knowledge representation formalism, dubbed *KRFPNr* is used. Vocabularies for automatic image annotation usually contain labels that correspond to objects (*building*, *bear*, *train*, etc.) and scenes (*forest*, *underwater*, *nature*, etc.) because these kinds of terms are typically used when searching for images.

If it is assumed that scenes are compositions of objects, then scene labels can be inferred from object labels. In this case, image annotation on object level should be as precise as possible to avoid the case where a misclassified object leads to wrong conclusion about the whole scene.

Increased precision can be achieved by including an image annotation refinement step in the annotation process. The approach used in the proposed procedure checks the consistency of each object label obtained with classification with respect to the most likely image context deduced from the rest of the labels. Detected inconsistent labels are assumed to be results of misclassification and are not used for annotation, thereby increasing precision. . The consistency checking procedure relies on the facts about objects and their relations stored in the knowledge base and the fuzzy inheritance algorithm. The functioning of algorithms is demonstrated on examples of images of outdoor scenes.

In the future work, instead of discarding the detected inconsistent labels, inference algorithms and facts from the knowledge base willl be adapted for finding suitable replacement labels that are useful for annotation.

5 References

[1] Hare JS, Lewis PH, Enser PGB Sandom CJ. 2006 January 17-19. Mind the Gap: Another look at the problem of the semantic gap in image retrieval. Multimedia Content Analysis, Management and Retrieval, San Jose, California, USA.

[2] Ivašić-Kos, M.; Ribarić, S.; Ipšić, I. Low- and High-level Image Annotation Using Fuzzy Petri Net Knowledge Representation Scheme. International Journal of Computer Information Systems and Industrial Management (IJCISIM). 4 (2012)

[3] J. Li and J. Z. Wang, ``Real-Time Computerized Annotation of Pictures," IEEE Transactions on Pattern

Analysis and Machine Intelligence, vol. 30, 2008, pp. 985-1002.

[4] Duygulu, P., Barnard, K., Freitas, J.F.G. de, Forsyth, D. A., 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, ECCV 2002, UK, 2002, pp. 97–112.

[5] Monay F. and Gatica-Perez D., "On image auto-annotation with Latent Space Models", Proc. ACM Multimedia, Berkeley, CA, 2003, pp. 275–278.

[6] Datta, R., Joshi, D., Li, J. 2008. "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Transactions on Computing Surveys, vol. 20, pp. 1-60, April 2008.

[7] Zhang, D., Islam, M. M., and Lu, G. (2012). A review on automatic image annotation tecniques. Pattern Recognition, 45(1), 346-362.

[8] Wang, Y., Gong, S. (2007). Refining image annotation using contextual relations between words. ACM CIVR 07', July 9–11, Nethelands, 425–432.

[9] Wang, C, Jing, F., Zhang, L., & Zhang, H.-J. (2006). Image annotation refinement using random walk with restarts. ACM MM 06', October 23–27, Santa Barbara, California, USA. 647–650

[10] Wang, C., Jing, F., Zhang, L., Zhang, H.-J. Content based image annotation refinement. CVPR'07.

[11] Liu, J., Li, M., Ma, W.-Y., Liu, Q., & Lu, H. (2006). An adaptive graph model for automatic image annotation. ACM MIR 06', Santa Barbara, California, USA, October 26-27, 61–70.

[12] Zhou, X., Wang, M., Zhang, Q., Zhang, J., Shi, B. (2007). Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. ACM CIVR 07', July 9–11, Amsterdam, Nethelands, 425–432.

[13] Yohan, J., Khan, L., Wang, L., & Awad, M. (2005) Image annotations by combining multiple evidence and WordNet. ACM conference on Multimedia (MM05'), Singapore, 706–715.

[14] Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance.Knowledge and Data Engineering, IEEE Transactions on, 19(3), 370-383.

[15] Ivašić-Kos, M.; Ipšić, I.; Ribarić, S. Multi-Level Image Annotation Using Bayes Classifier and Fuzzy Knowledge Representation Scheme, WSEAS transactions on computers. 13 (2014); 635-644

[16] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(8), 888-905.

[17] Ribarić, S., Pavešić, N., 2009. "Inference Procedures for Fuzzy Knowledge Representation Scheme", Applied Artificial Intelligence, vol. 23, January 2009, pp. 16-43

[18] Džeroski, S., Cestnik, B. and Petrovski, I., 1993. Using the m-estimate in rule induction. J. Comput. Inf. Technol, 1: 37-46

[19] Isabelle Bloch, Fuzzy spatial relationships for image processing and interpretation: a review, Image and Vision Computing, vol. 23(2), 2005, pp 89-110.

[20] Chen, S.M., Ke, J.S., Chang, J.F., 1990. "Knowledge Representation Using Fuzzy Petri Nets", IEEE Transactions on Knowledge and Data Engineering, vol. 2, 1990, pp. 311-3

A Multi-Sensor Information Fusion Approach for Efficient 3D Reconstruction in Smart Phone

Apurbaa Mallik[1], **Brojeshwar Bhowmick**[2] **and Shahnawaz Alam**[3]
Tata Consultancy Services, Kolkata, India
[1]apurbaa.mallik@tcs.com, [2]b.bhowmick@tcs.com,[3]shahnawaz.alam@tcs.com

Abstract— *Despite the recent advancements, camera pose estimation using smart phone's sensor data is still error-prone due to various environmental noise and variability of the force applied on the phone during data acquisition. Here, we propose a novel framework to mitigate these drawbacks of camera pose estimation using various sensor's data. The directional information of accelerometer is utilized to obtain reliable features for position estimation, rather than using its magnitude. Moreover, a multi-sensor data fusion approach is followed for robust rotation estimation. The proposed framework improves the epipolar geometry estimation and produces accurate stereo-correspondence points. This in turn reduces the reprojection error and enhances the quality of 3D reconstruction. Furthermore, the proposed framework utilizes only the CPU of the phone, instead of GPU and takes around 2-3 secs for processing an image. Extensive experimental results show supportive evidence of the effectiveness of the proposed framework in pose estimation problem.*

Keywords: 3D reconstruction, sensor data, pose estimation

1. Introduction

One of the fundamental problem in computer vision is to reconstruct a complete 3D scene from a set of multi-view 2D images. Multi-view 3D reconstruction systems have found applications in various domains including robotics, architecture, archaeology, surveillance, human computer interaction and entertainment industry etc. In this regard, the primary challenge is to find way of effectively fusing information from multiple views without a prior knowing the detailed calibration information and minimizing human intervention. In the last few decades, dense 3D reconstruction from multi-view has been explored widely by various researchers and numerous approaches have been proposed in the literature. One of the most important step in estimation of the 3D geometry of an object is obtaining the camera parameters.

Conventionally, camera parameters are estimated using image correspondence to get an accurate 3D model. Various schemes have been proposed by researchers to build 3D structure in near real time using image information [1],[2],[3]. Several industrial software such as Patch Based Multi-View Stereo [4] also provides different solutions to generate dense 3D model of a scene from multi-view images. But, the image based pose estimation schemes require high computation time. Moreover, presence of homogeneous texture and varying lighting condition in the scene may lead to failure of the image based 3D reconstruction scheme. Using cloud based 3D reconstruction framework described in [5], one can exploit the computation power of the server to generate the 3D model. But, it is not a convenient solution to the general users as access to a high processing servers is limited.

Nowadays, smart mobile (both high and low end) devices are equipped with build-in sensors like accelerometer, gyroscope and magnetometer which provide motion and orientation information. These additional sensor's data have proven to provide computationally efficient 3D reconstruction framework [6],[7],[8],[9]. Project Tango [6] uses these sensor information in their mobile device to model the 3D world. But, their mobile device is equipped with additional hardware such as motion tracking camera and integrated depth sensing tools. Therefore, it increases the infrastructure cost and is not easily available to the common users. Pan et al. have developed a 3D model on mobile phone which could not provide adequate structural information due to the sparse nature of the model [7]. In [8], Tanskanen et al. have utilized smart phone sensor's data along with image information to calculate accurate motion and orientation parameters for dense 3D reconstruction. But, usage of image information and epipolar geometry [10] increases the complexity of the reconstruction algorithm. For outlier filtering, they have also utilized computationally expensive methods like 5-point algorithm [11] and random sample consensus scheme [12] in their proposed algorithm. Moreover, they have used mobile GPU to achieve real time computation. Recently, Brojeshwar et al. have proposed a framework for real time 3D reconstruction, where mobile sensor's data is solely used for camera parameter estimation [9]. But, in practical scenarios smart phone sensor's data may get corrupted due to various environmental and sensor effects like magnetic vibrations, temperature etc [13]. Therefore, in most of the cases it becomes very difficult to accurately and precisely estimate the camera parameters.

Although, the magnitude of acceleration is greatly affected due to variability in applied force on the phone but the direction of movement is relatively robust. Keeping this in mind, we propose a robust camera pose estimation method leveraging the direction information obtained

from accelerometer values. The position estimation using accelerometer direction is inspired from [14]. Along with position estimation, the rotation of camera is also improved using fusion of accelerometer, magnetometer and gyroscope data. In this regard, the main contributions of the proposed framework can be enumerated as below-

- Use of directional component of noisy accelerometer values to compute camera position accurately. We show that, even in presence of noise in absolute value of acclerometer readings, the direction is a reliable statistic to compute position.
- Use of multi-sensor's data fusion to compute camera rotation robustly.
- Robust computation of position and rotation improves the estimation of epipolar geometry which in turn produce better geometric filtered point correspondence, thus improving the quality of reconstruction.
- The proposed 3D reconstruction framework executes in real-time on CPU of the phone. Therefore, it can run on inexpensive smart phones which may be devoid or have low end GPU and is not capable of executing high computational activities.

The rest of the paper is organized as follows : An overview of the proposed framework is presented in Section 2. Section 3 describes the data acquisition procedure of the proposed system. The novel camera calibration methodology based on sensor's data is explained in Section 4. The final 3D model estimation scheme is described in Section 5. Section 6, contains the experimental results and conclusion is drawn in Section 7.

2. Proposed Framework

In the proposed framework, a robust and computationally efficient camera pose estimation method based on relatively stable features of various sensor's data is described. The main stages of the 3D reconstruction pipeline are as follows-

- Data acquisition
- Camera calibration procedure
- 3D model estimation

Fig. 1, shows our 3D reconstruction pipeline. The user captures the sensor's data and images using a mobile data capture application. These acquired data are processed using the proposed 3D reconstruction framework to generate the dense 3D model in real-time.

3. Data Acquisition

The user moves their smart phone to get multiple view images of the subject. Simultaneously, data from different mobile sensors such as accelerometer, gyroscope and magnetometer is also captured. With respect to the mobile coordinate system, the accelerometer and gryscope provides the linear acceleration (m/sec^2) and the angular velocity

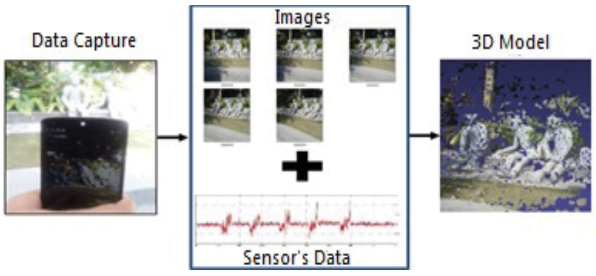


Fig. 1: Our 3D reconstruction pipeline.

(rad/sec) along X-axis, Y-axis and Z-axis, respectively. The sensor's data is used to compute the camera pose with respect to the world coordinate system. The internal clocks of the mobile sensors are time synchronized to obtain uniform data.

4. Camera Calibration Procedure

Camera calibration is of two types - internal and external calibration. Internal calibration estimates the focal length and principle point of a camera. The internal parameter remains the same for a particular resolution of an image of a camera. Therefore, the internal parameter is calibrated only once for a particular device. The external calibration estimates the rotation and position of the camera with respect to world coordinate system.

4.1 Rotation from Smart Phone Sensor Data

Orientation readings of the phone can be obtained by various combination of sensor's data. Here, orientation can be represented using the Euler angles $-$ azimuth angle (rotation of XY plane), row angle (rotation of XZ plane) and pitch angle (rotation of YZ plane). Using a combination of accelerometer and magnetometer data, we can estimate the orientation readings of the phone $\mathbf{o}^{am}$. Similarly, gyroscope data also provides the orientation of the phone $\mathbf{o}^{g}$ [13]. However, both $\mathbf{o}^{am}$ and $\mathbf{o}^{g}$ are prone to various noise. In presence of magnetic vibrations, the signal exhibits short interval variations and diverges from its true value. Therefore, in the proposed scheme low pass filter is applied on $\mathbf{o}^{am}$ to obtain a smoother signal. Though $\mathbf{o}^{g}$ does not get affected by magnetic field, but it is influenced by constant gyro bias and calibration noise. This is known as the gyro drift. We apply high pass filter on $\mathbf{o}^{g}$, to preserve the high frequency component of the signal and nullify the effect of gyro drift. Similar to [15], for every i^{th} sensor reading, we calculate the fused orientation parameter by combining the filtered $\mathbf{o}^{am}$ and $\mathbf{o}^{g}$ as,

$$\mathbf{o}_i = \lambda \mathbf{o}_i^{am} + (1-\lambda)\mathbf{o}_i^{g} \quad (1)$$

where λ is a tuning parameter and $\mathbf{o}_i$ is the i^{th} reading of the fused orientation of the phone. From every sample of $\mathbf{o}$,

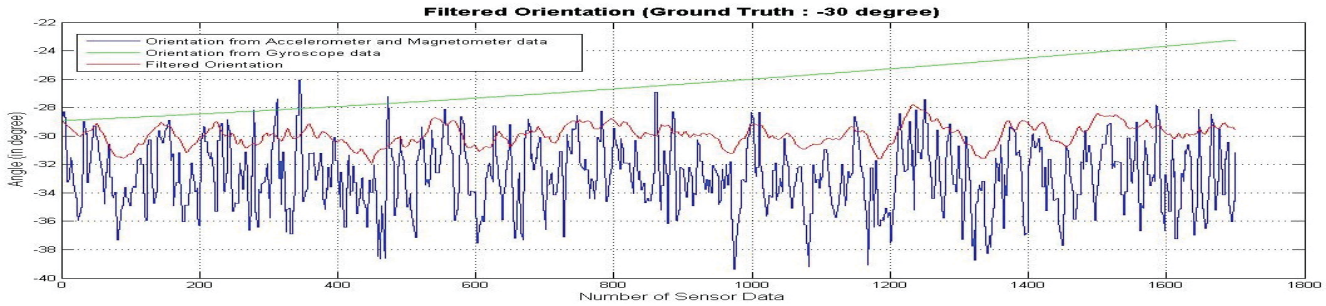


Fig. 2: Fused azimuth angles of **o** (denoted by red line) estimated using a combination of azimuth angle from accelerometer-magnetometer $\mathbf{o}^{am}$ (denoted by blue line) and gryoscope $\mathbf{o}^{g}$ (denoted by green line).

fused rotation matrix R is estimated and used for calculation of image positions.

Fig. 2, shows a graph in which the azimuth angle readings of **o** (in red colour), $\mathbf{o}^{am}$ (in blue colour) and $\mathbf{o}^{g}$ (in green colour) are presented. It is evident from Fig. 2, the azimuth angle of **o** is closest to the ground truth. Similar results are obtained using the pitch and row angles of the orientation parameters. In Table 1, we have compared the mean value of all orientation parameters $\mathbf{o}_i$, $\mathbf{o}_i^{am}$ and $\mathbf{o}_i^{g}$ in static condition where $i = 1, 2, .., 1700$. From both Fig. 2 and Table 1, it is evident that **o** provides higher accuracy and is closer to the ground truth than $\mathbf{o}^{am}$ and $\mathbf{o}^{g}$. The accuracy obtained using this filtering reduces reprojection error in the triangulation stage as illustrated in Section 6.

Table 1: Comparative result of $mean(\mathbf{o}_i)$, $mean(\mathbf{o}_i^{am})$ and $mean(\mathbf{o}_i^{g})$ with respect to ground truth (GT) in degree (deg).

GT (deg)	$mean(\mathbf{o}_i)$ (deg)	$mean(\mathbf{o}_i^{am})$ (deg)	$mean(\mathbf{o}_i^{g})$ (deg)
[15,0,0]	[15.6,0.0,0.1]	[17.3,0.4,1.6]	[9.9,-1.9,-8.6]
[30,0,0]	[30.3,0.0,-0.1]	[32.8,0.8,1.7]	[26.3,-4.3,-8.5]
[45,0,0]	[44.2,0.1,0.2]	[46.3,0.9,1.7]	[41.4,-4.0,-7.4]
[90,0,0]	[88.1,-0.3,0.2]	[93.5,1.5,1.7]	[85.5,-6.3,-7.7]

4.2 Position from Smart Phone Sensor Data

The accelerometer of the smart phone provides the raw acceleration readings $\mathbf{a}^{raw}$ in m/sec^2. But, environmental and sensor conditions affect the accelerometer data and make it noisy [13]. This ultimately results in erroneous position estimation. Some of these noises which influence the accelerometer data are static bias and random noise.

Theoretically, in static condition acceleration should be zero. But, due to the presence of static bias **B**, the accelerometer data deviates from its ideal value and in turn gives an error in estimated position. Therefore, we obtain the bias corrected acceleration value $\mathbf{a}^{\mathrm{B}}$, by subtracting **B** from $\mathbf{a}^{raw}$. Static bias **B** is estimated by averaging $\mathbf{a}^{raw}$ in static condition [13]. Fig. 3 shows $\mathbf{a}^{raw}$ and $\mathbf{a}^{\mathrm{B}}$ (represented using red and blue lines respectively) along with their estimated positions. The positions estimated from $\mathbf{a}^{\mathrm{B}}$ (represented using green circles) exhibit higher accuracy than the positions estimated using $\mathbf{a}^{\mathrm{raw}}$ (marked by red circles). Furthermore, computing image positions from acclerometer value using Newton's law of motion produce erroneous estimate as shown in Fig 3. Therefore, the accelerometer data needs further refinement.

While acquiring data, the user's hands can shake. It gives an illusion of movement in the accelerometer data, leading to wrong position estimate. Therefore, the region of active mobile movement (AMM) is determined from $\mathbf{a}^{\mathrm{B}}$ to estimate the position correctly. While the phone is in static condition, we calculate the standard deviation of the accelerometer data $\mathbf{B}^{std} = std(\mathbf{a}_i^{raw} - \mathbf{g}_i^{raw})$. Here, $std()$ represents the standard deviation function and $\mathbf{g}_i^{raw}$ is the gravitational projection of $\mathbf{a}_i^{raw}$ for the i^{th} sensor's data in static condition. While user is capturing the data, the region of AMM is identified as,

$$AMM = \begin{cases} 1 & \text{if, } 3\mathbf{B}^{std} < mean(\mathbf{a}_i^B - \mathbf{g}_i^B) < -3\mathbf{B}^{std} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where, $\mathbf{g}_i^B$ is the gravitational projection of $\mathbf{a}_i^B$ for the i^{th} sensor's data.

If "$AMM = 1$", the phone is said to be in motion; otherwise it is in static condition. In Fig. 4(a), the red box marks the region of AMM on $\mathbf{a}^B$ (denoted by green line). It is evident from Fig. 4(b), that the position estimated using the region of AMM (indicated by green circles) is more precise than the positions estimated without using the region of AMM (represented using red circles).

For every sample of sensor's data where the mobile is in motion, the acceleration reading is used to estimate the image position. Assuming, the time difference between two readings of acceleration data in motion be $\triangle t$ where, $\triangle t = t_{i-1} - t_i$. Using Newton's law of motion, we calculate the displacement of the phone from its previous position in t_{i-1} for the i^{th} sample of sensor's data reading as,

$$\mathbf{s}_i = \mathbf{u}_i \triangle t + 0.5 R_i (\mathbf{a}_i^B - \mathbf{g}_i^B - \mathbf{B}) \triangle t^2 \tag{3}$$

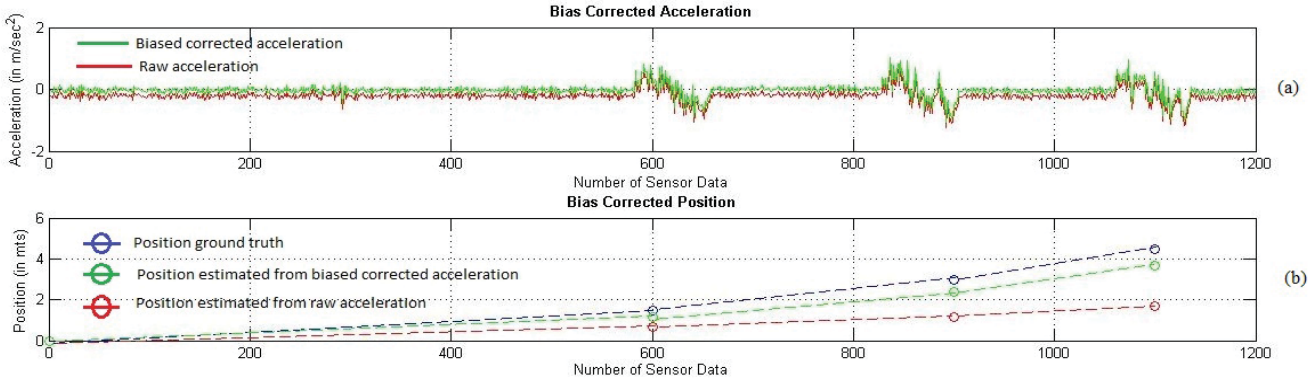


Fig. 3: (a) shows raw acceleration value $\mathbf{a}^{raw}$ and the bias corrected acceleration value $\mathbf{a}^{\mathrm{B}}$ using red line and green line respectively. (b) comprises of global image positions. The global images positions using $\mathbf{a}^{raw}$ and $\mathbf{a}^{B}$ is denoted by red and green circle respectively. The ground truth is given in blue circle.

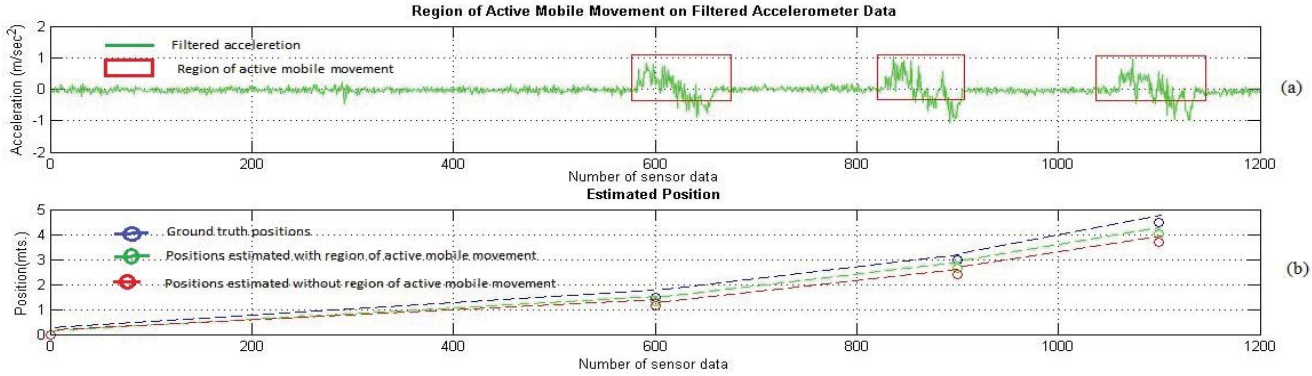


Fig. 4: (a) Here region of active mobile movement is represented by red boxes on $\mathbf{a}^{B}$ (marked by green line). (b) contains global image positions. The ground truth positions are in blue circles, the positions estimated using $\mathbf{a}^{B}$ in the region of active mobile movement are in green circles and the red circles denote the positions estimated from $\mathbf{a}^{B}$ without using region of active mobile movement.

where $\mathbf{u}_i$ is the initial velocity and R_i is the fused rotation matrix (as discussed in section 4.1) for i^{th} sensor's data. We use eq. (3) to estimate the relative position between two consecutive images. This process is similar to the relative position estimation scheme described in [9].

Fig. 5, shows a view-graph where its vertices represents the initial relative image positions. Here, d, e and f are the images taken in time t_d, t_e, and t_f where $t_d < t_e < t_f$. We use eq. (3) to estimate the relative image position between d and e as $\mathbf{c}_{de}$. Similarly, we estimate $\mathbf{c}_{ef}$. While calculating the relative image position, the former image is taken as the origin. Ideally, the acceleration at the former image should be zero. Therefore, we calculate $\mathbf{c}_{de}$ assuming $\mathbf{a}_d = 0$ where $\mathbf{a}_d$ is the acceleration at d. But, as the accelerometer provides noisy data, the acceleration at the image positions may not be zero. Therefore, $\mathbf{c}_{df}$ is calculated using non zero acceleration values at the intermediate image positions that is, $\mathbf{a}_e \neq 0$. In the same way, $\mathbf{c}_{dg}$ is estimated by taking $\mathbf{a}_e \neq 0$ and $\mathbf{a}_f \neq 0$. This concept is extended to estimated the global image positions $\mathbf{c}_d$, $\mathbf{c}_e$ and $\mathbf{c}_f$ that is, image positions with respect to the global origin. In this paper, the position of the first image of the dataset is considered as the global origin.

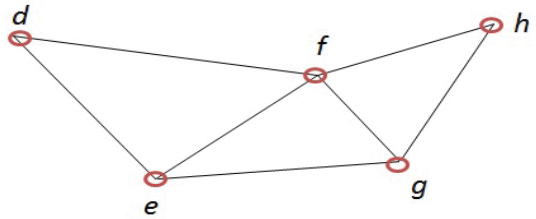


Fig. 5: View-graph of initial image positions. Images are taken in alphabetical order (ascending).

Due to the presence of various random noise, the sensor's data gets corrupted and as a result initial image positions become imprecise. Therefore, there is a need to optimize the initial image positions. In [9], the image positions are optimized using iterative reweighted least square position av-

eraging. Though position estimation methodology described in [9] gives good accuracy on datasets with smooth user transitions. But, we may not get precise position estimate using this method due to variability of force applied on the phone and environmental noise during data acquisition. The primary reason behind this is the force applied on the phone effects the accelerometer values. It is experimentally seen that in such conditions, the magnitude of the acceleration data gets noisy, but the direction of the movement can be reliably estimated. Using Newton's law of motion as given in eq. (3), the image positions are computed from the filtered accelerometer data. Then, we calculate the pair-wise unit direction vector between the mobile position at t_i and t_j , $\hat{c}_{ij}$ where $j = i + 1$ as,

$$\hat{\mathbf{c}}_{ij} = \frac{\mathbf{c}_{ij}}{||\mathbf{c}_{ij}||_2} \tag{4}$$

This concept can be extended for calculating the unit direction vectors between the camera positions. Therefore, for every pair of image positions d and e, we get a unit directional vector, $\hat{c}_{de}$. Then, similar to [14], the global position for those pairwise direction can be obtained using $\rho(\frac{(\mathbf{c}_d^o - \mathbf{c}_e^o)}{|(\mathbf{c}_d^o - \mathbf{c}_e^o)|}, \hat{\mathbf{c}}_{de})$ where, $\rho()$ is the pseudo huber loss function [10] and $\hat{\mathbf{c}}_{de}$ is unit directional vector between image position d and e. The earlier estimated initial global positions $\mathbf{c}_d$ and $\mathbf{c}_e$, is provided as initialization parameter to the optimization function.

In Table 2, using root mean square error we compare the accuracy of relative image positions estimations based on [9], our initial position estimation method and our direction based position estimation method respectively. The data is collected by moving the phone in a planer surface. We consider the euclidean distance between two image positions as the ground truth. In most of the cases, the error using the direction based position estimation method has reduced significantly compared to other two methods. This indicates improvement in the accuracy of our position estimation method. The usefulness of the optimized image positions are evaluated through triangulation and the obtained results are shown in Fig. 6. Fig. 6(a) is the original image. Fig. 6(b) is the triangulation result obtained from optimized image positions as against Fig. 6(b), where image positions are estimated using [9]. It is evident in Fig. 6(c), image positions are not accurate due to which the structure estimated is imperfect as marked in red. On the other hand, in Fig. 6(b) the human is clearly reconstructed.

5. 3D Model Estimation

Stereo-correspondence estimation is one of the essential stage in 3D reconstruction framework. A low computation stereo-correspondence finding algorithm using a combination of gradient and colour information [9] is used. The proposed precise camera pose computation yields accurate epipolar geometry [10] which leads to better stereo-correspondence estimation. This improves the quality of 3D reconstruction and reduces the reprojection error [10]. The results given in Fig. 10 and Table 4 supports our claim. A detailed explanation is provided in Section 6. Using the estimated correspondences and global camera parameters, the initial 3D points by an established method called triangulation [16] is generated. In the final stage of our framework, using well known bundle adjustment algorithm, simultaneously the dense point cloud and cameras parameters [17] is optimized.

Table 2: Comparison of relative image position using : [9], our initial position estimation method and direction based optimized position estimation method against the ground truth (GT) in terms of root mean square error (RMSE). The relative position is estimated in euclidean distance (in cms).

GT	RMSE using [9]	RMSE using our initial positions	RMSE using our optimized positions
10	2.38	1.27	0.20
15	6.67	4.45	3.33
25	2.46	2.49	1.63
30	3.86	3.36	1.86
60	8.47	7.34	1.99

(a) (b) (c)

Fig. 6: (a) is the original image; 3D models after triangulation stage where the image position is estimated (b) using our method (c) using [9]. The imperfections in the 3D model (marked in red) are due to inaccurate camera pose estimation.

6. Experimental Results

For testing the proposed 3D reconstruction method, LG Nexus 5 smart phone is used. It has Quad-core 2.3 GHz Krait CPU and 2GB RAM. The sensor's data is recorded at a frequency of 50 Hz and the images are captured with resolution 640x480. Extensive experiments are carried out for qualitative and quantitative evaluation of the proposed framework. Quantitative evaluation are performed based on the analysis of reprojection error [10] and time requirement for camera pose estimation. Whereas, qualitative experiments include assessment of visual quality of the estimated 3D model by testing the proposed framework on different datasets of varying shape and size in general illumination

condition.

In image based epipolar geometry estimation, computation time is dependent on number of stereo-corre-spondences. Table 3 shows a typical time requirement for computing fundamental matrix between pair of images using 8-point algorithm [10]. Instead of that, using smart-phone sensors for computing epipolar geometry is efficient as it is independent of number of stereo-correspon-dence points. Also, the computation time is very less. In Table 3, the epipolar geometry computation time is 0.0193 sec using sensor's data. As a result, for 1,00,000 stereo correspondence points the proposed epipolar geometry estimation method reduces the computation burden by approximately 701 times.

Table 3: Time requirement for epipolar geometry estimation time with varying number of stereo-correspondence points. Using sensor's data, our pipeline takes only 0.0193 sec and remains independent of number of stereo-correspondence points

Number of stereo corresponding points	Epipolar geometry estimation (sec.)
1,00,000	13.3253
1,25,000	15.9598
2,00,000	26.2119
5,00,000	62.2099

The accuracy of the proposed camera pose algorithm is evaluated using estimated reprojection error in the triangulation stage. Higher accuracy of the camera pose indicates lower reprojection error. Different combinations of rotations (R) and positions (P) calibrated using [9] and the proposed scheme are used to evaluate the camera pose estimation performance. These combinations are briefly described below -

- **S1** - Both R and P estimated using [9]
- **S2** - R estimated using [9] and P estimated using our method
- **S3** - R estimated using our method and P estimated using [9]
- **S4** - Both R and P estimated using our method.

In Table 4, the mean reprojection error using combination **S1**, **S2**, **S3** and **S4** with varying number of 3D points is shown. The reduction of reprojection error in combination **S2**, **S3** and **S4** against **S1**, indicates improvement of the proposed estimated camera pose with respect to the camera pose estimated using [9]. Apart from camera pose, the reprojection error is dependent on stereo-correspondences. We have used a simplistic stereo-correspondence finding method to have a real-time system. Therefore, in presence of homogeneous texture and varying lighting condition we may get noisy output. In these cases, despite of having accurate camera pose, the reprojection error is not reduced to a large extend. Usage of high computation stereo-correspondence finding algorithm will further reduce the reprojection error at the cost of increase in system execution time.

Table 4: Comparison of reprojection error using the camera pose combination – **S1**, **S2**, **S3** and **S4**

No. of 3D points	**S1** (pixel)	**S2** (pixel)	**S3** (pixel)	**S4** (pixel)
101994	10.74	7.23	7.22	5.08
112016	9.12	7.02	8.43	6.90
100124	18.79	14.54	14.23	13.37
84495	35.92	31.62	30.29	29.64

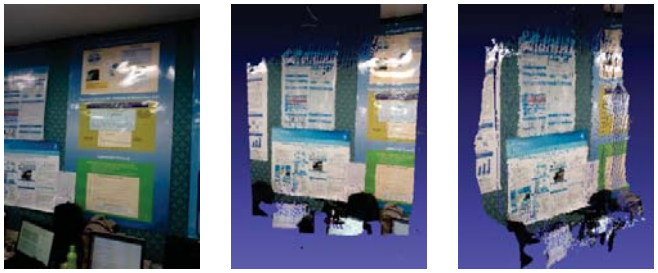
Different stages of our 3D reconstruction algorithm are executed parallely on the mobile CPU. The reconstruction stages takes around 2-3 seconds for every image. The stereo-correspondence algorithm takes less than a second and can be estimated while the user captures the data. The computation time for our camera pose estimation method is in the order of microseconds. The 3D reconstruction pipeline stages which require considerable amount of time are triangulation and global bundle optimization. The time requirement for both the stages is similar to [9].

The proposed framework is tested on different indoor and outdoor datasets of varying shape and size in general illumination. Fig. 7, shows a reconstructed dense 3D model of an outdoor structure of height 1.76 meters using 6 images. The 3D model of the objects in Fig. 8 of height 3.04 metres and Fig. 9 of height 0.25 meters are reconstructed using 8 images and 5 images respectively taken in indoor condition.

(a) (b) (c)

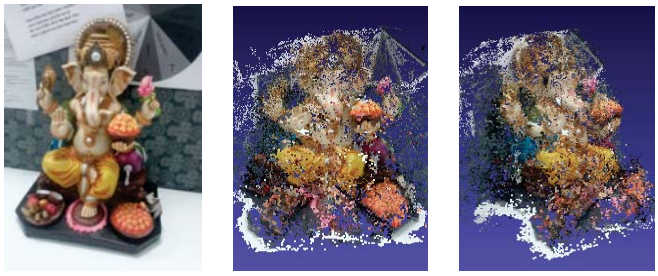
Fig. 7: (a) is an outdoor image of an object of height 1.76 meters; (b) and (c) are the front and side view the 3D model respectively.

The proposed method is compared with the scheme described in [9] and the comparisons are given in Fig. 10 using 3D model outputs from triangulation stage. The dense 3D models shown in Fig. 10(b) and Fig. 10(c) are the triangulation outputs where camera pose is estimated using [9] and our method respectively. Due to imprecise camera pose, the estimated 3D structure in Fig. 10(c) is imperfect as marked in red. Whereas, Fig. 10(b) gives comparatively accurate model because of better camera calibrations and stereo-correspondences. As, evident from Fig. 10(d-f), the accuracy of the final 3D model increases after our final stage of global bundle optimization.



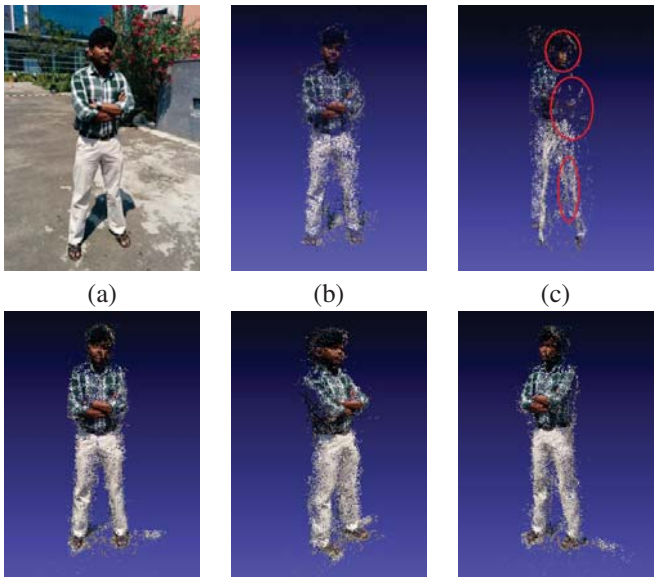
(a) (b) (c)

Fig. 8: (a) is an indoor image of an object of height 3.04 meters ; (b) and (c) are the front and side view the 3D model respectively.



(a) (b) (c)

Fig. 9: (a) is an indoor image of an object of height 0.25 meters; (b) and (c) are the front and side view the 3D model respectively.



(a) (b) (c)

(d) (e) (f)

Fig. 10: (a) is the original image; 3D models after triangulation stage (b) using the proposed method (c) using [9]. The imperfections in the 3D model (marked in red) are due to inaccurate camera pose estimation; (d) and (e-f) are the front and side views of the final 3D model using the proposed method.

7. Conclusion

In this paper, a computationally efficient on-device 3D reconstruction framework which runs on the CPU of the phone is proposed. The framework comprises of a robust and computationally effective camera pose estimation method based on relatively stable features of the smart phone sensor's data. The direction based position estimation method works robustly in practical scenarios where sensor's data is corrupted due to variability of force applied on the phone or environment noise. Rotation computation using fusion of multi-sensor's data also provides enhanced result. The estimated precise camera parameter improves the epipolar geometry which in turn gives accurate stereo-correspondence points. This reduces the reprojection error and improves the quality of 3D reconstruction. The proposed novel 3D reconstruction scheme can be extended in other applications like augmented reality, animation, robot vision, communication and bio-metric authentication.

References

[1] C. Wu, "Towards linear-time incremental structure from motion" *in Proc. IEEE Int. Conf. 3DTV*, pp. 127–134, 2013.

[2] N. Snavely, SM. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3D" *in Proc. ACM Int. Conf. TOG*, vol. 25, pp. 835–846, 2006.

[3] M. Klopschitz, A. Irschara, G. Reitmayr and D. Schmalstieg, "Robust incremental structure from motion" *in Proc. 3DPVT*, vol. 2, 2010.

[4] Y. Furukawa and J. Ponce, Available: http://www. di.ens.fr/pmvs/

[5] Y. Hong, L. Sun, K. Tang and Y. Li, "Real-time cloud-based 3D reconstruction and interaction with a stereo smartphone" *in Proc. ACM Int. Conf. Multimedia Systems Conference*, pp. 152–155, 2014.

[6] J. Lee, Available: https://www.google.com/atap/projecttango/

[7] Q. Pan, C. Arth, G. Reitmayr, E. Rosten, E., T. Drummond, "Rapid scene reconstruction on mobile phones from panoramic images" *in Proc. IEEE Int. Conf. ISMAR*, pp. 55–64, 2011.

[8] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys,"Live metric 3d reconstruction on mobile phones" *in Proc. IEEE Int. Conf. ICCV*, pp. 65–72, 2013.

[9] B. Bhowmick, A. Mallik and A. Saha, "Mobiscan3D: A low cost framework for real time dense 3D reconstruction on mobile devices"*in Proc. IEEE Int. Conf. UIC*, pp. 783–788, 2014.

[10] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision" *Cambridge university press*, 2003.

[11] D. Nistér "An efficient solution to the five-point relative pose problem", *in Proc. Pattern Analysis and Machine Intelligence*, vol. 26, pp. 756–770, 2004.

[12] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography", *in Proc. Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[13] O.J. Woodman, "An introduction to inertial navigation", *University of Cambridge, Computer Laboratory, Tech. Rep. UCAMCL-TR-696*, vol. 14, pp. 15, 2007.

[14] K. Wilson and N. Snavely, "Robust global translations with 1DSfM", *in Proc. IEEE Int. Conf. ECCV*, pp. 61–75, 2014.

[15] S. Ayub, A. Bahraminisaab and B. Honary, "A sensor fusion method for smart phone orientation estimation" *in Proc. Annual Post Graduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, Liverpool*, 2012.

[16] R. Hartley and P. Sturm, "Triangulation", *in Proc. Computer vision and image understanding*, vol. 68, pp. 146–157, 1997.

[17] C. Wu, S. Agarwal, B. Curless and S.M. Seitz, "Multicore bundle adjustment", *in Proc. IEEE Int. Conf. CVPR*, 2011.

SESSION

SIGNAL/IMAGING SCIENCE, COMPUTER VISION, AND APPLICATIONS + MEDICAL IMAGING + ASSISTIVE TECHNOLOGIES

Chair(s)

TBA

3D Closed Loop Boundary Detection and 6 DOF Pose Estimation

Huu Hung Nguyen[1] , **Jane Shi**[2] , **and Sukhan Lee**[1]
[1] School of Information and Communication Engineering of Sungkyunkwan University, Suwon, Korea
[2] General Motors Global R&D Center, 30500 Mound Road, Warren, MI, USA

Abstract - *For vision guided robotic assembly, one of the fundamental enablers is the robust estimation of 6 degree-of-freedom (DOF) pose of industrial parts or subassemblies. In this paper, we present a method to estimate 6 DOF pose of automotive sheet metal panels using 3D closed loop boundary (CLB) features from a stereo vision. The 3D CLBs extracted are used to identify the corresponding CAD model and estimate its 6 DOF pose with reference to the camera frame. The novelty of the proposed method lies in the fact that 3D CLBs are extracted efficiently by matching 2D CLBs from the stereo pair with its search space confined to the region of interest (ROI) and by reconstructing only the 3D data of the matched CLBs using the epipolar constraint. Our proposed method of the 6 DOF pose estimation using 3D CLBs has been demonstrated and applied to several decklid inner panels at GM Research Lab. Experimental results indicate that the proposed method offer computation efficiency less than one second and high performance under occlusion: over success rate 90% under 15% of occlusion.*

Keywords: 6DOF pose estimation, 2D/3D closed loop boundary, and stereo camera.

1 Introduction and related work

For vision-guided robotic assembly applications, a robust 6 DOF pose estimation is a critical enabler. Popular approach of object pose estimation [1] consists of 3 steps: (a) propose feature correspondences (matches) between model features and image features, (b) computing a hypothesized geometric transformation (hypothesis generation), and (c) check the agreement of image features and the transformed model features to confirm the suggested pose (hypothesis verification). This popular approach can be applied ideally to objects with rich 3D geometric features such as automotive inner panels shown in Figure 1 left.

Several methods of 6 DOF pose estimation from 3D shape features have been published recently. 3D planar surfaces and 3D cylinders [2] [3] are modelled using 3D point cloud data, and then these 3D features are used to determine the object pose by matching 3D mesh surfaces from CAD model [4]. On the other hand, several feature descriptors and matching algorithms have been extended from 2D to 3D such as Harris 3D [5] and 3D SURF [6] on 3D meshes or 3D point cloud data. Additionally, Rusu proposed the viewpoint feature histograms for fast 3D pose estimation [7]. These descriptors are invariant to rigid body transformation, however, sensitive to noise and occlusion. Additionally they are significantly expensive in computation for 3D than 2D.

Figure 1 An example of automotive inner body panel (left) outer body panel (right)

In automotive bodyshop applications, outer body panels, as shown in Fig. 1 right, generally have non-texture or few geometric features. Thus very low number of features, key points, or descriptors such as 3D Harris and 3D SURF can be detected even with high computation time. However, inner body panels with rich 3D geometric features are ideal objects to use 3D closed loop boundaries [8] where 3D images are generated from range images by applying morphology techniques [8]. However, reflective object surfaces are not well suited for the structured light camera. For this class of objects, the stereo vision is the best choice to construct 3D features from corresponding 2D features. Several stereo camera based circular detection have been proposed to determine location of object for real-time tracking [9].

Stereo 3D reconstruction can be divided into two main approaches: dense [10] and sparse [11] stereo correspondence. The "dense" approach produces a disparity estimate at every pixel that can provide 3D information for all image region. The "sparse" is based on the corresponding 2D features. Correct correspondence between 2D features of paired images is a critical step. To reduce searching space in finding corresponding pair, an image rectification is a step where 2D projective transformations are used to form an epipolar line for depth recovery in one dimension space.

Our method is based on the "sparse" approach with 3D CLB features. Five major steps are needed to estimate 6 DOF pose of an automotive inner panel:

1) Automatic 2D CLBs detection from edges extracted from two 2D images of stereo camera by Lanser's method [12] individually;
2) The region of interest (ROI) identification based on the epipolar constraints with working distance;
3) The stereo correspondence of 2D CLBs is established in its respective ROI using the shape and size indexes, and their 3D CLBs can be reconstructed quickly;
4) 6 DOF pose hypothesis generation between 3D model CLB and image CLBs;
5) A hypothesis and candidate transformation of 6 DOF pose for the object is generated using reconstructed 3D CLBs and 3D CLBs of CAD model. The final 6 DOF pose is selected to minimize least-square-fitting-error (LSE).

The remainder of this paper is organized as follows: we present the stereo vision setup and the algorithm overview in Section 2, followed by detailed description of 2D CLB feature extraction, ROI identification, and 3D CLB reconstruction in Section 3. Next, we outline the 6 DOF pose estimation approach in Section 4. The experimental results and algorithm performance are summarized in Section 5. We discuss our future work and conclude our paper in Section 6.

2 System and algorithm overview

Figure 2 shows the stereo camera setup with detailed parameters listed in Table 1. Each camera is a 5 megapixel digital camera, specifically Prosilica GC2450C GigE [17] from Allied Vision Technologies. The baseline distance is 140mm which was fixed for another project. This baseline distance can be increased for a better Z resolution for the decklid inner part. Similarly we can select a longer focal length than current 8.6 mm for a better X and Y resolution.

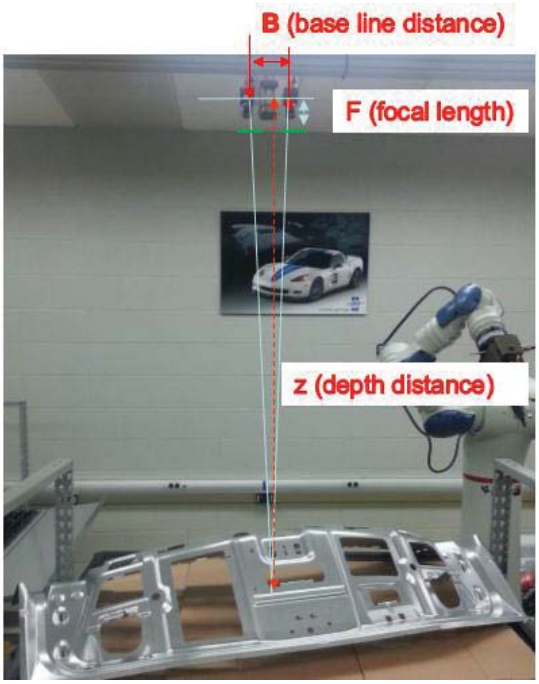


Figure 2 Camera system configuration

Table 1 Camera system configuration

Baseline Distance (B)	140 mm
Focal Length (F)	8.6 mm
Pixel Number	2448 x 2050
Pixel size (δd)	0.0034 mm
X and Y resolution	0.513 mm (at Z depth = 1.3 m)
Z resolution	4.3 mm (at Z depth = 1.3 m) 7.5 mm (at Z depth = 1.8 m)

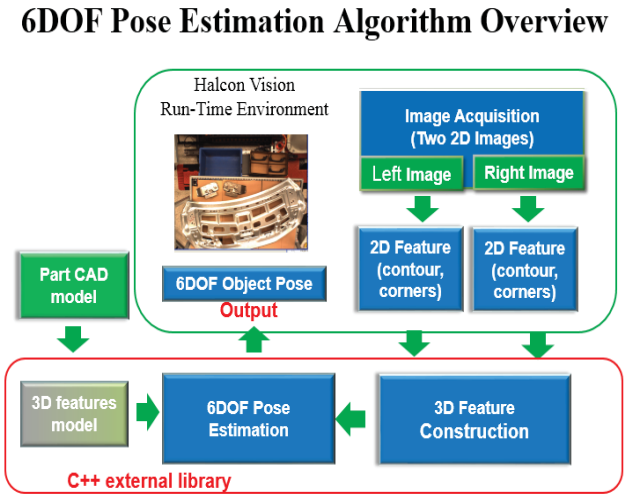


Figure 3 6DOF Pose Estimation Algorithm Overview

Fig. 3 above is the algorithm overview of our 6DOF pose estimation based on the stereo vision system. Halcon vision run-time environment [18] is used for 2D image acquisition from stereo cameras, 2D feature extraction, and final 3D pose display and update. We implemented an event loop with a fixed 0.5 second loop time within Halcom vision run-time environment. This means that 2D images and outputs are updated every 0.5 seconds. We developed an external library in C++ that is loaded into Halcon vision run-time event loop manager. 3D CLB construction algorithm, as described in Section 3, and 6DOF pose estimation algorithm, as described in Section 4, have been implemented in this external C++ library.

3 3D closed loop boundary extraction

To speed up the image processing time in later steps of edge detection, we used color images to segment the inner panel object (silver grey) from the background clusters. As shown in Fig. 4 below, we converted 2D images in RGB (Red-Green-Blue) color space to HSI (Hue-Saturation-Intensity) color space, and then filtered the converted images based on S value (0 to 120) and I value (120 to 255). The filtered image is then clustered to find the biggest connected cluster for the inner panel part. The smallest rectangle area that completely covers the biggest cluster is our region of interest (ROI) where all edges of the parts reside inside this rectangle.

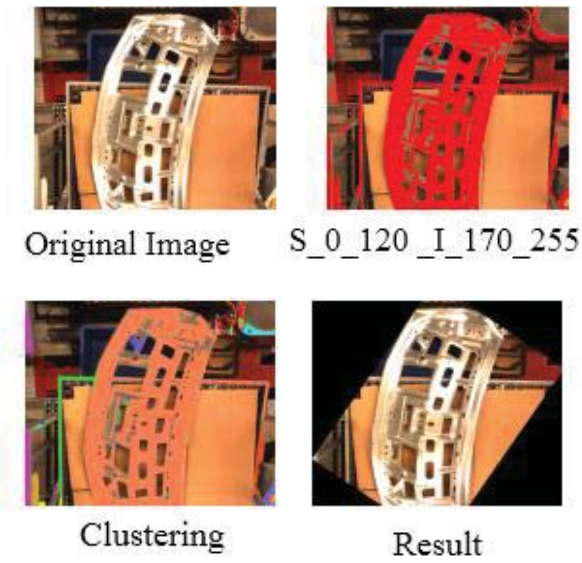


Figure 4 Original image in RGB is filtered and segmented based on HSI values to segment and identify the smallest rectangular ROI area for the inner panel part.

Our 3D CLB extraction algorithm consists of three major steps:

1) Automatic 2D CLBs detection from edges extracted from two 2D images of stereo;
2) the region of interest (ROI) identification based on the epipolar constraints;
3) 3D CLBs reconstruction based on the shape and size similarity;

We describe each of these steps in next two sections in detail.

3.1 2D closed loop boundary extraction

Closed edges usually are the openings on the inner panel whose start pixel and end pixel are the same. The definition is similar to the circle or ellipse but the shape of the closed edge are random. Edges detected by Lanser's method [12] could be closed, opened or mixed. For each edge, we perform following steps to extract closed loop:

1) Un-assign distance for all the point of a detected edge.
2) Choose a point randomly, push it into a pop queue, set its distance to zero, and assign it as its parent.
3) Take out a point from queue, search its neighbors and check each searched neighbor,
   a) if it is un-assigned, set its distance to pop distance plus one, assign the taken out point as its parent, push it to pop queue.
   b) On the other hand, if it is assigned already and assigned distance equal to distance of taken out point, then a closed loop exists. Go to Step 4.
4) From two points, we go back follow their parents, when their parent are same, we remove them from the edges as closed edge. Go Step 3.

Repeat (3) (4) until all points are assigned.

Fig.5 left shows the image with all edges with in ROI, the middle image is for the detected closed edges only, and the right illustration shows the points and their assigned pop distance to detect 2D CLB. The point marked *0* is root parent, points on the edge are pushed to queue to assign distance, when the distance of taken out point and the distance its neighbor point equal(*47*), we start go back to find closed loop.

Fig. 5. 2D Image with all edges (left) with detected CLBs only (middle) and the pop distance assignment for CLB detection.

3.2 3D closed loop boundary reconstruction

Until now, 2D closed loop edges are obtained for both left and right images captured from the stereo camera. Some of them appear on both images, otherwise others appear on left only or on right only. To construct 3D closed loop edges, corresponding pairs of 2D CLBs should be determined first. To reduce the search area and also to increase the matching accuracy, we use epipolar rule with the minimum (Z_{min}) and maximum (Z_{max}) distance to define a region of interest (ROI) in right image for each left CLB.

Given a point, x_L, of 2D CLB in the left image, we can project this point to a 3D point at Z distance in 3D space by (1), and then project it to a 2D point (x_L) in the right image by (2) as shown in Fig.6.

$$X(\mathrm{Z}) = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} ZK_L^{-1}x_L \\ 1 \end{bmatrix} \quad (1)$$

$$x_R = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \cong K_R[R|t]\begin{bmatrix} ZK_L^{-1}x_L \\ 1 \end{bmatrix} \quad (2)$$

Where:

K_L and K_R is the intrinsic matrix of left camera and right camera respectively

$[R|t]$ is the extrinsic matrix of right camera, while the extrinsic matrix of left camera is the identity matrix.

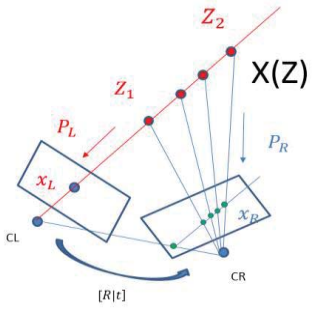


Figure 6 Epipolar rule

For each 2D CLB on the left image, we project two CLBs onto the right image using (2) at the minimum (Z_{min}) and maximum (Z_{max}) distance. Two new projected CLBs, *CLBmin and CLBmax*, are generated as shown in Fig.7 below. The corresponding 2D CLB on the right image should be within the range between CLBmin and CLBmax. This area is our region of interest (ROI) for 2D CLB correspondence.

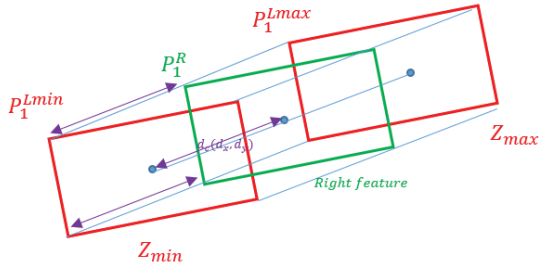


Figure 7 Two Projected CLBs at Z_{min} and Z_{max} distance is the Region of Interest (ROI) for the corresponding right CLB.

When several 2D CLBs exist in this ROI area, all of them should be considered as a candidate corresponding CLB. To identify precise correspondence we will search the CLBs in the bounded area using a shape similarity score $\left(e_{i1}^{\alpha} + e_{i2}^{\beta}\right)$ as in (3). In other words, exactly matched CLBs satisfy two shape similarity conditions: the similarity of boundary length (N_c in pixel counts) and similarity of enclosed interior area (A_c).

$$e_{i1} = \max\left(\frac{N_{CL}}{N_{CR_i}}, \frac{N_{CR_i}}{N_{CL}}\right)$$
$$e_{i2} = \max\left(\frac{A_{CL}}{A_{CR_i}}, \frac{A_{CR_i}}{A_{CL}}\right) \quad (3)$$
$$\min\left(e_{i1}^{\alpha} + e_{i2}^{\beta}\right) \quad i: 0 - n$$

Where:

N_{cL} *and* N_{cR} is the CLB boundary length in pixel counts for the left CLB and the right CLB respectively.

A_{cL} *and* A_{cR} is the CLB enclosed interior area in pixel counts for the left CLB and the right CLB respectively.

α, β are the control parameters (default to 2 for equal weight of both conditions).

Once the correspondence of left CLBs and right CLBs are established, we can determine point-to-point correspondence between two matched CLBs.

We first compute the central displacement d_C that is distance between the center point of right CLB and the center point of left CLB at $Z_{min.}$ For a point (P_j^R) on the right CLB, there is a corresponding point (P_i^L) on the left CLB that is on the line defined by two points (P_i^{Lmin}, P_i^{Lmax}) as shown in Fig. 7. Corresponding pair of points on two CLBs, P_i^{Lmin} and P_j^R, satisfies two conditions in Eq.(4) below: (1) their displacement is same as the central replacement d_C and (2) P_j^R is on the line formed by two points (P_i^{Lmin}, P_i^{Lmax}).

$$d_{i1} = \left| Dis(P_j^R - P_i^{Lmin}) - d_C \right|$$
$$d_{i2} = \text{Dis}\,(P_j^R, P_i^{Lmin} P_i^{Lmax}) \quad (4)$$
$$\min\left(d_{i1}^{\theta} + d_{i2}^{\omega}\right) \quad i, j: 0 - N$$

Where:

P_i^{Lmin}, P_i^{Lmax} are corresponding points of left CLBs at $Zmin, Zmax$.

P_j^R is a point on the right CLB.

θ, ω are the control parameters (default to 2 for equal weight of both conditions).

Once CLB to CLB correspondence and point-point correspondence on two corresponding CLBs are determined, we apply the triangulation rule to construct a 3D CLB with all boundary points. Fig. 8 below is one example result with two corresponding 2D CLBs and the resultant 3D CLB.

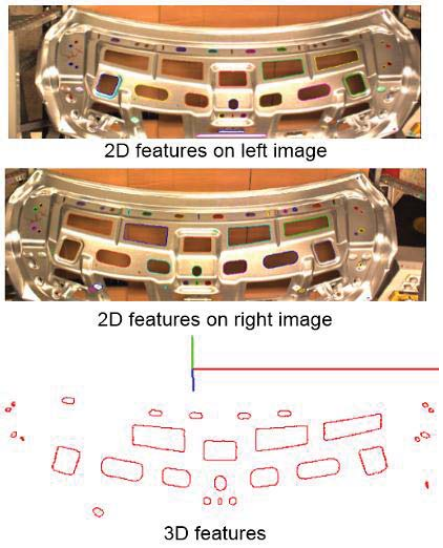


Figure 8 One example result of 2D CLBs and 3D CLBs

4 6 DOF pose estimation

Once we have 3D CLBs from previous steps, we can estimate 6 DOF pose of the decklid part with its CAD model.

Given the CAD model feature points M (where its i^{th} column is a 3D point M_i, i=1,…,n) and the image feature points F (where its j^{th} column is a 3D point F_j, j=1,…,m) from the 3D CLBs, we have a rigid body transformation relationship between these two sets of 3D points as illustrated in Eq. (5) below:

$$M = TF \qquad (5)$$

Where: T is a 4 by 4 homogeneous transformation matrix that includes a 3 by 3 rotation(R) and 3 by 1 translation (t).

In order to use (5) to estimate 6DOF pose, the corespondance between the model point M_i and the image point F_j has to be given or known. Without the prior known correpondenc, an iterative closest points (ICP) [14] is a well-known method to compute the transformation matrix T[R|t]. With nosiy feature data F, a least-square based fitting method [13][15] should be used. However, both methods take a long time to yield a result as they are interative methods. To speed up the 6DOF pose estimation algorithm, we can establish a good initial correspondence estimate among M and F using their shape similarity as shown in our algorithm flow chart in Figure 9 below.

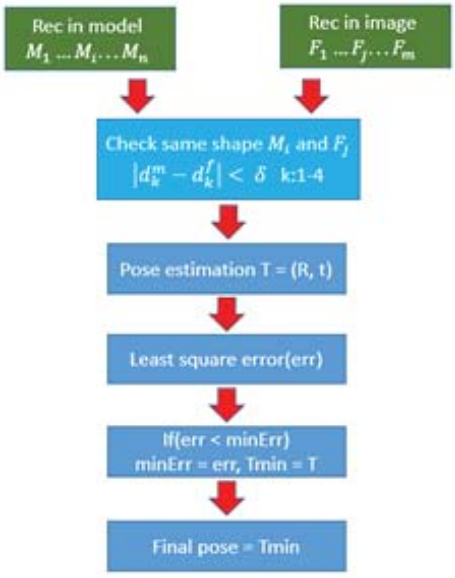


Figure 9 6DOF pose estimation algorithm

For each CLB, we compute its centeral 3D position $\bar{F}$ and $\bar{M}$ as the average position of all CLB boundary points in the feature set and in the model set respectively.

$$Mi' = [M_i - \bar{M}] \qquad (6a)$$
$$Fi' = [F_i - \bar{F}] \qquad (6b)$$

We use similar shape to check all possible cases. This will significantly reduce the number of candidate pairs. For each similar shaped CLBs, the least square error between model and object is computed. Assume that, M' is a matrix form of points M'_i of n points of model with each column for one point on model CLB, and F' is matrix form of points F'_j of m points of image features with each column for one point on feature CLB, Eq. (5) becomes Eq. (7) below:

$$M' = TF' \qquad (7)$$

When the correspondence of n points are known by the shape similarity test, we can compute the singular value of decomposition [16] of the least square error fitting as in (8) below:

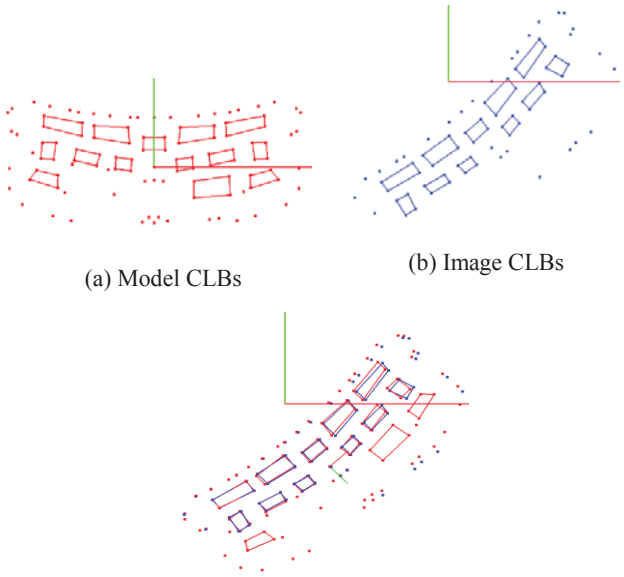
$$M' * (F')^t = VSU^t \qquad (8)$$

Where: V and U^t are orthonormlaized engien vectors associated with n largest eigenvalue in S.

Then the rotation matrix and the translation vector can be estimated by

$$R = VU^t \qquad (9)$$
$$t = \bar{M'} - R\bar{F'} \qquad (10)$$

The least square fitting error is computed for all candidate transformation T[R,t] and the one with the minimum least square error is chosen as final estimated 6DOF pose.



(a) Model CLBs (b) Image CLBs

(c) Estimated 6DOF pose with the minimum least square error

Fig. 10. One Example of 6DOF Pose estimation.

5 Experimental results and performance

In this section, we evaluate the performance of our algorithm for both computation and accuracy. The first major part in our algorithm is the 3D CLB reconstruction from two stereo images as detailed in Section II. Our 2D CLB matching and 3D CLB reconstruction has a better performance than two well-known methods, sum of absolute differences (SAD) and sum of squared differences (SSD), for stereo matching in term of measuring mean distance error (MDE). We use the window size 9x9 and find the minimum cost along the epipolar line from $Zmin$ to $Zmax$ with the 2 pixel gap. We set parameters $\alpha, \beta, \theta, \omega$ to 2 for equal weights for all four factors. Both 3D methods search the corresponding points in limited area defined by epipolar constraint and the known range of CLB

depth. We vary the object's position in two directions: linearly along camera Z axis up and down by 10 cm and rotationally about the Y axis by 3 degree. Figure 11 below shows the mean distance error (MDE) of our method in comparison with SAD and SSD methods for the translational position change (upper graph) and rotational position change (lower graph)

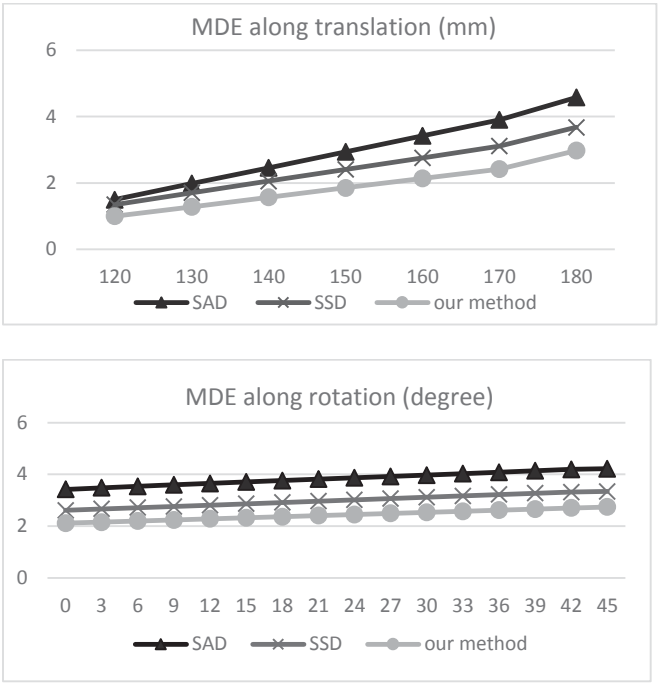


Fig. 11. Mean distance error (MDE) of our method in comparison with SAD and SSD

The second major part in our algorithm is the 6DOF pose estimation from the image 3D CLBs and the model 3D CLBs. To evaluate accuracy of 6DOF pose estimation, we rotate the object's position about the camera Z axis by 6 degree for 60 increments to complete the whole 360 degree rotation at the fixed Z distance (1.5m). Our model is consisted of 50 CLBs with 15 of them that are bigger than 2.5cmx2.5cm. These big CLBs play a bigger role to the 6DOF pose estimation than other remaining CLBs since they have significant shape information while smaller shapes are not distinguish enough and mainly used for calculate least square fitting error.

Table 2 Error in Estimated 6DOF Pose

	Euclidean Distance Error ratio	X (mm)	Y (mm)	Z (mm)	α (deg)	β (deg)	γ (deg)
Our	7.52 mm / 5 mm	3.21	3.12	6.14	1.67	1.74	2.56
Y.Lee[8]	1.8 mm/ 0.4 mm	1.4	0.8	0.5	-	-	-

Table 2 above lists the comparison of our estimated 6DOF pose with Lee's method [8]. Our method use stereo camera to test the decklid object shown in Fig 1.a at 1.5 m, Lee's method use high quality camera to test object in working distance from 1m to 1.5m.

The large positional error along X, Y, and Z in our method is most due to the resolution in our stereo camera setups. A high quality camera is used in Lee's case [8]. Therefore it is not a equivalent comparison. However, if we use a ratio to normalize the depth resolution in each camera system, the ratio of Euclidean distance error and the depth resolution, our method yields the ratio of 1.5 (7.52mm/5mm) which is better than the ratio of 4.5 (1.8/0.4) in Lee's method.

In addition, we also verified the performance of our 6DOF pose estimation with occlusion. As the number of CBLs are occluded, the estimated pose success rate will decrease as expected. At fixed Z distance of 1.5 m, we vary the object's position in X direction as a portion of the object is out of the view gradually. Fig.12 is the graph which shows the decreased number of valid CLBs (upper) where Num of REC indicate big CLBs and the reduced success rate (below) for the estimated 6 DOF pose. The success rate from 0 to 30% of occlusion obtained from experiments, and this value from 30% to 50% of occlusion is estimated.

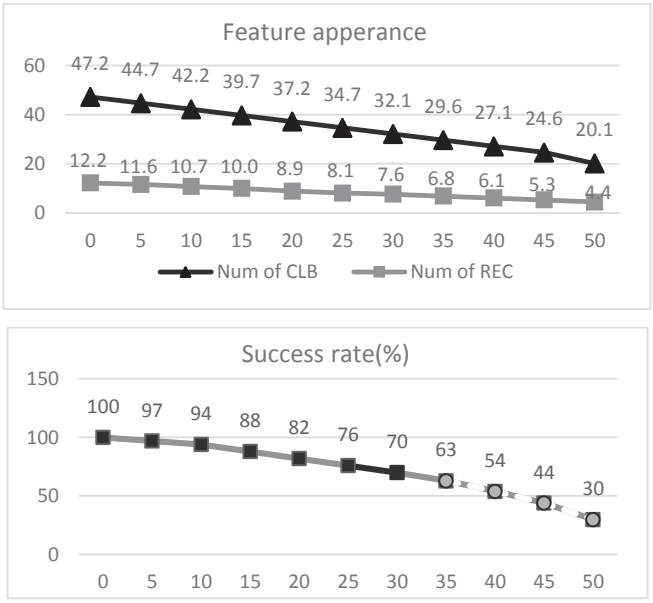


Figure 12 Estimated 6DOF Pose Performance with occlusion

Our method can be used in real-time application with the total computation time at 1.0 second. Roughly 0.85 second is for two 2D CLB extraction (0.5 seconds) and CLB correspondence (0.35 seconds) in Halcon run-time environment. 6DOF pose estimation using 3D CLBs takes 0.15 seconds with a DLL library written in C++.

6 Conclusion

In this paper, we present a fast and robust method to estimate 6 DOF pose of automotive inner panels using 3D CLBs from a stereo vision. First, 2D closed loop boundaries are extracted from two RGB images of the stereo pair. Next, the region of interest (ROI) in the paired image is determined

based on the epipolar constraints within the known working distances. Then, the stereo correspondence of 2D CLBs is established in its respective ROI using the shape and size indexes, and their 3D CLBs can be reconstructed quickly. Finally, a hypothesis and candidate transformation of 6 DOF pose for the object is generated using reconstructed 3D CLBs and 3D CLBs of CAD model. The final 6 DOF pose is selected to minimize least-square-fitting-error (LSE).

We evaluate the performance of our 6 DOF pose estimation algorithm and demonstrate that the mean distance error (MDE) of our method is better than two well-known methods, SAD and SSD. However, the absolute error in our results is worse than Lee's method due to the poor depth resolution (the z resolution is 5mm at z = 1.5 meters). When measured by normalized error, i.e. with mean distance error to depth resolution ratio, our method performs better. We also evaluate the occlusion effect on the appearance of 3D CLB features and pose correct rate. As expected, these performance deteriorates as the number of distinguish 3D CLBs decreases from 12 to 4.

Our 6DOF pose estimation algorithm is fast, within 1 second, to be used for real-time applications. We have applied this method to several decklid inner panels at the manufacturing research lab, GM Global R&D Center, Warren, MI, US.

7 Acknowledgments

This work was supported by MEGA science research and development projects, funded by Ministry of Science ICT and Future Planning (NRF-2013M1A3A3A02042335), and also by the Technology Innovation Program (10048320), funded by the Ministry of Trade, Industry and Energy (MI, Korea). This work was performed at Manufacturing Systems Research Lab, GM Global R&D Center, Warren, MI with support and help from Gurdayal. S. Koonjul, currently with General Electric, Lance Ransom at GM Manufacturing Engineering, and Neil McKay at GM Global R&D Center.

8 References

[1] Haralick, R. M., Joo, H., Lee, D., Zhuang, S., Vaidya, V. G., & Kim, M. B. "Pose estimation from corresponding point data". Systems, Man and Cybernetics, IEEE Transactions on, 19(6), 1426-1446, 1989.

[2] Nguyen, H. H., Kim, J., Lee, Y., Ahmed, N., & Lee, S. "Accurate and fast extraction of planar surface patches from 3D point cloud". In Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication (p. 84). ACM, Jan ,2013.

[3] Schnabel, R., Wahl, R., & Klein, R. "Efficient RANSAC for Point-Cloud Shape Detection". In Computer graphics forum (Vol. 26, No. 2, pp. 214-226). Blackwell Publishing Ltd, June, 2007.

[4] Kim, J., Nguyen, H. H., Lee, Y., & Lee, S. "Structured light camera base 3D visual perception and tracking application system with robot grasping task". In Assembly and Manufacturing (ISAM), 2013 IEEE International Symposium on (pp. 187-192). IEEE, Jul 2013.

[5] Sipiran, I., & Bustos, B. "Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes". The Visual Computer, 27(11), 963-976, 2011.

[6] Knopp, J., Prasad, M., Willems, G., Timofte, R., & Van Gool, L. "Hough transform and 3D SURF for robust three dimensional classification". In Computer Vision–ECCV 2010 (pp. 589-602). Springer Berlin Heidelberg, 2010.

[7] Rusu, R. B., Bradski, G., Thibaux, R., & Hsu, J. "Fast 3d recognition and pose using the viewpoint feature histogram". In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on (pp. 2155-2162). IEEE, October , 2010.

[8] Lee, Y., Lee, S., Kim, D., & Oh, J. K. "Improved industrial part pose determination based on 3D closed-loop boundaries". In Robotics (ISR), 2013 44th International Symposium on (pp. 1-3). IEEE, Oct 2013.

[9] Yoon, Y., DeSouza, G. N., & Kak, A. C. "Real-time tracking and pose estimation for industrial objects using geometric features". In Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on (Vol. 3, pp. 3473-3478). IEEE. Sep, 2003.

[10] Scharstein, D., & Szeliski, R."A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". International journal of computer vision, 47(1-3), 7-42, 2002.

[11] Loop, C., & Zhang, Z."Computing rectifying homographies for stereo vision". In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. (Vol. 1). IEEE, 1999

[12] Lanser, S., & Eckstein, W. (1992). "A modification of Deriche's approach to edge detection". In Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis, Proceedings., 11th IAPR International Conference on (pp. 633-637). IEEE. 1992.

[13] Arun, K. S., Huang, T. S., & Blostein, S. D. "Least-squares fitting of two 3-D point sets". Pattern Analysis and Machine Intelligence, IEEE Transactions on, (5), 698-700, 1987.

[14] Paul J. Besl and Neil D. McKay. "A method for registration of 3-D shapes". IEEE Transaction on Pattern Analysis and Machine Intelligence, 14(2):239–256, 1992.

[15] Estépar, R. S. J., Brun, A., & Westin, C. F. "Robust generalized total least squares iterative closest point registration". In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004 (pp. 234-241). Springer Berlin Heidelberg, 2004.

[16] Golub, G. H., & Reinsch, C. Singular value decomposition and least squares solutions. Numerische Mathematik, 14(5), 403-420, 1970.

[17] Prosilica GC2450 Camera from AVT GugE http://www.alliedvision.com/en/products/cameras/detail/2450.html

[18] Halcom Vision Run-time Engine http://www.halcon.com/halcon/hdevelop/hdevengine.html

Cable Footprint History: Spatio-Temporal Technique for Instrument Detection in Gastrointestinal Endoscopic Procedures

Chuanhai Zhang[1], Wallapak Tavanapong[1], Johnny Wong[1], Piet C. de Groen[2], and JungHwan Oh[3]
[1]Dept. of Computer Science, Iowa State University, Ames, IA, USA
[2]Mayo Clinic College of Medicine, Mayo Clinic, Rochester, MN, USA
[3]Dept. of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Abstract - *We propose a new fast spatio-temporal technique that detects an operation scene---a video segment corresponding to a single purpose diagnosis action or a single purpose therapeutic action. The technique utilizes (1) color contrast of the cable region and the background, (2) the new area-based coordinate system to compute spatial features, and (3) the history of locations of detected cables of the instrument in a video to discard false regions. The proposed technique and software are useful for (1) automatic documentation of diagnostic or therapeutic operations at the end of the procedure, (2) a second review for causes of complications due to these operations, and (3) a building block for an effective content-based retrieval system to facilitate endoscopic research and education. On 38 full-length colonoscopy and upper endoscopy video files with different cable colors and four different insertion directions of the instruments, the average percentage of false positive duration is small at 2.23%. The average percentage of true positive duration is high at 92.9%. The average analysis time per image is 13 milliseconds on an inexpensive off-the-shelf PC.*

Keywords: Colonoscopy, Polypectomy, Instruments, Image Analysis, Image Features

1 Introduction

Gastrointestinal endoscopy is a procedure using an endoscope to diagnose or treat a condition in the digestive system. For instance, colonoscopy enables inspection of the inside of the human colon and diagnostic and therapeutic operations. Colonoscopy is currently the gold standard for colorectal cancer screening. Upper Endoscopy (EGD) is the procedure for inspection of the stomach. In the US, Colorectal cancer is the second leading cause of cancer-related deaths behind lung cancer [1], causing about 50,000 annual deaths. Colorectal cancer and stomach cancer are the third and the fifth most common cancer in the world [2].

During the insertion phase of an endoscopic procedure, a flexible endoscope (with a tiny video camera at the tip) is advanced under direct vision (via the anus for colonoscopy) and (via the mouth for EGD). The video camera generates video of the internal mucosa of the organ. The video data are displayed on a monitor for real-time analysis by the endoscopist. During the withdrawal phase, the endoscope is gradually withdrawn with careful examination of the mucosa. Necessary biopsy and therapeutic operations (e.g., polypectomy) are performed. During these operations, an instrument is inserted via a working channel of the endoscope through the shaft. A variety of instruments (e.g., Fig. 1(a-c), cytology brushes, and sclerotherapy needles) can be used. Within a single procedure, the head and the cable of the instrument typically appears in the field of view (FoV) of the camera from the same approximate direction and location in the image. We call them *insertion direction* and *insertion location,* respectively. For instance, in Fig. 1(d-e), the instrument is in the lower right corner. In Fig. 1(f), the instrument is from the lower left corner.

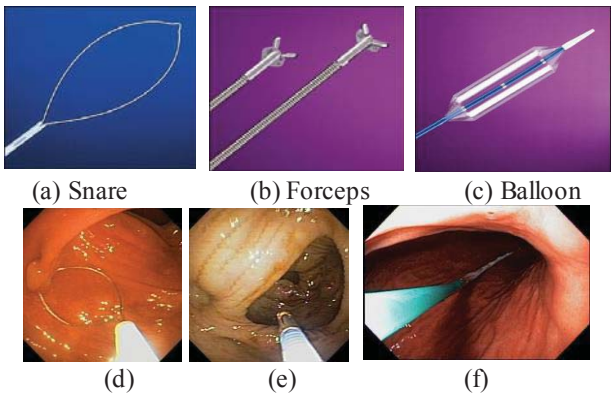


Fig. 1. (a-c) Instruments; (d-f) Appearance of a snare, a biopsy forceps, and an inflatable balloon in endoscopic images

Previously, we defined an *operation shot* in an endoscopic video as a segment of visual data that corresponds to a diagnostic or therapeutic operation [3]. For instance, we count each biopsy as one operation. We proposed an algorithm that detects these shots automatically [3] based on key visual properties of the cable of the instrument. However, the technique is very slow, making it impractical for automatic documentation at the end of the procedure to help

saving physician's reporting time. Furthermore, it is not uncommon that multiple biopsies are performed consecutively on the same or nearby colon mucosa.

Our contribution in this paper is as follows. First, we define an *operation scene* as *a video segment corresponding to a single purpose diagnosis action or a single purpose therapeutic action*. For instance, a series of consecutive biopsies is considered in the same scene with the biopsy purpose. Second, we propose a new, fast spatio-temporal technique for detection of operation scenes. The crux of the technique is (1) the detection of the footprint of the cable of the instrument utilizing the contrast of the cable from the background, (2) a new area-based coordinate system to compute spatial features, and (3) the history of detected footprints of the cable to automatically detect the insertion direction which is effective for eliminating false positives.

The highlight of our findings is as follows. On our data set of 38 full-length Gastrointestinal (GI) endoscopic video files covering a total of 20 hour worth of video data with different cable colors and four different insertion directions, the average percentage of false positive duration is very small at 2.23%. The average percentage of true positive duration is high at 92.9%. On 5 additional full-length GI video files without any operations, the technique does not report any false operations. Most significantly, the average analysis time per image was small, about 13 milliseconds (*ms*) on a PC with 3.4 GHz Intel® Xeon® and 16GB of RAM on our data sets. The analysis time is less than the time interval between two consecutive frames. Reporting detected operation scenes right at the end of the procedure is achievable without slowing down video capturing or other processing.

The remainder of the paper is organized as follows. Section 2 discusses related work on video segmentation techniques for endoscopic videos. We present our proposed technique in Section 3 and discuss the evaluation method and results in Section 4. We give the conclusion and the description of the future work in Section 5.

2 Related work

We limit our discussion of related work in optical endoscopic procedures. We exclude those of wireless capsule endoscopy (WCE) since therapeutic intervention with instruments during WCE procedures is not possible.

2.1 Endoscope hardware and endoscopic images

Within a single GI endoscopic procedure, instruments appear in the same general area in the video because the instruments are inserted through the same working channel from the top of the endoscope. The cable of the instrument has similar tubular shape regardless of various instrument types. The cable appears from one of the borders of the image. The cable may have different colors (e.g., green, blue, orange, and red), but it usually has some parts with high contrast. While the cable has a higher chance to be detected reliably, some operations do not have any cable present in the image because only the head of the instrument appears. A biopsy is typically very short between 2-4 seconds. Multiple biopsies in nearby mucosa are not uncommon.

2.2 Methods for endoscopic video segmentation

We proposed algorithms for blurry frame detection and shot segmentation of colonoscopic videos by color difference in 2004 [4]. Frames with similar color histograms in RGB color space are grouped in the same shot after blurry frames are discarded. These shots do not correspond to a biopsy or therapeutic operation. We proposed a method to segment shots based on camera motion into forward-camera-motion and backward-camera-motion shots [5]. Cumulative camera motion over time is used to find the frame separating the insertion and withdrawal phases. We later proposed a faster technique using motion vector templates [6].

We introduced algorithms for detection of operation shots [3, 7]. Hessian matrix and hierarchical clustering are used in [3] for detection of the insertion direction. This initial step is accurate, but is very slow. The detected insertion direction is used to discard regions outside the area of the detected direction. Moment invariants and Fourier shape descriptors are used in [3] and [7], respectively, for matching the detected regions with the cable template regions. The average processing time per frame with 390×370 pixel resolution once the insertion direction was identified was about 7*s* (seconds) on a PC with 3.40 GHz Pentium(R) 4 and 1GB of RAM. JSEG took 6*s* for segmentation of images into regions. The average true positive fraction and false positive fraction are 0.94 and 0.10, respectively. Only 3% of the actual shot boundaries was missed with 7 false shots on 25 videos of full-length colonoscopic procedures.

Shot segmentation by significant motion changes was proposed [9]. The method uses Kaneda-Lucas Tomasi (KLT) tracking on feature points in nine non-overlapping areas of the images. Motion within each area is calculated and smoothed over the same area in a temporal window. For each frame, the standard deviation *s* of the motion of all the areas is computed. The high value of *s* over a threshold signifies an object movement. Camera movement is determined with the high mean motion and low standard deviation. A candidate motion shot boundary is detected when the change in the standard deviation *s* over a time window is at peak. The detected boundaries are then further refined. The reported average recall and precision are same at 0.86. The processing time was not reported. The technique was not specifically designed to detect instruments in colonoscopy or EGD where instruments appear infrequently or not at all.

In 2004, we introduced a method for detection of endoscopic scenes: rectum, sigmoid, descending colon, transverse colon, ascending colon, and cecum scenes, each corresponding to different sections of the colon [8]. In 2014, a method to detect endoscopic scenes, each defined as one stable feature track of a tissue on the organ surface, was

proposed [10]. This method uses an optical flow enhanced with forward-backward tracking of SIFT features. Proposed Scale-Invariant Distance and Rotation Invariant Angle of pairwise relationships between landmarks in the same frame are used as features to compute a likelihood score to include a subsequent image in the same scene. Scene segmentation was performed on the fly. Average precision between 0.74 and 0.99 and maximum recall between 0.40 and 0.84 of endoscopic scenes were reported on four videos from Olympus Narrow Band Imaging endoscopes.

3 Cable footprint history technique

3.1 Preprocessing

We sample t images per second from the input video, forming what we called *reduced colonoscopy video*. The smaller the value of t, the larger the reduction in the processing time, but the larger the difference between the actual and detected scene boundaries. To reduce processing time, we sub-sample and classify pixels in each input image I_i of the reduced video into two sets: I_{ND}--- non-dark pixel set using Equation (1) where T_c is a constant in the range of 0 and 1 and I_D---dark pixel set for all the pixels excluded from I_{ND}.

$$I_{ND} = \{p |\ p\ is\ a\ pixel\ of\ I_i\ with\ all\ its\ normalized\ \ R,G,B\ values > T_c\ \} \qquad (1)$$

Next, for each image, we compute the median values of the non-dark pixel values in *CIE LUV* color space denoted as M_{LUV}. For each pixel p in the image, we compute the Euclidean distance $d(p, M_{LUV})$ [14] between the pixel values in *CIE LUV* color space of p and M_{LUV}. *LUV* is one of the uniform color spaces for which the distance function maps to perceptual distance well. We separate the foreground and the background using Equation (2) where T_F is a contrast threshold.

$$d(p, M_{LUV}) > T_F,\ \ d(p, M_{LUV}) = \begin{cases} \sqrt{\frac{(p-M_{LUV})^2}{3}} & if\ p \in I_{ND} \\ 0 & if\ p \in I_D \end{cases} \qquad (2)$$

We perform an erosion with a disk structuring element and remove regions that are too large or too small. Let R_i represent the *i*-th remaining connected component. We justify these threshold values based on experiments discussed in Section 4. This step replaces the time-consuming image segmentation method used in the previous algorithms [3,7]. We did not use the well-known Otsu method [14] to obtain a dynamic threshold value for each image because it wrongly considered cable regions as background for images with strong light reflection with higher contrast than cable regions in our training sets.

3.2 Spatial feature extraction

A number of shape features have been proposed [11] with varying degrees of computational complexity. Instead of using invariant moments [12] as in [3] or Fourier shape descriptors [7] to represent region shape, we introduce a new method based on the domain knowledge to calculate a new Cartesian coordinate system to derive region shape features.

By examining the cable regions from 17 endoscopic videos in our training video set, we further refine possible insertion directions into twelve general triangular areas as shown in Fig. 2(a). For each area, two of the borders are denoted by the two-headed arrow in the figure. The third border is a portion of the image border intersecting the first two borders. We define a new Cartesian coordinate system for each corresponding Area k as shown in Fig. 2(b-d). For instance, for $Area_{(2,5,8,11)}$, we choose X' perpendicular to the diagonal line and Y' perpendicular to X' as shown in Fig. 2(d). We derive four spatial features to represent the cable shape as follows.

First, we assign a cable region R_i to Area k denoted by the triangle $Area_k$ if more than half of the pixels of the region are in $Area_k$ as shown in Equation (3). The $|x|$ denotes the number of pixels in region x.

$$Area\ number\ = \{k | \frac{|R_i \cap Area_k|}{|R_i|} > \frac{1}{2}\} \qquad (3)$$

Next, we calculate the features using the area-based coordinate system. Equation (4) shows the eccentricity defined as the ratio of the cable region height in the longest extension of R_i along the Y' axis to the width (the longest extension of R_i along X') of a region R_i.

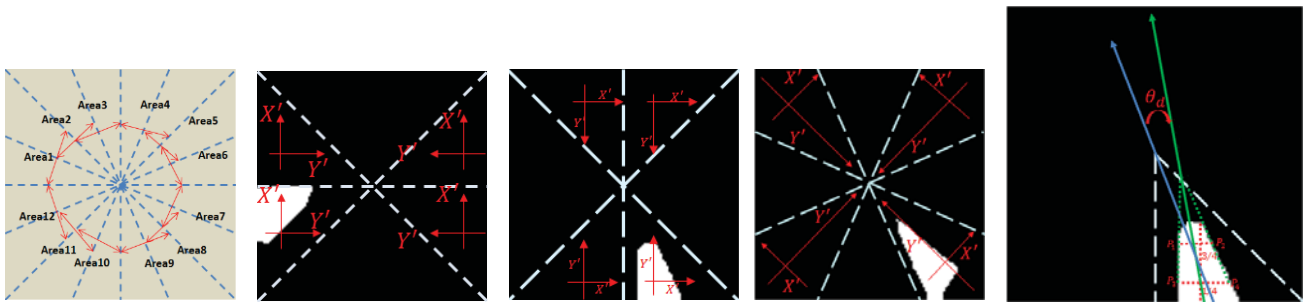
$$eccentricity = \frac{height}{width} = \frac{\max_{x'} \sum_{y'} R_i(x', y')}{\max_{y'} \sum_{x'} R_i(x', y')} \qquad (4)$$

Next, we calculate the orientation of the region. To get a reliable orientation, we only use the middle part of the region. We define two lines: P_3P_4 at the one-fourth height of the region and P_1P_2 at the third-fourth height of the region as shown in Fig. 2(e). The orientation vector of the region is calculated using Equation (5).

$$orientation\ vector\ = \frac{\overrightarrow{P_3P_1} + \overrightarrow{P_4P_2}}{2} \qquad (5)$$

$$d = \min_{(x',y') \in R_i} \{d(R_i(x', y'), border\ of\ Area_k)\} \qquad (6)$$

$$S = \frac{\sum_x \sum_y R_i(x,y)}{Number\ of\ all\ pixels\ in\ the\ image} \qquad (7)$$



(a) (b) for Areas 1,6,7,12 (c) for Areas 3,4,9,10 (d) for Areas 2,5,8,11 (e)

Fig. 2. (a) 12 general insertion areas; (b-d) new Cartesian coordinates; (e) insertion direction calculation

Now, we compute the angle difference θ_d between the region orientation vector and the angle bisector of $Area_k$. See an example θ_d in Fig. 2(e). We calculate the distance d from a region R_i to the border of its $Area_k$ defined in Equation (6) and the normalized area s of the region in Equation (7).

3.3 Feature classification

We investigated the effectiveness of the classification of non-cable/ cable regions using Support Vector Machine (SVM) with the linear kernel and the radial-basis-function kernel as well as J48 Decision Tree on the spatial feature vector ($\theta_d, d, s,$ and *eccentricity*). The parameter values of each type of classifiers were optimized to achieve the best performance as described in Section 4.

3.4 Cable footprint history

In Equation (8), we compute the weight w_i of the pixel at the coordinate (x,y) on image i using the corresponding binary image B_i where only pixels of the detected regions by the classifier have the values of one and the rest have zeros.

$$w_1(x,y) = B_1(x,y)$$
$$w_i(x,y) = B_i(x,y) * (w_{i-1}(x,y) + 1) \quad (8)$$

In other words, the weight of a pixel depends on whether it is part of the detected cable region in the current frame multiplied by one plus the weight of the corresponding pixel in the previous frame. The implication of this recursive equation is that the weight of this x-y location increases when it is part of the detected cable regions in consecutive frames. The weight is reset to zero whenever this location is not part of any detected cable regions. Other positive constant positive values instead of 1 can be used. We chose 1 for simplicity.

We compute the cable footprint history for each frame i from the first frame to the last frame of the video using Equation (9). Fig. 3(a-d) shows cable regions of the same video. Fig. 3(e) shows the cable footprint history of the last frame of the entire video. The brightest region marks the most common pixel locations of detected cable regions in the video.

$$H_1(x,y) = w_1(x,y)$$
$$H_i(x,y) = H_{i-1}(x,y) + w_i(x,y) \quad (9)$$

We use a binary threshold T_H to segment the cable footprint history of the last frame t of the video into a set of connected components. Let HR_j represent the j-th connected component in the set. We choose the brightest connected component $R_\#$ as the insertion area of this video as illustrated in Equation (10).

$$R_\# = \{HR_j \mid \max_{(x,y)\in HR_j} H_t(x,y) = \max_{(x,y)\in I_t} H_t(x,y)\} \quad (10)$$

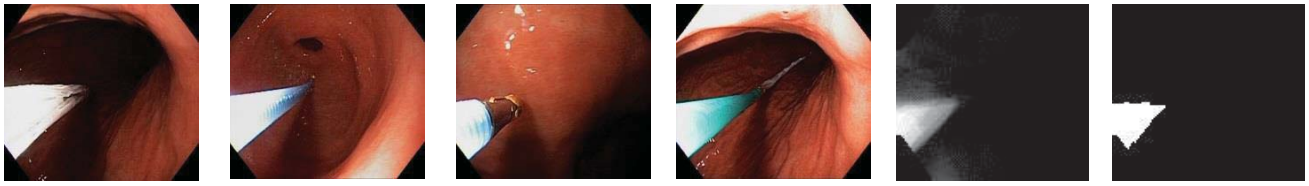
After locating the insertion direction for the video, we discard all candidate regions that do not intersect with $R_\#$ since they are not likely a true cable region. Finally, we assign each frame either 0 or 1. A frame is assigned a 0 if it does not have any remaining cable candidate region. Otherwise, we label it as 1.

3.5 Operation scene detection

This step utilizes temporal information and domain knowledge to identify operation scenes. This step accepts L, a sequence of 0 and 1 from the previous step, as input and outputs the frame numbers indicating the boundaries of the detected operation scenes.

3.5.1 Eliminate falsely detected cable images

This step corrects the misclassification results. We initialize the output sequence L^* with zeros. We slide a window of W frames on L from the beginning to the end of L one digit at a time. Each time, we compute the sum of all the



(a) (b) (c) (d) (e) (f)

Fig. 3. (a)-(d) Cable regions in one colonoscopy video; (e) cable footprint history of this video; (f) detected cable region

numbers under the sliding window. When the sum is equal to W (i.e., all the frames under the window are cable images), we copy all the numbers under the sliding window in L to L*. We set the window size W to $t/2$ where t is the temporal sampling rate in frames per second used in the pre-processing step. This window size covers frames within half a second since we observe that true cable frames typically appear consecutively more than half a second.

3.5.2 Locate cable scene boundaries

Like in our previous techniques [3,7], we scan L^* from the beginning to the end. We first determine a sequence S of consecutive frames from L^* with all the following properties.

1. The sequence S starts and ends with a 1, followed by at least $K * t$ consecutive 0s. In other words, the first and the last frames in S are cable images. The value of K should be the maximum temporal distance between consecutive cable shots of the same scene learned from training.
2. The sequence S must have the ratio between the total number of 1s (cable images) and the length of S greater than a threshold r_1.
3. The sequence S lasts at least 2 seconds based on a consultation with our endoscopist and our observation. A biopsy is typically short about 2-4 seconds. A scene can have multiple sequences.

If the temporal distance between two consecutive sequence S is less than T_t seconds, we group them in the same operation scene. We select the values of r_1 and T_t based on experiments discussed in Section 4.

4 Performance evaluation

The software for operation scene detection was implemented in Matlab. Weka was used for selecting the best classifiers for cable region classification. All the experiments were conducted on a PC with 3.4 GHz Intel® Xeon® and 16GB of RAM.

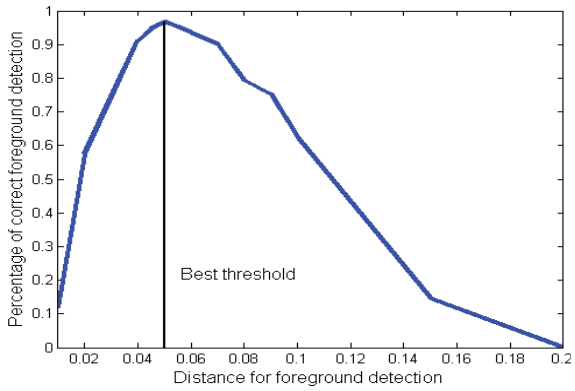


Fig. 4. Sensitivity analysis of the color contrast threshold values tested on cable images in the image

4.1 Data sets and parameter values

Table 1 shows the description of the data sets used in this study. The training data set consists of 17 endoscopic videos that are not in any of the test data sets. All the videos are of full length procedures captured in MPEG2 format. The first and last frames of each ground truth operation scene have cable images in them. The image set was for evaluation of classification effectiveness of cable regions using new features and existing features. The video sets I, II, and III do not overlap. The testing video sets were used to evaluate the effectiveness of the cable region detection and operation scene detection.

Table 1. Description about the data sets

Name	Number	Purpose
Training set		
Image set	1,000 cable images and 1,100 non-cable images	For cable region detection evaluation
Training video set	17 videos	For temporal parameters such as t, K, and T_t
Testing set		
Video set I	38 videos	Strong contrast between cable colors (blue, green, white)and the background
Video set II	18 videos	Weak contrast between cable colors (orange and red) and the background
Video set III	5 videos	No operation scenes

Table 2. Parameters and values used in experiments

Parameter	Value
Sampling rate t fps (frames per second)	6
Image resolution after spatial subsampling (pixels x pixels)	112x112
Dark pixel threshold T_c for preprocessing	0.3
Color contrast threshold T_F for preprocessing	0.05
Disk structuring element for erosion	5
Range of acceptable region size R_S in pixels	$50<R_S<$image area /14
Threshold T_H for determining true cable area	0.5
Ratio of cable frames in an operation shot (r_1)	0.1

The parameter values in Table 2 were determined from the training data in Table 1. We chose the temporal sampling rate of 6 *fps* to minimize the minimum distance between the true scene boundary and the detected scene boundary to 16 *ms*. For spatial subsampling rate, any higher rate, resulting in a smaller image resolution does not provide good classification result though it reduces the processing time. For the optimal color contrast threshold value, we plotted the percentage of correct foreground (cable region) detection with different threshold values using the cable images in the image

set. The plot in Fig. 4 shows that the color contrast threshold of 0.05 gives the highest correct foreground detection result. We observed that many cable images in an operation scene are difficult to detect for several reasons such as blurry images, strong light reflected regions with tubular shape, use of dye color, and too small cable regions. Therefore, we set the ratio of cable frames in an operation shot, r_1 to a small value of $\frac{1}{10}$.

Table 3. Performance and metrics

Metric	Definition
Sensitivity (SE)	Ratio of correctly detected cable images to cable images in the ground truth
Specificity (SP)	Ratio of correctly detected non-cable images to non-cable images in the ground truth
Precision	Ratio of correctly detected cable images to detected cable images
Number of false scenes (#F)	Number of software detected scenes not overlapped with any operation scenes in the ground truth
Number of missed scenes (#M)	Number of operation scenes in the ground truth for which the software miss both boundaries
True Positive Fraction (TPF)	Ratio of correctly detected images as part of operation scenes in the ground truth (true positives) to images of operation scenes in the ground truth
False Positive Fraction (FPF)	Ratio of incorrectly detected images as part of operation scenes (false positives) to images of operation scenes in the ground truth

4.2 Performance metrics

Table 3 shows the performance metrics used in this study. These metrics are similar to the metrics defined on operation shots in the previous work [3]. Note that if one of the two boundaries of an operation scene is incorrect, the detected operation scene still captures part of an operation. In such cases, we did not treat the detected operation scene as a false or a missed scene, but used TPF and FPF to quantify the effectiveness. High TPF is desirable as the algorithm uncovers most frames in the scenes. Low FPF indicates that a small fraction of a detected operation scene is not part of an operation scene.

4.3 Results

We first evaluated the effectiveness of our proposed spatial features. We used the grid search method to find optimal SVM parameters [15] in this experiment. Incorporating the cable insertion direction into the spatial high sensitivity (SE) of 99.9% and specificity (SP) of 97.9% with the decision tree classifier. Since colonoscopy has many more non-cable images than cable images, the 97.9% specificity is still inadequate, causing a large number of false cable images. This is where our cable footprint history is helpful.

Table 4. Performance of 10-fold cross-validation classification of cable images on the image set I

Classifier		SE(%)	SP(%)
New features	SVM + Linear	97.9	94.2
	SVM + RBF	99.6	97.7
	Decision Tree	**99.9**	**97.9**
Existing Features	Invariant moments	62.0	55.3
	Fourier descriptors	67.3	55.2

Table 5. Effectiveness of cable image detection on the video sets I & II

Method	Precision
Using proposed spatial features only	0.45
Proposed spatial features + cable footprint history	0.91

Table 4 shows that Decision Tree outperforms the two SVMs for detecting cable regions using Weka[16]. We chose Decision Tree as the classifier. Our new features perform better than the two existing features used in previous work [3,7]. The angle difference of the insertion direction is quite effective in discarding false regions.

Table 5 shows that the cable footprint history significantly reduces false cable regions by about half compared to using our spatial features only, resulting in doubling the average precision. In this experiment, we did not apply the operation scene detection step described in Section 3.5.

We use the video set I and II to test the effectiveness of our entire operation scene segmentation. The duration of these videos ranges from $4.7m$ to $60.9m$, which illustrates that the algorithm works well. The cable footprint history detects the insertion area correctly in all the videos in both video test sets. The TPF and FPF for the video set I are 0.929 and 0.022, respectively. TPF of 0.929 is very high. The TPF and FPF for video set II with weak cable contrast are 0.813 and 0.027, respectively. In video set II, the TPF is lower and the FPF is higher as the weak color contrast makes it difficult to differentiate the cable region from the background. The weak color contrast mainly results from two cases. First, the cable color is orange or red, which has very weak color contrast with the colon mucosa background. Second, the cable is exposed to the strong white light so that it is mixed with the background. There are 8 false scenes because of bright tubular looking colon surfaces or colon folds. The average number of false scenes per video is very low, 0.14. There are 12

missed scenes which were not detected mainly in two cases. The average number of missed scenes per video is 0.21. On video set III without any operation scenes, the algorithm performs very well. It did not generate any false operation scenes. The average processing time per frame from video sets I and II using our algorithm is 13*ms*. This is a great improvement by at least 38 times when compared to over 500*ms* using the previous method to detect operation shots given a known insertion direction [3]. This great improvement results from the cable extracting method based on the color contrast feature which costs only linear time.

5 Conclusion and Futrue Work

We introduce a new fast operation scene detection technique. The technique is based on color contrast to separate cable regions, new Cartesian coordinates for computing spatial features, and the cable footprint history. Our study shows that the technique is effective and at least 38 times faster than the existing technique. This is because detailed image segmentation or complex foreground-background subtraction is not needed. The experiments on real endoscopic videos demonstrate that the proposed technique misses a small number of operation scenes and it generates a very small number of false video segments that do not correspond to diagnostic or therapeutic operations. The algorithm is operation equipment, endoscope brand and procedure independent.

In our future work, we will improve the cable region extraction method to better separate the orange or red cables from the background. We will also design an online cable detection technique which outputs detected operation scene boundaries after each operation is complete. Such a method has the potential to be useful for analysis of quality of therapeutic operations where appropriate feedback is given for sub-optimal therapeutic operation

6 Acknowledgement

This work is partially supported by Iowa Regent Innovation Funds and EndoMetric Corp. Johnny Wong, Wallapak Tavanapong, and JungHwan Oh hold positions in EndoMetric Corporation, Ames, IA 50014, U.S.A, a for profit company that markets endoscopy-related software. De Groen is the medical advisor of EndoMetric.

7 References

[1] American Cancer Society. Colorectal Cancer Facts & Figures, 2014-2016.

[2] World Cancer Research Fund International. Colon cancer statistics and stomach cancer statistics. http://www.wcrf.org/int/data-specific-cancers.

[3] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Computer-Aided Detection of Diagnostic and Therapeutic Operations in Colonoscopy Videos. IEEE Transactions On Biomedical Engineering, 54(7), July 2007.

[4] J. Oh, S. Hwang, W. Tavanapong, P. C. de Groen, and J. Wong. Blurry Frame Detection and Shot Segmentation for Colonoscopy Videos. In Proc. of IS&T/SPIE Conf. on Storage and Retrieval and Applications for Multimedia, pp. 531-542, San Jose, CA, USA, January 2004.

[5] S. Hwang, J. Oh, J. Lee, Y. Cao, W. Tavanapong, D. Liu, J. Wong, and P. C. de Groen. Automatic Measurement of Quality Metrics for Colonoscopy Videos. In Proc. of ACM Multimedia 2005, pp. 912-921, Singapore, November 2005.

[6] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, and P. C. de Groen. Real-time Phase Boundary Detection for Colonoscopy Videos using Motion Vector Templates. In Springer Lecture Notes in Computer Science (MICCAI Workshop on Computational Abdominal Imaging, Computational and Clinical Applications). Vol. 7601, pp. 116-125, France, 2012.

[7] Y. Cao, D. Li, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Parsing and Browsing Tools for Colonoscopy videos. In Proc. of ACM Multimedia 2004, pp. 844-851, New York, NY, USA, October 2004.

[8] Y. Cao, W. Tavanapong, D. Li, J. Oh, P. C. de Groen, J. Wong. A Visual Model Approach for Parsing Colonoscopy Videos. In Proc. of Int'l Conf. on Image and Video Retrieval (LNCS 3115), pp. 160-169, Dublin, Ireland, July 2004.

[9] M. J. Primus, K. Schoeffmann, and L. Laszlo Böszörmenyi. Segmentation of Recorded Endoscopic Videos by Detecting Significant Motion Changes. In Proc. of CBMI 2013, June 2013.

[10] M. Ye, E. Johns, S. Giannarou, and G.-Z. Yang, Online Scene Association and Endoscopic Navigation, In Proc. of MICCAI 2014, Boston, MA, USA, Sept. 2014.

[11] Mingqiang Yang, Kidiyo Kpalma, Joseph Ronsin. A Survey of Shape Feature Extraction Techniques. Peng-Yeng Yin. *Pattern Recognition*, IN-TECH, pp.43-90, 2008.

[12] Invariant moments: N. Sebe and M. S. Lew, "Robust Shape Matching," In Proc. of IEEE International Conference on Image and Video Retrieval, London, UK, 2002, pp. 17-28.

[13] Poynton, Charles (2003). Digital Video and HDTV Algorithms and Interfaces, Morgan Kaufmann Publishers, page 226.

[14] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66.

[15] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, Tech. Rep., Taipei, 2003.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Accurate Liver Extraction Using a Local-Thickness-Based Graph-Cut Approach

Yasuhiro Kobayashi[1], **Masanori Hariyama**[2], **Mitsugi Shimoda**[3], **Keiichi Kubota**[3]
[1]National Institute of Technology, Oyama College, Japan
[2]Graduate School of Information Sciences, Tohoku University, Japan
[3]Second Department of Surgery, Dokkyo Medical University, Japan

Abstract— *This article presents an accurate and automatic approach to extract a liver from CT images for oncologic surgery planning. Our algorithm exploits graph cut segmentation, which perform global optimization. The quality of graph cut segmentation strongly depends on the edge image that gives prior knowledge about locations of foreground and background regions. In order to get a good edge image, a liver candidate region is extracted based on three types of liver structure models: intensity model, shape model, and blood vessel model. Moreover, the "local-thickness" image of the liver candidate region is used as the edge image. The experimental results show that the use of local-thickness edge image can avoid the over-extraction of anatomical structures surrounding the liver.*

Keywords: Medical imaging, 3D simulation analysis, anatomic hepatectomy, local thickness

1. Introduction

For liver cancer surgery, 3D simulation before surgery operations, recently, is getting one of the crucial tasks since a liver has complex structure. Liver segmentation from CT images is a challenging task since the variations of the liver shape is large and since there exist some other anatomical structures with the CT values similar to the liver around the liver. For example, Fig. 1(a) shows a CT image. Rib muscle contacts Liver partly, and they have almost similar CT values. There have been several researches on liver extraction[1]-[5]. The researches [1] and [2] use, respectively, statistical shape models and probabistic atlases, and both methods suffers from large variations of liver shapes. The active contour approach [3] are dependent on image gradient, and leads to over-extraction into organs with CT values similar to the liver. Moreover, its quality strongly relies on the location and shape of the initial contour. The intensity-based approach [4] usually exploits a simple intensity model, and miss the vessels and non-homogenous texture regions inside the liver. The structure-model-based approach[5] exploits a shape and vessel models as well as an intensity one. Although it successfully extracts the major volume of liver, there is still over-extraction at regions touching the liver tightly like rib muscle as shown in Fig. 1.

2. Algorithm

Figure 2 shows the total flow of extracting a liver region. For simplicity, we use 2-D images in the figure although a 3-D image is used in fact. First, a liver candidate region is extracted from CT images using structure-model-based method[5]. The upper-left and upper-right figures shows a CT image and the liver candidate region. Readers can see rib muscle is mis-extracted as the liver region. Next, for the liver candidate region, the thickness feature is measured by computing “Local Thickness“ [6], where the local thickness of a point is defined as the diameter of the largest sphere that fits inside the object and contains the point. The lower-right figure of Fig. 2 shows the local-thickness image of the liver candidate image. Generally speaking, the core region of the liver has a large LT value whereas mis-extracted regions surrounding the liver tends to have a small LT value. This observation indicate that the local thickness image is a good metrics to give the possibility which voxel is foreground(or background) one. Figure 3 shows the difference between the binary and LT images of the liver candidate region. In the binary image, the mis-extracted region like rib muscle has the same weight value as correctly-extracted region. In the LT image, the mis-extracted region has smaller weight value than the correctly-extracted region. Finally, graph cut segmentation is done using the LT image as the edge image as shown in the lower-left figure. The graph-cut segmentation [7]-[9] is a energy-based approach which allow much more robust segmentation than simple techniques such as region growing or split-and-merge. The CT image and the LT image are used as inputs for the graph-cut segmentation. The graph-cut segmentation works in such a way that regions with small LT values are segmented into background.

3. Experimental results

For programming, we use the Java-based image processing platform called ImageJ[10]. ImageJ has a lot of plugins like a DICOM reader, basic 2D/3D image processing and visualization, and can be extended easily by adding use-defined plugins. Moreover, ImageJ runs on multi-platforms like Microsoft Windows, Linux and MacOS since it is based on Java. To be specific, we use Fiji, one distribution of ImageJ, since it has much more plugins such as graph cut

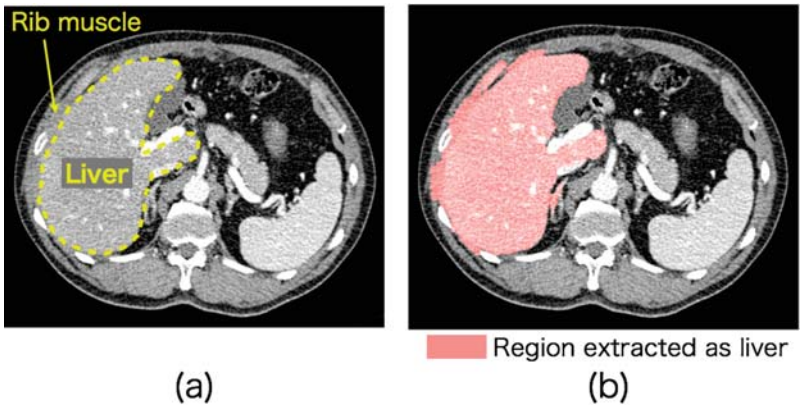


Fig. 1: CT image including the liver and extraction result of conventional approach[5].

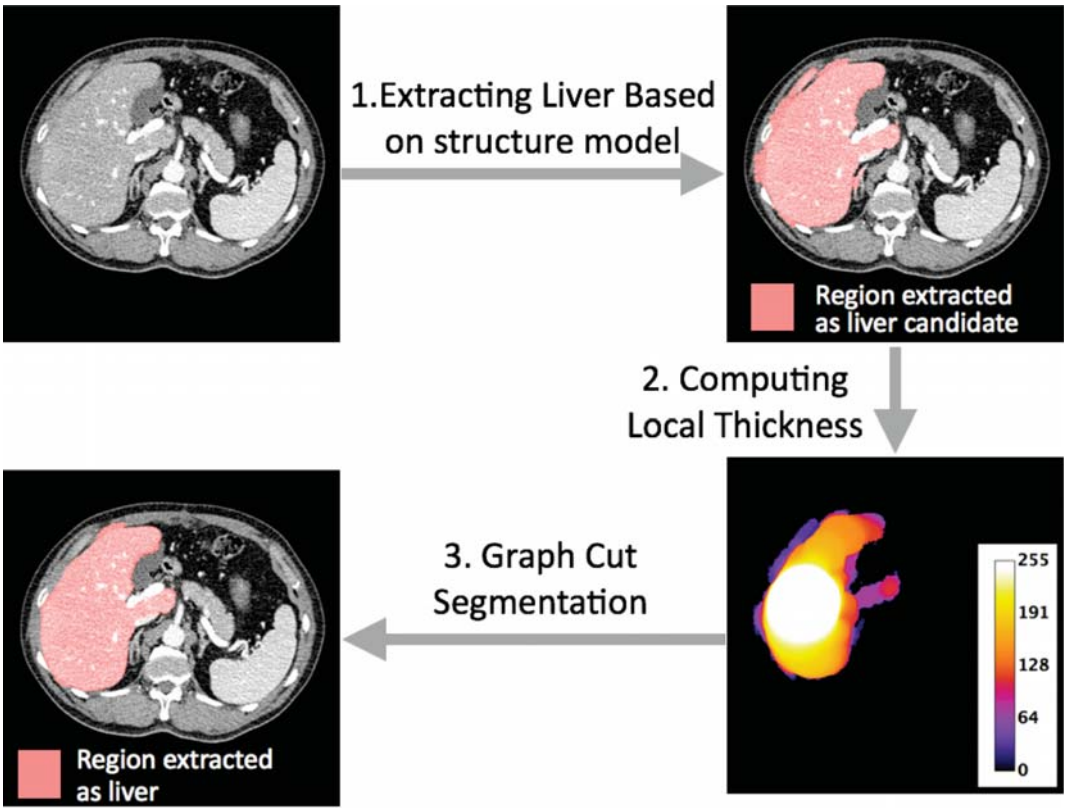


Fig. 2: Flow of extracting the liver regions.

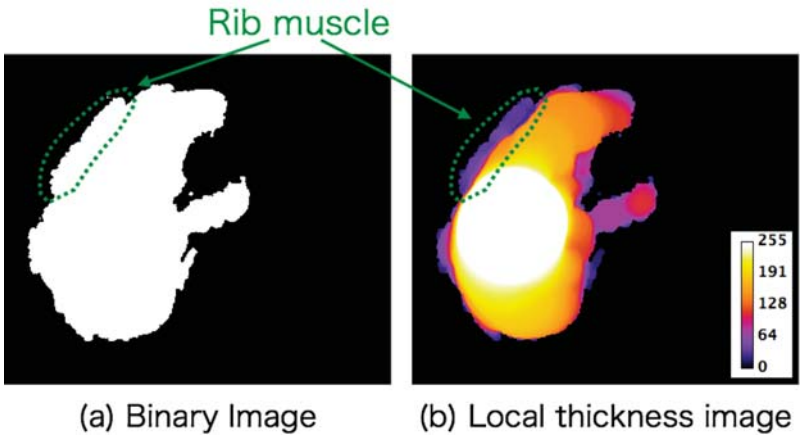


Fig. 3: Edge images: binary image and local-thickness image.

segmentation than original ImageJ. For the experiments, we use the plugin of graph cut segmentation which is implemented based on [8]. This plugin has for parameters to be preset as follows:

- the expected numbers of foreground pixels,
- the smoothness of the segmentation,
- the influence of the edge image,
- the variance of the edge image.

We adjust these parameters as needed.

Let us compare the proposed method(graph-cut segmentation using LT edge images) with segmentation based on structure models[5], and with graph-cut segmentation using binary edge images. Note that the segmentation based on structure model is used to obtain the liver candidate region in our extraction flow shown in Fig. 2. The comparison results are shown in Figures 4-6; the regions bounded with dotted lines indicate the mis-extracted regions; note that there are only over-extracted regions in this experiments. Each of these figures have (a) Grand truth, (b) Extraction based on structure model, (c) Graph-cut segmentation using binary edge images, and (d) Graph-cut segmentation using LT images(proposed). In these results, there are over-extracted areas in (b), and they are improved in the proposed method (d). In contrast, the results of (c) are almost same as those of (b) in Figs. 5 and 6, and is even worse in Figure. 4. The another advantage of the proposed method is that it is less sensitive than the Graph-cut segmentation using binary edge image.

4. Conclusion

The proposed method can obtain the accurate boundary of liver region based on the information of local thickness. As future work, setting parameters automatically is important issue. Moreover, simultaneous recognition of other organs around the livers is on-going to improve the accuracy.

References

[1] Y. Song, A.J. Bulpitt, and K. Brodlie, "Liver segmentation using automatically defined patient specific B-Spline surface models," MICCAI 2009 London, pp.43-50(2009).

[2] G.Linguraru,J.K.Sandberg,Z.Li,F.Shah,andR.M.Summers, "Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation," Medical Physics, vol.37, no.2, pp.771-783(2010).

[3] R.S. Alomari, S. Kompalli, and V. Chaudhary, "Segmentation of the liver from abdominal CT using Markov random field model and GVF snakes," Proc. 2008 International Conference on Complex, Intelligent and Software Intensive Systems, pp.293-298(2008).

[4] A.H. Foruzan, R.A. Zoroofi, M. Hori, and Y. Sato, "A knowledge-based technique for liver segmentation in CT data," Computerized Medical Imaging and Graphics, vol.33, no.8, pp.567-587(2009).

[5] Masanori Hariyama, Riichi Tanizawa, Mitsugi Shimoda, Keiichi Kubota, Yasuhiro Kobayashi, "Liver Extraction from CT Images Based on Liver Structure Models", Proc.International Conference on Image Processing, Computer Vision, and Pattern Recognition(IPCV), pp.170-173, (2014).

[6] R. Dougherty and K. Kunzelmann, "Computing local thickness of 3D structures with ImageJ", Proc. Microscopy & Microanalysis Meeting, www.optinav.com/LocalThicknessEd.pdf (2007)

[7] Y. Boykov, and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", In proc. International Conference on Computer Vision, vol. I, pp.105–112(2001).

[8] Y. Boykov, and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.26, no.9, pp.1124–1137(2004).

[9] Y. Boykov and G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation", International Journal of Computer Vision, vol.70, no.2, pp.109-âĂŞ131(2006).

[10] Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997-2015.

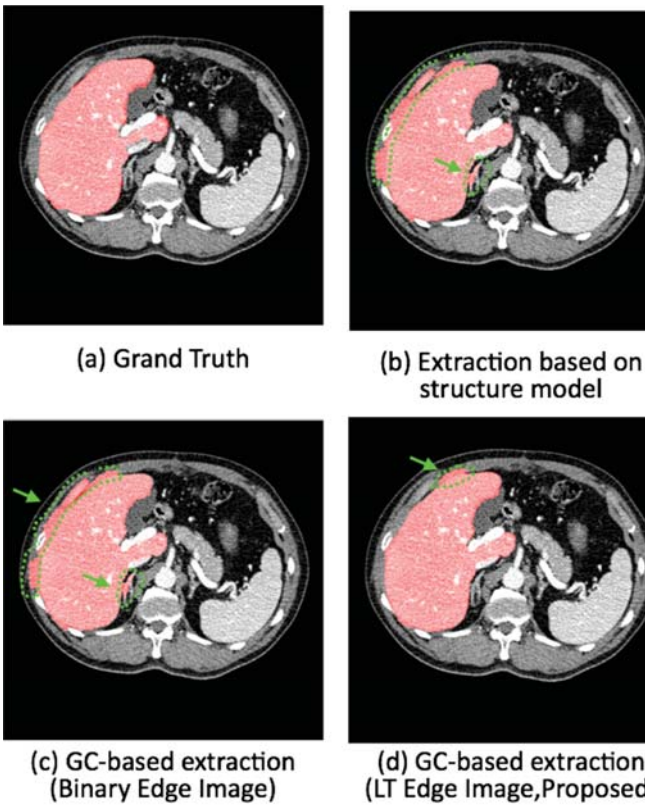


Fig. 4: Comparison 1.

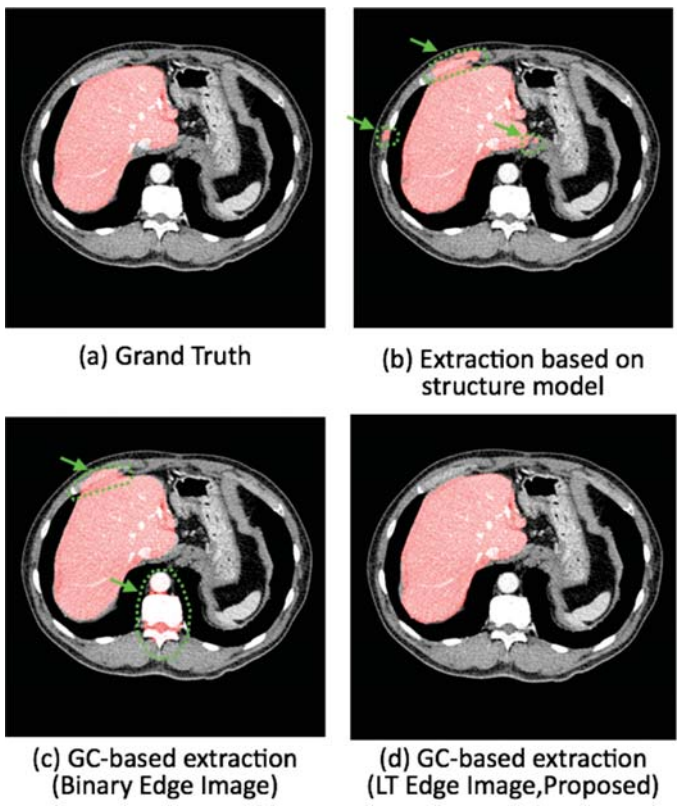


Fig. 5: Comparison 2.

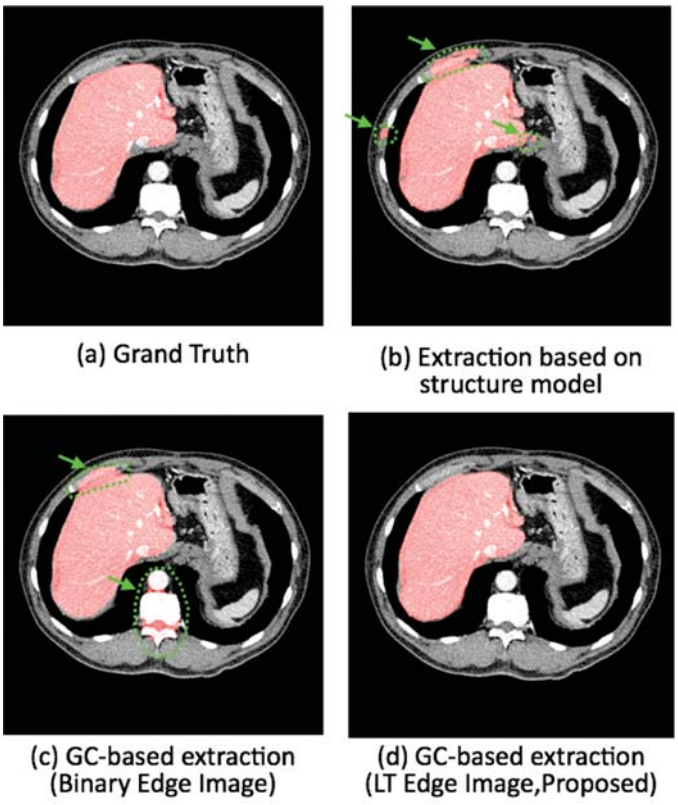


Fig. 6: Comparison 3.

A SMART "VIRTUAL EYE" MOBILE PROTOTYPE SYSTEM FOR THE VISUALLY IMPAIRED

David Zhou
Cypress Ranch High School
10700 Fry Road, Cypress, TX 77433

Yonggao Yang & Hanbing Yan
Department of Computer Science
Prairie View A&M University, Prairie View, TX 77446

Abstract: *This article presents a wearable mobile smart prototype system that can be used as the "virtual eyes" to assist visually impaired to navigate their surroundings. The smart system consists of two units: (1) an imbedded wearable smart sensor unit that uses ultrasound sensor to detect objects in front and sends the data to the other unit through Bluetooth; and (2) a smartphone that processes the detected object data from the other unit, utilizes the smartphone GPS unit to obtain orientation and location information, and feeds the visually impaired the critical surrounding information in real-time through Text-to-Speak (TTS) and/or vibrate warning signals.*

Keywords: *smart mobile device; wearable smart system; visually impaired*

I. INTRODUCTION

With over 285 million blind or visually impaired people in the world, technology serves and will continue to advance the welfare of these individuals. It is believed that 90% of the information the brain receives is through sight alone[1], thus an efficient yet economical device to serve as a distance sensor as well as provide other blind-assistance features and audibly relay the information back to the user would be highly significant.

Research in the field of mobile systems for the benefit of the visually impaired has primarily focused on specific object interaction through RFID[2] and MATLAB ultrasonic detection algorithms[3]. However, previous research does not provide users with a constant stream of object distances that the eyes usually provide. Furthermore, the use of smartphone capabilities has not been implemented in these devices which neglect the advantages of widespread smartphone applicability. Basic visual challenges such as knowing where the location of a car is or actual interpersonal human interactions are common in the lives of the visually impaired[4]. Thus, basic knowledge of these circumstances would greatly advance the lives of these individuals.

The latest relatively inexpensive smartphone devices, small and cheap imbedded microprocessor systems, and various smart sensor systems together make it feasible now to develop lost cost wearable smart system to assist visually impaired people to better "feel", "see", "hear", and "conceive" their surroundings, thus helping them move around and making their daily easier. The imbedded microprocessor allows us to develop wearable tiny device while still having enough data processing and computing capability. Various object detecting and distance sensors can help visually impaired people "see" the objects in front of them and "conceive" their distance. GPS chips can now easily and precisely tell these people their orientation and geographic location. Color sensors may be utilized to tell people with visual problem the traffic lights. Motion sensors can provide the probability of helping these people "feel" and "see" the moving objects along their moving direction. The smartphone is a mobile computer that now has enough data processing, computing and storing capability and excellent human interface to present visually impaired people critical information about their surroundings.

This project describes the application of imbedded microprocessors with ultrasound distance sensors in order to detect objects around the user. After the information is collected, the data is relayed via Bluetooth to smartphone that prompts the user with audio cues when an object is in their near proximity. Making use of other applicable features of the microprocessor and smartphone, further features such as voice guided GPS navigation and voice commands can be implemented in order to give visually impaired users a visual experience that has never been conceived before. Given the low costs of these devices, the design of this type of device is an innovative visual supplement for the blind and visually impaired with limited economic resources. This paper reports the novel conception of such a design, the optimization of the application, discussion of the results and future research development of this system.

In this article, we discuss and present a wearable smart system that is designed as "virtual eyes" to assist visually impaired to navigate their surroundings. The remaining of this article is organized as follows. Section II discusses the "virtual eyes" system structure. Section III concentrates on the software implementation of the system. In section IV, we present and discuss the system testing data. Section V concludes this work.

II. SYSTEM STRUCTURE

Continuously feeding the visually impaired with the necessary information about his/her surroundings is critical for him/her to "feel" and "sense" nearby and thus move around without bumping into objects and losing orientation. The wearable system presented in this section is designed to provide visually impaired such assistance.

The system consists of two major units: the imbedded wearable sensor unit and the smartphone unit. Figure 1 below is the system organization. The imbedded wearable small sensor unit has four major units: power unit, CPU unit, sensor unit, and the communication unit.

The power module relies on one 9-volt DC battery to provide this wearable unit the power.

The imbedded microprocessor (CPU) we selected for developing this prototype is the 32-bit "mbed NXP LPC1768" running at 96MHz from ARM Ltd [5] due to its lower price, very small size and lower power consumption, and most of all relatively very high computing performance. This microcontroller in particular is for prototyping all sorts of devices. It is especially armed with the flexibility of lots of peripheral interfaces and FLASH memory. The microprocessor includes 512KB FLASH, 32KB RAM and lots of interfaces including built-in Ethernet, USB Host and Device, CAN, SPI, I2C, ADC, DAC, PWM and other I/O interfaces. Most pins can also be used as DigitalIn and DigitalOut interfaces, which allow to easily adding various sensors. The online C++ software developing and compiling environment allows the development team to work on the project anywhere and anytime.

The communication module is a Bluetooth chip (or a WiFi chip) that allows the wearable unit to talk to the smartphone unit to exchange data and instructions. The LPC1768 microprocessor offers flexible hardware interfaces to for us to easily hookup various lower power consumption Bluetooth and WiFi chips that are available on the market, such as RN42-XV Bluetooth and XBee WiFi module.

The sensor module consists of various smart sensors we want to imbed to the system to obtain data. The ultrasound sensor detects objects and measures object distance from the unit. This is very similar to those sensors that are widely installed now on vehicles for preventing collision. In this prototype, we used "Ultrasonic Range Finder XL-MaxSonar-EZ4". Other sensors we plan to but have not plugged in yet include color sensors to detect traffic lights, motion sensors to detect moving objects in front of the visually impaired user, and a camera that can be used to provide front view for image processing to identify objects in interest.

The smartphone receives data obtained from the wearable smart sensor unit, processes the data in real-time, and generates and provides visually impaired user with critical warning and other information to assist him/her navigate the surrounding area in the format of voice and/or vibrating signal. For instance, when the sensor detects an object in front of the user within 3 meters, the smartphone sends beeps to user in the way that it beeps faster when the object becomes closer. At the same time, the smartphone utilizes its GPS and map function to "talk" to the visually impaired user his/her moving orientation and location (such as street names, building names, etc.).

The wearable smart sensor unit is designed to be clipped on a pocket, the belt, or the hat, and faces the user's moving direction so that the ultrasound sensor can "see" the objects in front of the user. The user should also carry the smartphone appropriately so that he/she can receive the correct moving orientation information. These two units talk to each other via Bluetooth; therefore there is no wire connection between them.

III. Software Implementation

The system software has two parts: the software running on the wearable smart sensor unit and the app running on the smartphone. The data and information flow within the system is depicted in Figure 2. Objects that are detected within three meters of the ultrasonic range finder have their distance relayed to the smartphone unit. Furthermore, further wearable sensors such as motion detection and color sensing can be implemented to provide a wider range of available sensing power. On the application side, GPS maps and walking directions can be applied using current Android GPS software.

3.1 Wearable Smart Sensor Unit Software

The NXP LPC1768 is programmable with C++ programming language through the online www.mbed.org

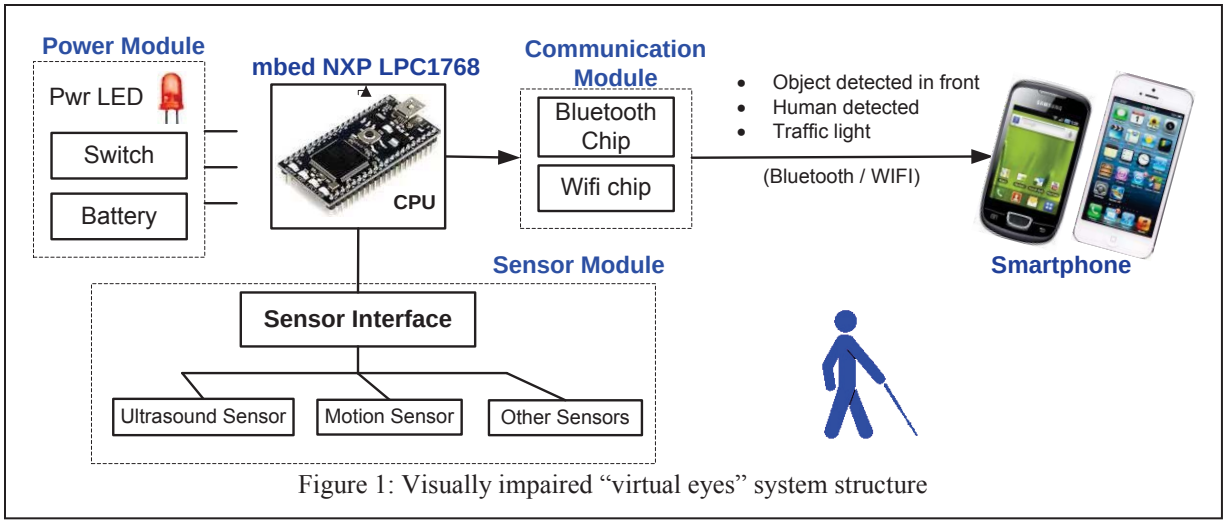


Figure 1: Visually impaired "virtual eyes" system structure

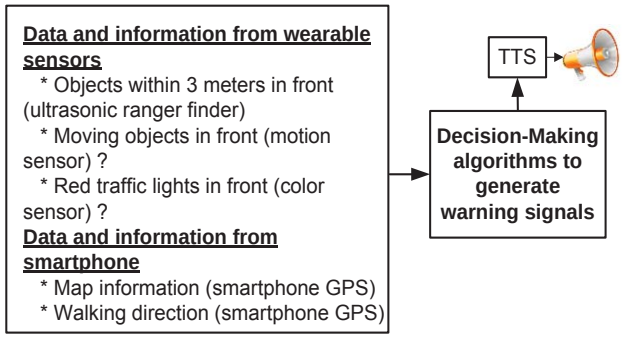


Figure 2: Software architecture

website. The successfully compiled executable code can be downloaded to the CPU through a USB cable.

The software at the wearable smart sensor unit has two major modules: the sensor module and the Bluetooth communication module. The sensor module initializes and controls the sensors, and read measurements from them. The raw readings from sensors are converted appropriately to meaningful data, such as distance, etc by using the sensors' formula. These data then are sent to smartphone unit through the Bluetooth communication module upon the request from smartphone.

3.2 Smartphone App

The prototype system at this stage is targeted on Android platform, which can be easily expended to iOS iPhone platform. The Android application was built in the Eclipse IDE with ADT package in order to fully utilize all the Android functionalities. Optimizing for the needs of the blind, the device uses simple commands via both touch for testing purposes and voice for the final product in order to communicate the sensor data to the user audibly.

The application begins with a click on the application icon in order to launch it. Once the application is launched, the user is provided with a variety of Bluetooth devices that can be programmed to remember previously used devices as well as automatically linking with the sensor. The entire process as well as future work is shown in Figure 1.

The life of the application begins with a click on the application in order to launch it. Once the application is launched, the user is provided with a variety of Bluetooth devices that can be programmed to remember previously used devices as well as automatically linking with the sensor.

Once linked, the application continuously listens for any sensor data relayed to the device. The sensor board is programmed in order to relay distance readings from one to ten feet. This way, the user experiences a minimal hands-free interaction with the device which is highly desirable for the visually impaired.

Within the project package, the application uses three classes for functionalities. The BlueDuino.java extends the Activity class to serve as the hub of the application and to perform all its functions. Another class, the BluetoothConnection.java, creates, receives, and sends the flow of data between the Android application and the imbedded microprocessor. The third class, DeviceListActivity.java, saves the list of all previously detected Bluetooth devices.

Once a successful connection is created, the application automatically receives distance notifications at set time intervals to the nearest tenths of an inch. The application then relays the information to the user in an audible fashion to ensure hands free functionality. For initial testing purposes, simple buttons are used to send the request for data (Send "D" Button). Figure 3 shows a screenshot of the application at work.

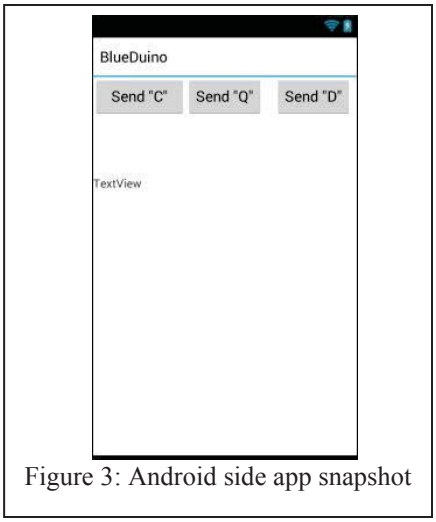


Figure 3: Android side app snapshot

IV. System Experimental Data

4.1 Initial Testing

In this section, the system was tested for its accuracy and precision. The implications of such results are discussed below. Optimization of the sensing data is also shown in this section. The sensor used its maximum threshold of 254 inches with a .25 inch diameter width range in either direction[6].

The Android application system was tested along with the MaxSonar sensor in an experimental environment. Testing was implemented through distance detection of a standard trash can (18 by 23 by 36 inches) in order to replicate real life obstacles the user might experience. The object was placed at varying lengths ranging from one foot up to ten feet away at one foot increments in order to test the accuracy and consistency of the sensor data. The results obtained are shown in Table 1 (the unit of all the measurements is in inch). The results obtained above indicate that the MaxSensor's distance data has low residuals from the average of the detected distances. However, as distance increases, the value of the residuals from the actual distance increase in a positive fashion as represented in Figure 4. Therefore, a mathematical model is needed in order to minimize the deviations from the actual in order to produce a more accurate reading.

Table 1: Actual distance vs. reading from sensor

Obj. Dist.	Reads					Avg	STD Dev.
	1	2	3	4	5		
12	13.3	13.2	12.6	12.0	13.0	12.8	0.47
24	27.3	27.6	27.5	27.6	26.0	27.2	0.61
36	41.2	41.6	41.4	41.5	41.6	41.5	0.15
48	55.0	56.0	55.9	56.1	55.7	55.7	0.39
60	69.2	69.0	69.1	68.6	68.6	68.9	0.25
72	82.4	82.5	82.8	82.8	82.3	82.6	0.21
84	96.5	93.4	96.5	97.0	97.0	96.1	1.36
96	110.1	110.2	110.2	110.5	110.1	110.2	0.15
108	129.2	129.2	128.9	127.7	128.2	128.6	0.60
120	139.5	139.8	140.4	139.9	139.9	139.9	0.29

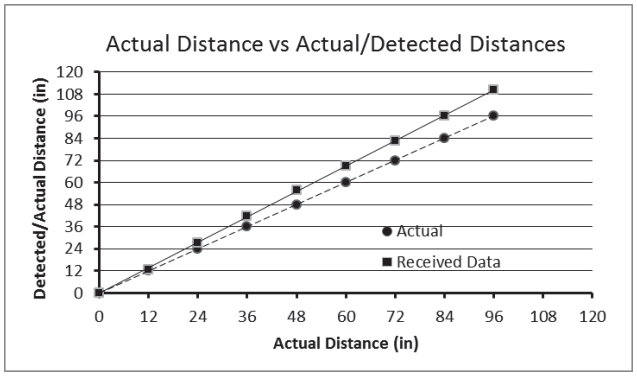


Figure 4: Actual object distances vs detected distances

4.2 Optimization of Sensing Data

As shown in Figure 4, the residuals increase as the actual increases, thus allowing us to use a linear relationship to optimize the detected value line to match the actual values. With the values of the residuals obtained, a graph of the actual distances vs residual values was made with a trend line in order to describe the overall progression of the residual values. The 96 and 120 inch measurements were discarded due to their extraneous values given the other more linear data.

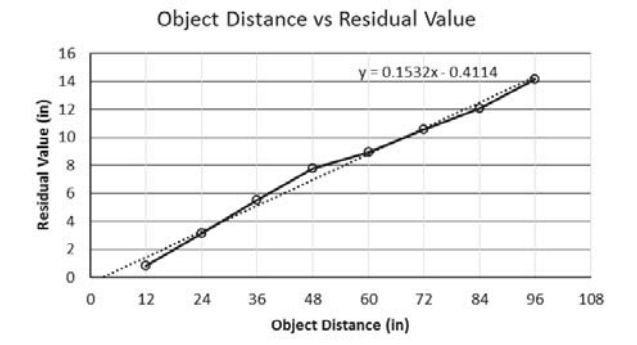


Figure 5: Object distance (in) vs Residual values (in)

The graph of the residuals is show in Figure 5. By using a trend line to show a good linear equation, **`y = 0.1532 x - 0.4114`**, between the distances, the value of the residual is able to be predicted by the received data value. Thus, by subtracting the predicted residual value from the received data value, the final result has far less residual to the actual distance.

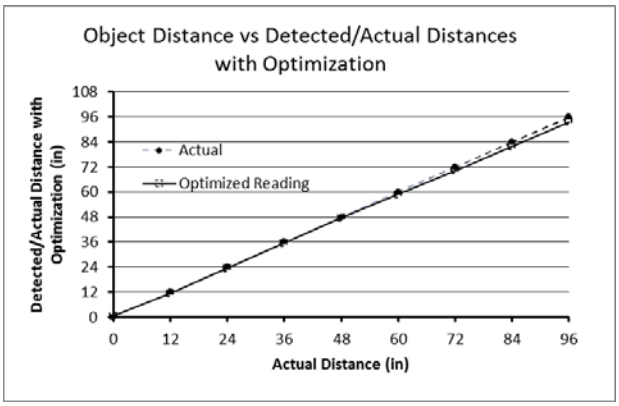


Figure 6: Obj. dist. vs detected dist. with optimization

The graph of the object distance vs detected/actual distances after optimization is applied is shown in Figure 6**.**

V. CONCLUSION

A novel Android application system with distance sensing for the visually impaired has been successfully designed. The application is robust and able to accurately detect objects up to ten feet away, thus giving users an interactive audible display of nearby proximity objects. The efficient design coupled with its low costs makes it an intriguing device with worldwide application. Future research will primarily focus on developing the application system outside of a controlled laboratory environment. New focus will be placed on designing the system to be particularly robust in everyday user interactions with the real world, including GPS navigational systems and voice commands.

REFERENCES

[1] "Visual Impairment and Blindness." WHO. N.p., n.d. Web. 10 Nov. 2014. <http://www.who.int/mediacentre/factsheets/fs282/en/>.

[2] Dionisi, A.; Sardini, E.; Serpelloni, M. "Wearable object detection system for the blind", *Instrumentation and Measurement Technology Conference (I2MTC), 2012 IEEE International*, On page(s): 1255 – 1258.

[3] Rastogi, R.K.; Mehra, R. "Application of temperature compensated ultrasonic ranging for blind person and verification using MATLAB", *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, On page(s): 1154 – 1158.

[4] Brady, Erin, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. "Visual Challenges in

the Everyday Lives of Blind People." (n.d.): n. pag. *Https://cs.rochester.edu/hci/pubs/pdfs/chi2013-vizwiz.pdf*. University of Rochester. Web. Nov.-Dec. 2014.

[5] ARM mbed microprocessor, http://developer.mbed.org/platforms/mbed-LPC1768/; accessed on Nov. 15, 2014.

[6] Owner. "MaxSonar Product Description." PD10001d-MB1000 Datasheet(n.d.): n. pag. MaxSonar. Web. 12 Nov. 2014.

Book reader optimization using a time of flight imaging sensor

L. Galarza, Z. Wang, and M.Adjouadi
Center for Advanced Technology and Education, College of Engineering and Computing
Florida International University Miami, Florida, USA

***Abstract*—This study expands on our previous approach for correcting warping in images of a book spread, which in turn can be utilized to improve character recognition. The correction is made possible via depth map extraction using a time of flight sensor. Curvature corrections are achieved by modifying the lens equation to take into account different height points on the book spread. The corrections have been expanded not only to flatten the page, but to extend accordingly the characters in the image. In addition, current resolution limitation from the time of flight device is overcome by scaling and matching the pixel depth data to an image taken from a camera with higher resolution. The Implementation results support the assertion of improved reading accuracy, which in turn highlights the merits of using this approach based on the depth maps in order to correct for the book curvature.**

Keywords—Time of flight; depth maps; curvature correction; book reader; reading accuracy, OCR

I. INTRODUCTION

The proper digitalization of information from books and documents is a great undertaking; to this end different tools to capture and convert the information have been proposed [1-3]. Furthermore, once the contents of a book are digitized, it can be made more readily available to visually impaired individuals as part of automated book readers or other reading systems [4-7]. However, during the capturing process, warping effects can occur, which can make the character recognition process very challenging. Consequently, there are a number of varying approaches that are proposed to correct for the warped effect on book spread images in order to improve character recognition. These correction mechanisms are essentially made using 1) various rectification algorithms [8-10] on the basis of perspective effect, structure and geometric measure; 2) using image restoration [11-13] through 3D shape reconstruction and modeling; and 3) other specific de-warping or flattening mechanisms [14-19] through the use of grids and the so-called developable surfaces which take into consideration the concepts of parallelism between the lines and equal spacing in order to project the image onto a flat surface.

In this particular book reader design, a different method employing a time of flight device augmented through a lens correction approach is utilized to improve the character recognition rate (reading accuracy) of an Optical Character Recognition (OCR) engine. The approach relies on the ability to obtain an accurate depth map of the book spread through the use of a rather inexpensive and lightweight 3D camera. Possible depth recovery devices include those that are based on stereo imaging, scatter light, laser scanning and time of flight. In this case, the time of flight device was chosen due to its higher processing speed and light weight while still maintaining a good level of depth accuracy. These characteristics have led to a proliferation in the use of the time of flight devices in a number of applications [20-22]. A current drawback from this type of device is that it currently produces a low resolution, which can be overcome by pairing it with a higher resolution device, as has been proposed in other methods [23], [24]. The corrections will be made on the higher resolution image, where the warped effect is most noticeable. In the following sections, it will be shown how the height was incorporated as part of the modified lens equation to determine the unwarped pixel locations. The validity of this approach for the book reader will be made possible via an experimental setup, which should produce a page flattening and character expanding effect to improve the legibility of the characters of the book spread's image. The results obtained proved the feasibility of the proposed approach.

II. METHODS

A. Book spread correction and updates

Previously, the warping effect in an image was ameliorated by keeping the distance fix from the camera to the object so that the factor affecting the display can be attributed to the height of the object [25]. The desired non-affine transformation to flatten the book spread was derived and yielded the following equations:

$$x = x^* \times \big(U - h(x^*, y^*)\big)/U, \text{and}$$

$$y = y^* \times \big(U - h(x^*, y^*)\big)/U \quad (1)$$

To further improve the results obtained by using equations (1) and make the correction more fluid and natural, it was observed that the image of a book spread should be extended. This behavior can be observed as a book is pressed on a scanner or copier to have its curvature reduce which in turn causes the book to be flattened and the pages to be extended. The extend location can be obtain by using the change in height

$$dh_i = h_{i+1}(x^*, y^*) - h_i(x^*, y^*) \quad (2)$$

As well as the change along the x direction

$$dx_i = x_{i+1} - x_i = 1\ pixel \quad (3)$$

The new extended lengths (L_i) can then be calculated by using (2) and (3):

$$L_i = (dh_i - dx_i)^{(1/2)} \quad (4)$$

Using (4) it is possible to express the new extended (xt_i) locations as:

$$xt_{i+1} = \sum_{i=0}^{n}(xt_i + L_i) \quad (5)$$

Where $xt_0 = 0$ and n = the number of rows. This extension adjustment is applied after the flatting transformation along all points has been performed. In turn, this will allow the image to be corrected and the warping effect on the book spread to be attenuated to a greater extent than was previously noted as the OCR results would confirm.

B. Experimental setup

The experimental set up remains as in Fig 1 where the Argos3D-P100 depth sensor, developed by Bluetechnix [26] serves as means of depth extraction. The depth perception of this sensor is based on the Time of Flight (ToF) principle. Ideally this would be the only device that would be need for the book reader. However, the resolution is only 160 x 120 pixels and for the book's characters to be legible a higher resolution image is required. In this case the Cannon G6 is used to obtain the higher resolution image. A second L shape brace can be added to the top of the setup and can serve to adjust the position of the cameras independently which is very useful for alignment purposes. However, the two devices must remain as close as possible in order to reduce the number of occlusions which can occur. Also, excessive and large portions of black colors should be remove from the book viewing area since this can cause signals to be absorbed that the Argos 3D-P100 uses for depth estimation which is why a brown color was selected for the base of the book reader.

C. Implementation

The first step which only has to be done once to ensure the accuracy of the depth map is the calibration process which can resemble [27]. Specifically, for this application, the distance for the sensor was set to 100cm which is a little more than the distance between the Argos3D-P100 and the surface of the book holder platform. The native bilateral filter was used since this has been shown to improve depth resolution [28]. Also, to reduce extreme variances from the depth measurements which can lead to inaccuracies, the average of 50 depth maps samples was used as the raw depth map.

Once all the alignment and calibration has been performed the height and pixel amplitude can be obtained. The height of the book is calculated by taking the averaged depth map from the book holder platform and subtracting it from the averaged depth map of the book spread. This height is the raw height:

$$h(x^*, y^*) = Raw\ Height \quad (6)$$

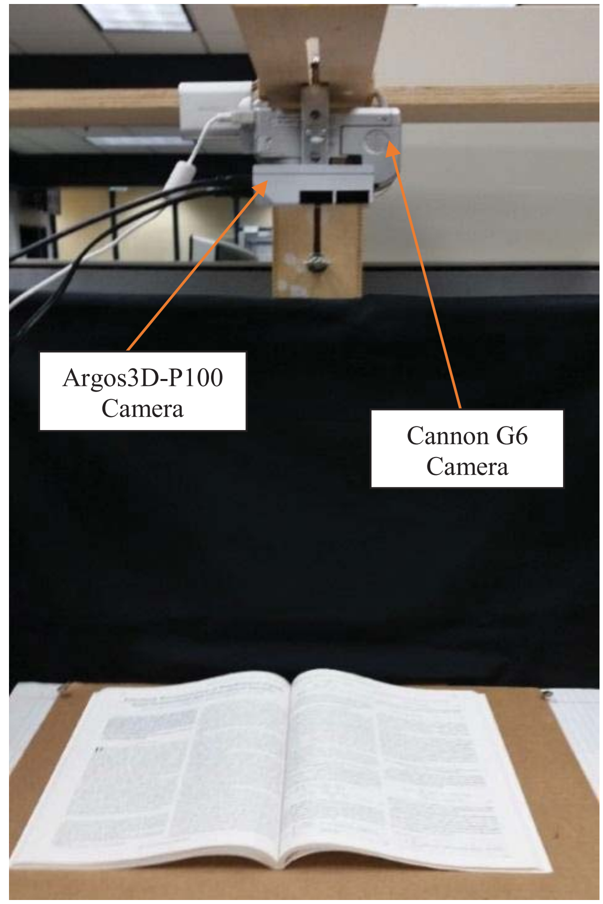


Fig. 1. Experimental Setup

In these initial steps, the height can also be estimated using polynomial approximations of order n with coefficients p_{ij} as best fit of the height points for each row, on the basis of the Argos3D-P100 resolution, as given in equation (7) [25]:

$$h(x^*, y^*) = \begin{cases} p_{11}x^{*n} + p_{12}x^{*n-1} + \cdots + p_{1n}x^* + p_{1n+1} & , y^* = 1 \\ \vdots & \vdots \\ p_{m1}x^{*n} + p_{m2}x^{*n-1} + \cdots + p_{mn}x^* + p_{mn+1} & , y^* = m \end{cases} \quad (7)$$

The additional values that are needed to perform the corrections are the value of U (for equation 1) and the actual distance cover between pixels in the high resolution image (for converting units before applying equation 4). Earlier experimental evaluations showed that an appropriate value for U in this case is calculated to be 58.05 cm and the distance cover between pixels is 2668/58 (Pixel/cm) for this experimental setup.

Another critical step is to match the low resolution image of the Argos3D-P100 to the higher resolution image of the Canon G6 camera in order to improve the OCR results. Image registration is based on matching 3 landmark points in both images (two along the spine of the book and one at a corner). With the registered images it is then possible to interpolate and obtain the new high-resolution height map, $h(x^*, y^*)$. The height at these points can be used as raw data as given in (6) or

via polynomial approximation as expressed in (7) to then initiate the page flattening process. The high resolution image is reconstructed via 2D cubic interpolation, and the flattened image can then be extended by calculating the change in height and the change along the x direction using equations (2) and (3), respectively. It should be noted that either one of these equations should be converted so that the units match (i.e. in pixels or in centimeters). In our case, we previously determined, given the geometry of the experimental setup, the conversion value to be 2668/58 (Pixel/cm). Hence, once converted, equation (4) can be used to calculate the extended lengths as expressed in (5). The image can then be reconstructed via scattered interpolation using the methods of linear interpolation, natural neighbor interpolation, and nearest-neighbor interpolation. Then, the resulting images can then be introduced to the OCR engine for character recognition and for the book to be read for a person with visual impairment or digitized and saved in memory for future use. In our case the ABBYY FineReader 12 OCR engine was used.

III. Results

The initial findings for the flattening correction have been presented in an earlier study [25]. Thus, the following results will focus on the effects of the OCR on the flattening correction and as well as the page extension approach. In Fig. 2 we can see an illustrative example of a warped image from the book curvature.

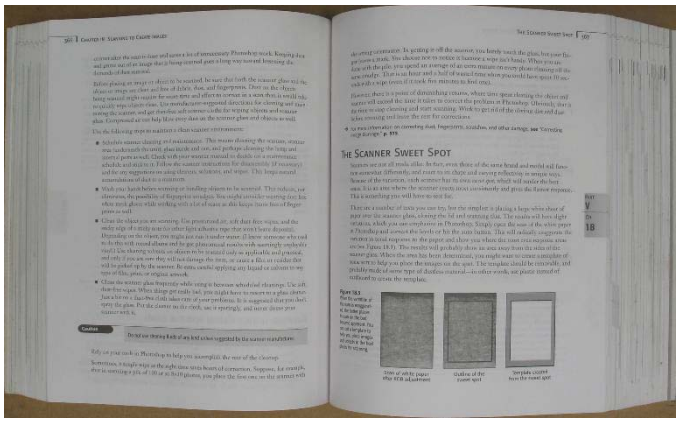
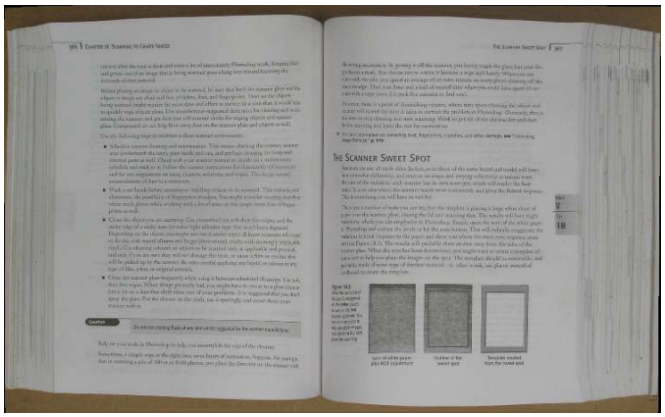
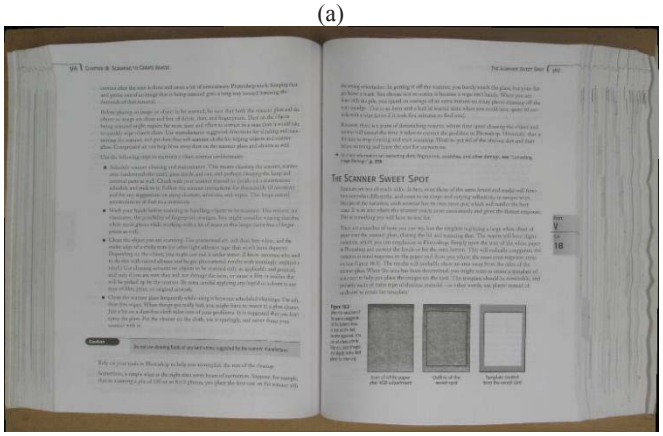


Fig. 2. Higher resolution image (warped)

The flattening transformation process is applied using both the raw height data and the polynomial approximation of the curvature by means of equation (1) in order to produce the corrected images as exemplified in Fig.3 (a) and (b), respectively. A zoom in on the text of these two examples is provided in the following figures for visual appreciation as well as for the assessments of the results of the OCR before and after curvature correction. The results for warped book spread source, the corrected image with the raw height and the corrected image with polynomial fitted heights can be seen in Fig.4, Fig.5 and Fig.6 respectively. Table 1 summarizes the improvements obtained from the OCR engine from these image captures which for this particular example contained 129 words.



(a)

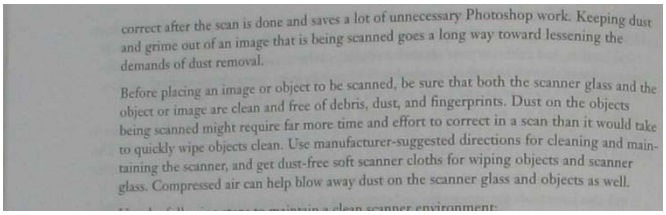


(b)

Fig. 3. Corrections using (a) raw height, and (b) polynomial fitting

Table 1.Sumary of OCR results for Fig.4, Fig.5 and Fig. 6

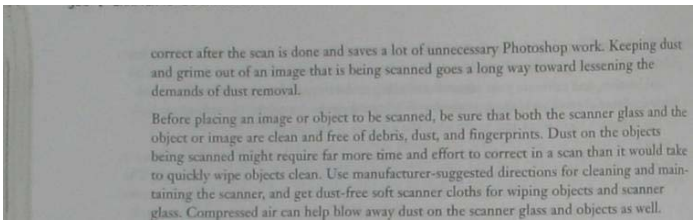
	Warped	Flat Raw	Flat PolyFit
# word errors	19	5	7
correct %	85.27	96.12	94.57



correct after the scan is done and saves a lot of unnecessary Photoshop work. Keeping dust and grime out of an image that is being scanned goes a long way toward lessening the demands of dust removal.

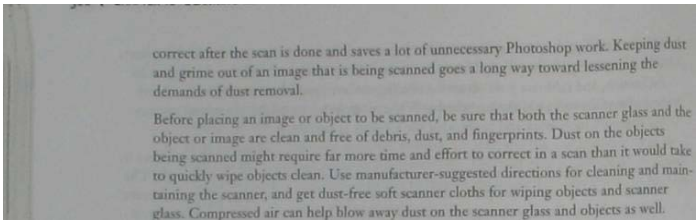
Before placing an image or object to be scanned, be sure that both the scanner glass and the object or image are clean and free of debris, dust, and fingerprints. Dust on the objects being scanned might require far more time and effort to correct in a scan than it would take to quickly wipe objects clean. Use manufacturer-suggested directions for cleaning and maintaining the scanner, and get dust-free soft scanner cloths for wiping objects and scanner glass. Compressed air can help blow away dust on the scanner glass and objects as well.

Fig. 4. Results of the OCR engine on the original warped image



correct after the scan is done and saves lot of unnecessary Photoshop work. Keeping dust and grime out of an image that is being scanned goes a long way toward lessening the demands of dust removal.
Before placing an image or object to be scanned, be sure that both the scanner glass and the object or image are clean and free of debris, dust, and fIngerpiints. Dust on the objects being scanned might require far more time and effort to correct in a scan than it would take to quicIJv wipe objects clean. Use manufacturer-suggested directions for cleaning and maintiiming the scanner, and get dust-free soft scanner cloths for wiping objects and scanner gla.cs. Compressed air can help blow away dust on the scanner glass and objects as well.

Fig. 5. Results of the OCR engine on the raw corrected image



correct after the sea n is done and saves a lot of unnecessary Photoshop work. Keeping dust and grime our o lin image that is being scanned goes a long way toward lessening the demands of dust removal.
Before placing an image or object to be scanned, be sure that both the scanner glass and the ubiect or image are clean and free of debris, dust, and fingerprints. Dust on the objects being scanned might require far more time and effort to correct in a scan than it would take to quickly wipe objects clean. Use manufacturer-suggested directions for cleaning and maintaining the winner, and get dust-free soft scanner cloths for wiping objects and scanner glass. Compressed air can help blow away dust on the scanner glass and objects as well.

Fig. 6. Results of the OCR engine on the polynomial corrected image

The results do show that there is a significant improvement on the OCR recognition when using the raw height data. However, when using the polynomial fitting approach, the results improved slightly. In this last case, although the intent was to have a more continuous rendering of the height data, the mathematical approximations of the polynomial may have affected the raw heights for some of the characters in contrast to using the raw data as they are. The flattening approach was further tested under poor lighting conditions and with a thinner glossy book spread, the Argos3D-P100 was still able to retrieve an adequate depth profile of the book spread, but with some errors. This particular deficiency of the Argos3D-P100 camera is expected since it uses active IR illumination. Fig. 7 shows an example of the high-resolution image under the assumed poor conditions.

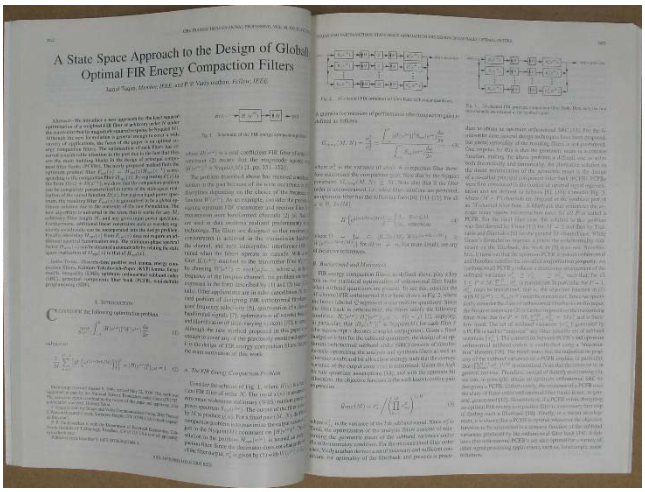
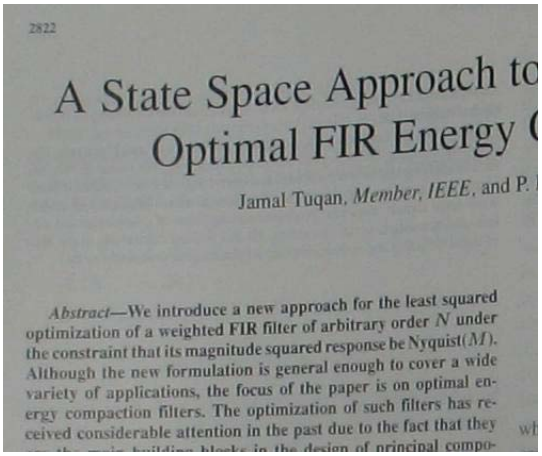


Fig. 7. Higher resolution image (warped) with inadequate lighting

Again, in order to better appreciate the effect of the correction, a small section of the text was introduce into the OCR engine which the visual results can be seen in Fig. 8, Fig. 9 and Fig. 10. Table 2 summarises the improvements obtained from the OCR engine from these images, which in this case contained 89 words.

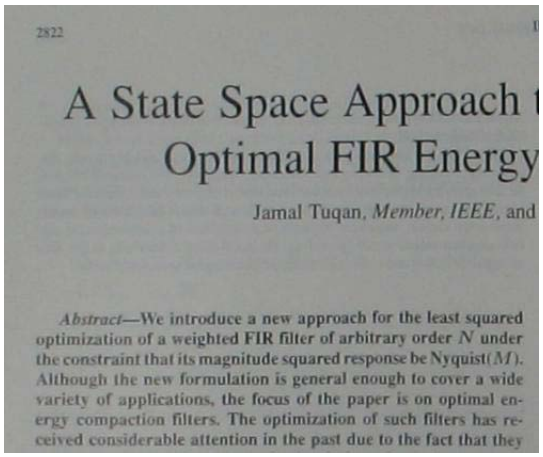
Table 2.Result summary of OCR engine with spell check

	Warped	Flat Raw	Flat PolyFit
# word errors	23	20	19
correct %	74.16	77.53	78.65



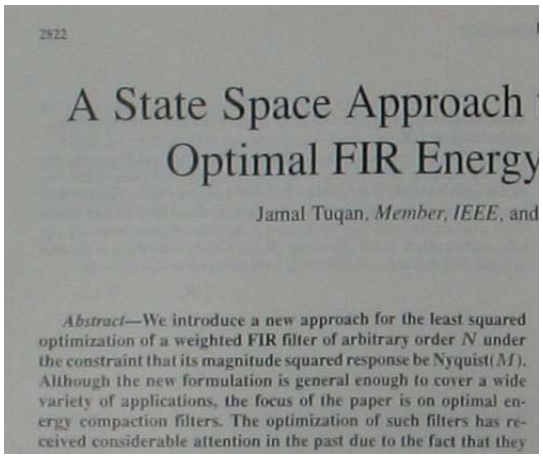
A State Space Approach t Optimal FIR Energy
Jamal Tuqan. &h'mbt'r. IEEE. and P.
4bssma—We introduce a nels approach for the least St4LIaITd optimization of a ueighted FIR filler of arhitrari order N under the constraint that iL magnitude quared response be Nquttt!IIlI). lthough the new formulation is general enough to cnsier a alide .ariet ui applications, the locus of the paper is on optimal C0 er compaction filters. The optimizaiion or such Pliers has cccei ed considerable attention in the past due In he fact that they w

Fig. 8. Results of the OCR engine on the original warped image



A State Space Approach 1 Optimal FIR Energy
Jamal Tuqan. Member, IEEE. and
.4bihacl—Sr introduce a new approach for the least squared optimiz ution of a weighted FIR filter of arbilrar order N under (he constraint that u.s magnitude squared response he %quisll ?1i. .lthough the ness formulation is general enough to cos er a iside sarich of dpplications. the focus or the paper is on optimal enerv compaction filters. The optimization of such fitters has ite-eised considerable altcnhion In the pasi due to the fact that tbt

Fig. 9. Results of the OCR engine on the raw corrected image



A State Space Approach Optimal FIR Energy
Jamal Tuqan. Me,nher. IEEE. and
4bWT,ci—S iniroduce nw approach for the least squared optimization of a weighted FIR ruler of arhitrars order N under thv constraint that uLs magnitude squared response be iquist MI. Although the ness formulation is general enough to casey a ide sariet or applications, the focus at the paper is on optimal enerv compaction filers. The optimization of such fliers has receised cons.derable attention In the past due to the fact that (hes

Fig. 10. Results of the OCR engine on the polynomial corrected image

Earlier work by a research group with the Center for Advanced Technology and Education has made use of the ABBYY engine. However, that previous approach relied on using two cameras to estimate the height and perform a different correction [29]. The following is an excerpt of their finding using the older OCR engine of ABBYY FineReader 8.0

Table 3. Excerpt of results for old approach using ABBYY FineReader 8.0

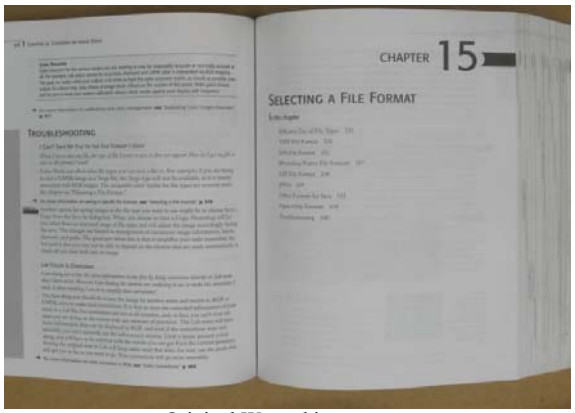
Text #	Word count	errors original image	Reading accuracy	errors corrected Image	Reading accuracy
1	464	267	42.46%	8	98.28%
2	383	85	77.81%	15	96.08%
3	944	400	57.63%	15	98.41%
4	963	300	68.85%	24	97.51%
5	496	192	61.29%	9	98.19%
6	456	180	60.53%	21	95.39%
7	513	243	52.63%	17	96.69%
8	898	560	37.64%	29	96.77%
9	567	320	43.56%	11	98.06%

The ABBYY OCR engine has been greatly improved since , and it may now include some forms of geometrical distortion correction [30] as part of their new versions. Such improvements can be seen in the many examples that were considered in Table 4. The results shown in Table 4 are obtained using the newer version of ABBYY FineReader 12.0, focusing only on the raw heights, which as we have seen earlier, provided better results. In retrospect, the newer ABBYY version performed rather well; still the proposed method which corrected the page curvature using the heights as extracted by the Argos3D-P100 Camera performed in general better in the majority of cases considered in Table 4.

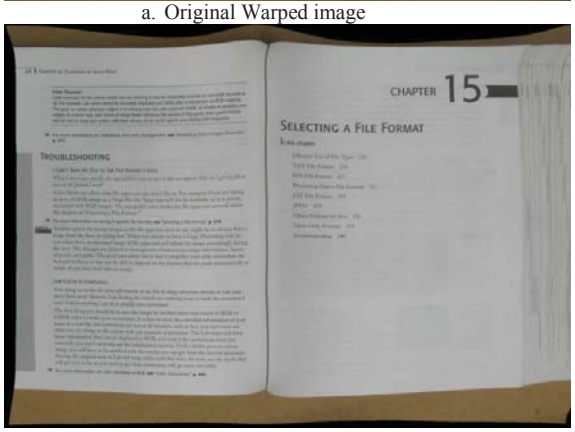
Table 4. Results of just Flattening using ABBYY FineReader 12.0

Text #	Word count	errors original image	Reading accuracy	errors corrected Image	Reading accuracy
1	368	57	84.51%	19	94.84%
2	601	48	92.01%	27	95.51%
3	480	70	85.42%	30	93.75%
4	614	76	87.62%	16	97.39%
5	286	15	94.76%	4	98.60%
6	610	83	86.39%	10	98.36%
7	305	3	99.02%	4	98.69%
8	472	20	95.76%	11	97.67%
9	506	32	93.68%	7	98.62%

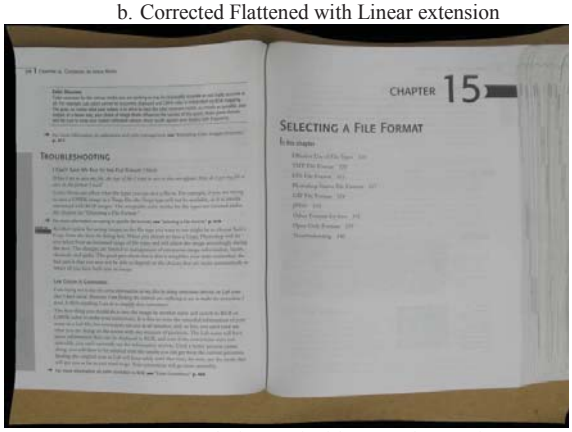
Fig. 11 shows examples of the original and corrected via flattening with extension. The OCR results for these can be seen in entry #3 of Table 5, Table 6, and Table 7 respectively.



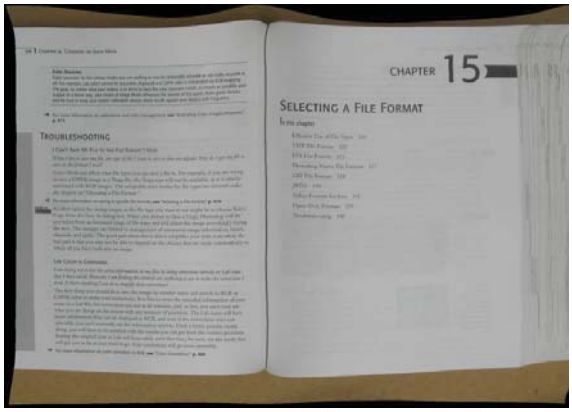
a. Original Warped image



b. Corrected Flattened with Linear extension



c. Corrected Flattened with Natural extension



d. Corrected Flattened with Nearest extension

Fig. 11. Example of the Original and corrected images respectively

To the naked eye all 3 types of corrections of flattening with extension (linear, nearest, and natural) as shown in previous figures look the same, but as we can see in the following tables they do have different impacts on the OCR performance.

Table 5. Results of Flattening with extend (linear) using ABBYY FineReader 12.0

Text #	Word count	errors original image	Reading accuracy	errors corrected Image	Reading accuracy
1	368	57	84.51%	22	94.02%
2	601	48	92.01%	42	93.01%
3	480	70	85.42%	32	93.33%
4	614	76	87.62%	24	96.09%
5	286	15	94.76%	4	98.60%
6	610	83	86.39%	10	98.36%
7	305	3	99.02%	2	99.34%
8	472	20	95.76%	12	97.46%
9	506	32	93.68%	9	98.22%

Table 6. Results of Flattening with extend (natural) using ABBYY FineReader 12.0

Text #	Word count	errors original image	Reading accuracy	errors corrected Image	Reading accuracy
1	368	57	84.51%	16	95.65%
2	601	48	92.01%	33	94.51%
3	480	70	85.42%	39	91.88%
4	614	76	87.62%	12	98.05%
5	286	15	94.76%	2	99.30%
6	610	83	86.39%	9	98.52%
7	305	3	99.02%	4	98.69%
8	472	20	95.76%	14	97.03%
9	506	32	93.68%	7	98.62%

Table 7. Results of Flattening with extend (nearest) using ABBYY FineReader 12.0

Text #	Word count	errors original image	Reading accuracy	errors corrected Image	Reading accuracy
1	368	57	84.51%	16	95.65%
2	601	48	92.01%	29	95.17%
3	480	70	85.42%	23	95.21%
4	614	76	87.62%	14	97.72%
5	286	15	94.76%	1	99.65%
6	610	83	86.39%	10	98.36%
7	305	3	99.02%	8	97.38%
8	472	20	95.76%	7	98.52%
9	506	32	93.68%	7	98.62%

From these tables, it can be observed that the image reconstruction does affect the result. Both nearest and natural methods of interpolation appear to have the best results.

IV. Conclusion

The results showed that with the proposed mathematical derivations, the image can be corrected and the warping effect is attenuated, which in turn improves the legibility of the characters and the reading accuracy of the OCR results. Furthermore, even though the accuracy of the OCR engines such as ABBYY has improved in its reading accuracy, it is still possible to make additional enhancements for a higher reading accuracy. However, it should be noted that this approach could have negative effects on those OCR engines that already embed within their design certain kinds of geometry correction algorithms since this more natural correction process based on raw height (depth) data does not assume uniformity of the distortion or curvature across the image.

Acknowledgments

This work is supported by the National Science Foundation through grants CNS-0959985, CNS-1042341, HRD-0833093, and IIP-1230661. The support of the Ware Foundation is greatly appreciated.

References

[1] Stephen Pollard, Maurizio Pilu. "Building cameras for capturing documents." *International Journal of Document Analysis and Recognition*. 7.2-3 (123-137) (2005).

[2] Karen Coyle. "Mass digitization of books." *The Journal of Academic Librarianship*. 32.6 (641-645) (2006).

[3] Muhammad Muzzamil Luqman, Gomez-Krämer Petra, Jean-Marc Ogier. "Mobile phone camera-based video scanning of paper documents." *Camera-Based Document Analysis and Recognition*. (164-178) (2014).

[4] Malek Adjouadi, Eddy Ruiz, Lu Wang. "Automated book reader for persons with blindness." *Computers Helping People with Special Needs*. (1094-1101) (2006).

[5] Elia Contini, Barbara Leporini, Fabio Paternò. "A Semi-automatic Support to Adapt E-Documents in an Accessible and Usable Format for Vision Impaired Users." *Computers Helping People with Special Needs*. (242-249) (2008).

[6] Antonello Calabrò, Elia Contini, Barbara Leporini. "Book4All: A tool to make an e-book more accessible to students with vision/visual-impairments." *HCI and Usability for e-Inclusion*. (236-248) (2009).

[7] Bassam Almasri, Islam Elkabani, Rached Zantout. "An Interactive Workspace for Helping the Visually Impaired Learn Linear Algebra." *Computers Helping People with Special Needs*. (572-579) (2014).

[8] Liang, Jian, Daniel DeMenthon, and David Doermann. "Geometric rectification of camera-captured document images.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 30.4 (591-605) (2008).

[9] Hyung Il Koo. "Segmentation and rectification of pictures in the camera-captured images of printed documents." *Multimedia*. 15.3 (647-660) (2013).

[10] Yihong Wu, Zhanyi Hu, Youfu Li, "Radial distortion invariants and lens evaluation under a single-optical-axis omnidirectional camera." *Computer Vision and Image Understanding*. 126 (11-27) (2014).

[11] Michael S. Brown, W. Brent Seales. "Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents."*Computer Vision, ICCV 2001*. 2 (367-374) (2001).

[12] Atsushi Yamashita, et al., "Shape reconstruction and image restoration for non-flat surfaces of documents with a stereo vision system." *Pattern Recognition, ICPR 2004*. 1 (482-485) (2004).

[13] Zheng Zhang, Chew Lim Tan, Liying Fan. "Restoration of curved document images through 3D shape modeling." *Computer Vision and Pattern Recognition, CVPR 2004*. 1 (10-15) (2004).

[14] Nikolaos Stamatopoulos, Basilios Gatos, Ioannis Pratikakis, Stavros J. Perantonis. "A two-step dewarping of camera document images." *Document Analysis Systems*. (209-216) (2008).

[15] Lili Song, Yadong Wu, Bo Sun. "A Robust and Fast Dewarping Method of Document Images." *E-Product E-Service and E-Entertainment, ICEEE 2010*. (1-4) (2010).

[16] Likforman-Sulem Laurence, Jérôme Darbo, Elisa H. Barney Smith "Enhancement of historical printed document images by combining Total Variation regularization and Non-local Means filtering." *Image and Vision Computing* 29.5 (351–363) (April 2011).

[17] Michael Patrick Cutter, Patrick Chiu. "Capture and dewarping of page spreads with a handheld compact 3D camera." *Document Analysis Systems*. (205–209) (2012).

[18] Kazim Pal, Melissa Terras, Tim Weyrich. "Interactive exploration and flattening of deformed historical documents." *Computer Graphics Forum*. 32 2pt3 (327-334) (2013).

[19] Chelhwon Kim, Patrick Chiu, Surendar Chandra. "Dewarping Book Page Spreads Captured with a Mobile Phone Camera." Camera-Based Document Analysis and Recognition. (101-112) (2014).

[20] Sergi Foix Salmerón, Guillem Alenyà Ribas, Carme Torras Genís. "Exploitation of time-of-flight (ToF) cameras." Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Tech. Rep. IRI-TR-10-07 (2010).

[21] Loren Arthur Schwarz, Artashes Mkhitaryan, Diana Mateus, Nassir Navab "Human skeleton tracking from depth data using geodesic distances and optical flow." *Image and Vision Computing*. 30.3 (217–226) (March 2012).

[22] David Jiménez, Daniel Pizarro, Manuel Mazo, Sira Palazuelos "Modeling and correction of multipath interference in time of flight cameras." *Image and Vision Computing*. 32.1 (1–13) (January 2014).

[23] Frederic Garcia, Bruno Mirbach , Bjorn Ottersten, Frederic Grandidier, Angel Cuesta "Pixel weighted average strategy for depth sensor data fusion", *Image Processing*. (2805-2808) (2010).

[24] Sebastian Schwarz , Marten Sjostrom, Roger Olsson. "A weighted optimization approach to time-of-flight sensor fusion." *Image Processing*. 23.1 (214-225) (2014).

[25] Luis Galarza, Zhenzhong Wang, and Malek Adjouadi. "Book Spread correction using a time of flight imaging sensor." IPCV 2014 (2014).

[26] https://support.bluetechnix.at/wiki/Argos%C2%AE3D_P100_Camera

[27] Stefan Fuchs, Gerd Hirzinger. "Extrinsic and depth calibration of ToF-cameras." *Computer Vision and Pattern Recognition, CVPR 2008*. (1-6) (2008).

[28] CholSu Kim, Huimin Yu, Gang Yang. "Depth super resolution using bilateral filter." *Image and Signal Processing, CISP2011*. 2 (1067-1071) (2011).

[29] Lu Wang, Ph.D. Dissertation Topic: "An Automated Book Reader Design as an Assistive Technology Tool for Persons with Blindness, *Department of Electrical and Computer Engineering, FIU,* (December 2007)

[30] http://www.abbyy-developers.eu/en:fre:new-v11

Design and Implementation of Smart Vehicular Camera for Real-Time Visual Metadata Extraction and Sharing

Sanghyun Son, Beomjun Kim, Yeonsu Jung and Yunju Baek
School of Computer Science and Engineering, Pusan National University, Busan, Republic of Korea
sonsang@eslab.re.kr, bjkim@eslab.re.kr, yeonsu@pusan.ac.kr, yunju@pusan.ac.kr

Abstract - *The number of vehicles has grown, accompanied by an increase in accidents. Thus, advanced driver assistance systems providing the status of a vehicle and the surrounding environment using various sensor data are being studied. We designed and implemented a smart vehicular camera device for real-time extraction and visual metadata sharing. Moreover, we propose a technique for extracting the metadata from an input image and sensor data using an image recognition algorithm. Moreover, we propose the S-ROI and D-ROI techniques, which set the region of interest in an image frame to improve the image processing, and a pattern check technique for increasing the recognition rate. We developed the main server for information sharing and a real-time road view service. We also evaluated the performance of the two ROI methods, and confirmed that the video processing speed of S-ROI and D-ROI are 3.0- and 4.8-times better than a full-sized frame analysis, respectively.*

Keywords: region of interest; metadata extraction; smart vehicular camera; image recognition; information sharing

1 Introduction

Along with urban development, the number of the vehicles on the roads has increased, with the greater complications accompanying such development [1]. The need for a technologically-based increase in vehicle safety to mitigate the risk of traffic accidents is increasing in proportion to the increasing number of vehicles [2]. An advanced driver assistance system provides a variety of information for solving traffic and safety problems [3]. Initially, driver assistance systems were studied in terms of driver convenience. Recently, this system provides safety services such as alerts of lane departure and directly vehicular control. In major advanced countries, a trend in enforcing the mandatory installation of such safety-related systems has begun.

IT companies such as Google and Apple, along with automobile manufacturers such as BMW, WV, and Mercedes Benz are researching autonomous driving and driver assistance techniques. Mobileye [4] ADAS provides more information using a computer vision algorithm and a proprietary hardware device than other types of sensor-based ADAS. In image-processing based ADAS, the front and rear images are recorded and analyzed to provide road condition information.

Although the existing driver assistance systems are different from utilizing of sensors and algorithms, the vehicle can recognize and utilize the information separately. Vehicles are able to collect various types of road information using image processing and an existing sensor technique; however, information collected by a vehicle is limited when compared to shared information. Therefore, it is necessary to allow vehicles to share information through a network.

For vehicle-to-vehicle wireless communication, research based on wireless sensor networks and ad-hoc networks [5] is underway. Moreover, IEEE 802.11p WAVE [6] is a related standard that defines the PHY and MAC layer for V2X communication. However it is difficult to apply an ad-hoc network because such a network requires a high penetration of networking devices and infrastructure. To solve this problem, a mobile cellular network can be used. Although the use of a mobile network is limited by fee it is possible to use such a network through the driver's smart phone.

When a mobile cellular network is used, a metadata extraction technique is required to minimize the data transmission. For visual data, visual metadata analyzing an input image are extracted. For sensor data, specific information from various sensors is extracted. Therefore, extracting metadata minimizes the data size.

An image analysis is conducted using a computer vision algorithm [7]. Such algorithms find specific patterns based on learned data within a frame. An algorithm used for an image analysis searches sequentially by applying various mask filters, and thus requires a great deal of computing power. For image processing, we applied a high-performance processor and reduced ROI technique to improve the processing speed.

For this paper, we designed and implemented a computer-vision based smart vehicular camera for real-time extraction and sharing of visual metadata. The smart vehicular camera extracts various types of visual metadata for analyzing an input image and shares the metadata through a network. Moreover, we propose a D-ROI technique for minimizing the computational load, and a pattern check technique for increasing the recognition rate. We developed a main server for information sharing and a real-time road view service. Finally, we evaluated the implementation of the smart vehicular camera device and confirmed its image processing speed and recognition rate.

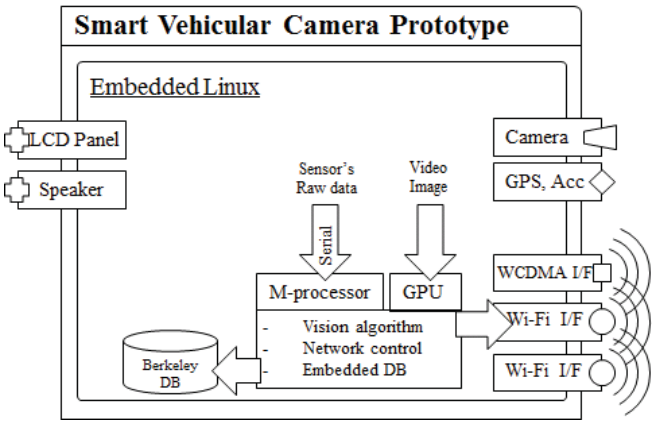


Fig. 1. Block diagram of smart vehicular camera device

2 Related Works

This section describes the advanced driver assistance systems associated with the proposed smart vehicular camera and image recognition technologies.

ADAS provides a variety of audio-visual information to the driver to prevent accidents that may occur during operation. The system alarms to the driver using sensors to recognize risks on the road, and the driver confirms the warning and response. The main functions of ADAS consist of forward collision avoidance, lane departure warning function, blind spot monitoring and rear monitoring.

ADAS with these features are likely to continue to grow due to the changes of the social structure, regulatory strengthening. Consumers are increasingly in demand, and since technological development and mass production are reduce the production cost. Therefore, ADAS market is expected to grow an annual average of 25% by 2017 [8].

The image recognition technologies for the vehicular environment is a part of computer vision technology. According to research of driver assistance system is underway, the image recognition technologies has attracted attention as complementary elements for vehicular sensors. Computer vision that is a branch of artificial intelligence aims to understand scene or feature on the image.

In order to perform these functions on the smart vehicular camera, many researchers use the computer vision library which is OpenCV (open computer vision). This library is possible that various recognitions such as face, pedestrian and motion tracking on a variety of platforms, and provides sample codes in various program languages. [9]-[12] The proposed smart vehicular camera uses this library to recognize pedestrian, license plate, lane and lane departure detection.

TABLE I
SPECIFICATION OF THE SMART VEHICULAR CAMERA DEVICE

Modules	Specification
Board	Odroid-XU (Hardkernel)
CPU	Samsung Exynos5410 Octa 1.6Ghz
GPU	PowerVR SGX544MP3 533MHz
Memory	LPDDR3 2GB
Wi-Fi	WN111v2 (Netgear)
WCDMA	DTW-400W (AnyData)
Camera	Odroid USB-CAM (720p)
GPS	GPS680(Acen Korea) / NMEA0183
LCD	LTN101AL03-8 (10.1 inch)

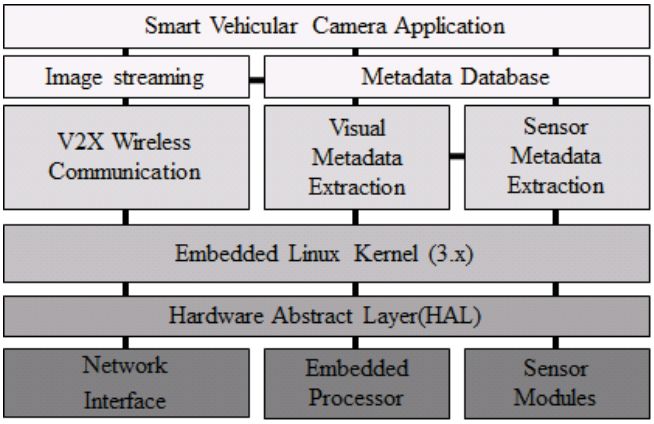


Fig. 2. Software architecture of smart vehicular camera

3 Smart vehicular camera

In this section, descriptions of the hardware prototype of the proposed smart vehicular camera device and the image processor, as well as the region-of-interest technique used to extract the visual metadata, are provided.

3.1 Smart camera device

A smart camera device must be capable of analyzing data from an input image and sensor data from onboard modules, and share the collected metadata over a network. Computer vision algorithms require high computational power, and thus high-performance embedded processors are required for the image processing device. We selected the high-performance Samsung Exynos 5410 core processor, which contains an ARM-based 1.6 GHz octo-core processor and a PowerVR SGX544MP3 GPU. The proposed prototype device includes an Exynos core, 2 GB of DDR3L RAM, an IEEE 802.11n Wi-Fi module, an HSDPA + WCDMA module, and various sensors. Table 1 shows the specifications of the proposed device.

The proposed device uses an HD class camera to record the road images, 3GPP Release 5 for the WCDMA module, and a Wi-Fi module for wireless networking capability. In addition, the device recognizes the current state of the vehicle using a GPS device, an acceleration sensor, and an OBD scanner. This information is forwarded to the driver over an LCD screen and speaker, which provide warning messages and alarms. Figure 1 illustrates the block diagram of the proposed smart vehicular camera device.

For operation of the smart camera device, the required software components are as follows: an embedded operating system for control of the peripheral devices, modules and libraries of the metadata extraction and wireless networking, and an application constituting all functions and the forwarding of information to the user. Figure 2 shows the software architecture of the proposed device.

Fig. 3. Prototype device of proposed smart vehicular camera

We developed a prototype device based on the design of the smart vehicular camera. Figure 3 shows the implementation of a smart vehicular camera device. The device, based on a Hardkernel Odroid-XU [12] development board, includes several modules such as a camera, GPS device, Wi-Fi, a WCDMA module, an OBD scanner, an LCD panel, and a speaker. The OBD scanner was developed using a CAN transceiver and STM32F103 microcontroller.

3.2 Visual Metadata Extraction Technique

For image recognition, we developed visual metadata extraction module by applied OpenCV library. In the various recognition functions, we implemented image recognitions such as pedestrian, license plate, lane, lane departure. To recognize pedestrian, we applied a HOG algorithm in the library and used SVM classifier with learning data for pedestrian recognition.

For vehicular license plate recognition, we applied a rectangle detection algorithm and an optical character recognition (OCR) algorithm to our recognizing application. After detect rectangles in image frame, analyze candidate rectangles using OCR to recognize a vehicular number. For lane recognition, the application finds straight lines and selects lanes that straight lines toward the vanishing point in the region. The application uses Hough transform to find straight lines and calculates the vanishing point using random sample consensus (RANSAC) algorithm. The lane departure function is verifying the gradient of lane continuously, and the function determines lane departure condition when the gradient changes more than a predetermined value.

Fig. 4. Size of static ROI to recognize objects on the road

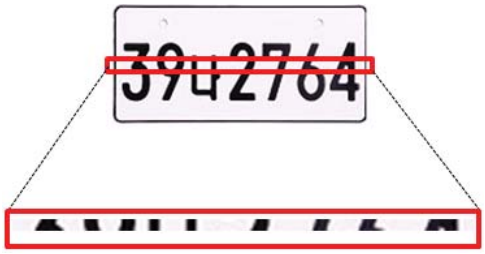


Fig. 5. Example of color pattern check technique to improve license plate recognition-rate

3.3 ROI Technique

The image recognition technique is conducted using many simpler comparison calculations compared to other applications, and the computing time required for the feature extraction is increased when the size of the target region of the image frame is increased. Thus, it is recommended to reduce this region by setting the ROI according to the recognized target. Therefore, the application sets the ROI according to the camera attached to the vehicle. A vehicular camera that consistently monitors the same area is able to set the ROI intuitively.

Owing to the use of a narrow surveillance area, the computing time is reduced and recognition errors occurring outside of the search region are avoided. The camera attached to the front of the vehicle monitors the center region, and thus the application is able to increase the image processing time by excluding the top and bottom regions, such as the sky and hood of the vehicle. Likewise, the side regions such as the other side of the road and the pedestrian area are excluded from the ROI.

In this paper, for pedestrian and license plate recognition, we set the ROI to 33% of the frame. For the lane and lane departure recognition, we set the ROI to a narrower region than that of the S-ROI used for pedestrian recognition. For the lane recognition, we set the ROI to 12% of the frame, and for the lane departure recognition, the ROI is set to 4%. The ROIs used for recognition are defined as static ROIs (S-ROIs). An S-ROI is always the same size regardless of the recognition result, and the exact position and size of the ROI can change according to the camera conditions.

Although the S-ROI technique is simple to set up and easy to apply, it is inefficient for a road environment, such as when monitoring the same region for a predetermined period time. In particular, when pedestrian and license plate recognition algorithms find a target object, the probability that the target will be found near the same region in the next frame is high. Considering these features, the application searches the S-ROI for recognition by default, and minimizes the ROI fitting the recognized target when the algorithms find their target. We defined and implemented this technique as a dynamic ROI.

If a target such as a pedestrian or license plate is recognized in a previous frame, the application temporarily reduces the ROI as follows. For the case of a pedestrian, most pedestrians appear in front of the vehicle and are crossing the road at a constant speed. Thus, the application sets a new D-

(a) (b) (c)

Fig. 6. Example of the proposed viewer screens of road condition and real-time street view based on the google map

ROI by considering the pedestrian speed. Considering recognition-able human size, the size of the D-ROI for a pedestrian is 7% of the frame. For recognizing a license plate of the front vehicle, because the vehicle can move in all directions, the application sets a new D-ROI by considering the lane-to-lane width. The size of the D-ROI for a license plate is 11% of the frame.

The D-ROI condition is changed to normal when the target disappears for longer than a D-ROI timeout. The D-ROI technique reduces the image processing time; however, this technique has a problem in that it cannot recognize new targets during a narrow region search. Thus, the D-ROI technique has a timeout value, and this is designed searching S-ROI after a predetermined time. The timeout value varies according to the application or recognized target, and is set to achieve the optimal performance in terms of the processing time and recognition rate.

3.4 Patten Recognition Techniques

For the proposed license plate recognition technique, rectangles are recognized within the frame to collect candidate license plates, after which the technique determines the existence of a license plate by checking the aspect ratio of the target rectangle. However, the error rate of this technique, i.e., determining that a rectangle is not a license plate, is high.

Table 2 shows the results of license plate recognition for various input image sizes. The results confirm that the recognition rate is low and the error rate is high. To solve this problem, we proposed a pattern check technique, which checks the center line of the candidate rectangle and counts the number of color pattern changes. The background color of a Korean license plate is white, and the characters are black, and thus the technique determines the existence of a license plate when the number of white-black pattern changes is over the threshold value. Figure 5 shows an example of pattern check technique for increasing recognition-rate.

3.5 Metadata Sharing Technique

To utilize the metadata collected from a smart vehicular camera device, information sharing is required. Research into wireless ad-hoc-based vehicle-to-vehicle communication is underway; however, it is difficult to apply to the ad-hoc network because such a network requires a high penetration of the networking device and infrastructure. Therefore, we used a mobile cellular network to communicate this information.

The number of methods using a mobile network is two. In the first method, the smart vehicular camera device uses a WCDMA module for direct communication. In the second method, the device connects to the smart phone of the driver and thus connects to its mobile network. Although the use of a WCDMA module is convenient, an additional network charge must be paid. When using a smart phone, the two devices are connected using a wireless local area network technique such as Wi-Fi or Bluetooth. Although the devices can connect through tethering, such a connection process is inconvenient.

The proposed smart vehicular camera device is able to use these two methods, including a WCDMA module and Wi-Fi module, and set the communication directly by default. If metadata are extracted, the information is stored in an internal database and the updated information is transmitted to the main server for information sharing. In addition, the server forwards this information to the nearby vehicles based on their GPS position information.

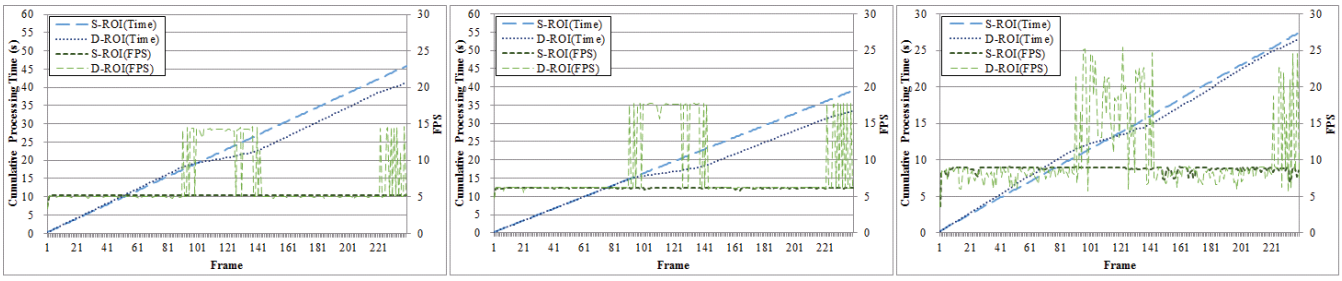
The main server used for information sharing collects the data from the vehicles and provides road condition information. This server stores the vehicular data such as the time, position, speed, road conditions, snapshot images, and video sequences. Moreover, the server transmits the information to the requesting vehicles and provides road images and condition information to the user by utilizing a Web-based viewer. Figure 6 shows the viewer screens such as the road condition and real-time street view. In figure 6 (a) and (b), the arrows on the Google map show the road conditions in accordance with the direction based on color, and the road conditions are confirmed through the uploaded

TABLE II
RECOGNITION-RATE AND ERROR-RATE OF OBJECT RECOGNITION

	VGA	WXGA	FHD
Recognition-rate	45.5%	49.4%	49.0%
Error-rate	30.6%	68.3%	75.5%

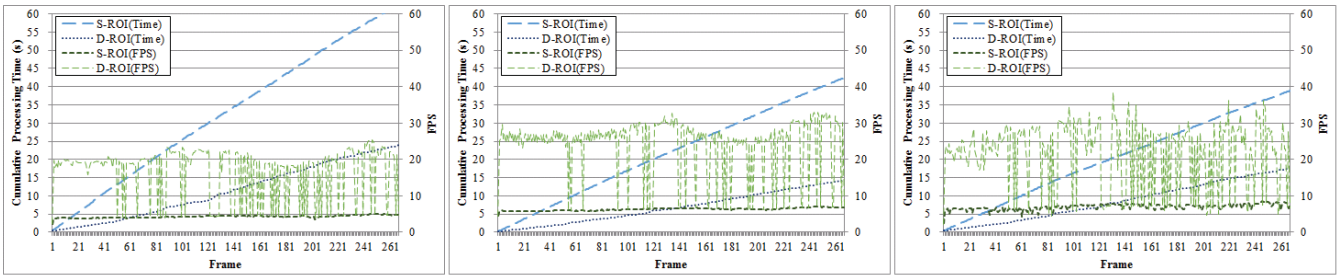
TABLE III
SPECIFICATION OF THE METADATA MEASURING PERIOD

Device	Technique	Kind of metadata	Period
Camera	Record Image	Image, video	-
	Computer Vision (with ROI)	Pedestrian	0.20s
		Lisence plate	0.25s
		Lane, lane departure	0.05s
GPS	Satellite positioning	Position, direction, time	1.0s
Accel.	Sensor measuring	Impact, acceleration	0.1s
OBD	CAN comm.	RPM, speed, pedal position	0.5s



(a) Arndale (b) Odroid-X2 (c) Proposed device

Fig. 7. Pedestrian recognition performance according to each ROI method and each each test board



(a) Arndale (b) Odroid-X2 (c) Proposed device

Fig. 8. License plate recognition performance according to each ROI method and each each test board

road images. Figure 6 (c) shows the speed change of the target vehicle in the viewer.

4 Performance Evaluation

In this section, an evaluation based on the hardware performance, image processing speed, and recognition-rate is conducted.

The smart camera device uses various modules such as a camera, GPS device, an accelerator sensor, and an OBD scanner to extract the metadata. A still image is a type of visual metadata and is analyzed as data using computer vision. The still image is stored and forwarded when a request from the main server is received. If image processing has been conducted, the measurement time is changed based on the computational load.

Image processing of a WVGA (800 x 480) input image using the D-ROI technique takes 0.33 s for the proposed device. When sensor metadata are extracted, the extraction period is changed according to each sensor module. The GPS module is able to change the measurement period based on the setup value, and we set the GPS measurement period to 1 s. The acceleration sensor used for shock recognition is able to measure approximately 15,343 raw datum per second. We set the measurement period to 0.1 s using an average of ten values, and determine a sudden change of acceleration using 150 average values. The OBD scanner used for collecting the sensor data of a vehicle is able to take approximately 22 measurements per second. We set the measurement period to 0.5 s using the average of ten values. Table 3 shows the specifications of the metadata measurement period.

We developed the proposed pattern-check technique to increase the license plate recognition rate. For a VGA sized image, two cases, i.e., rectangle detection and a check of the aspect ratio of the candidate rectangles, and a color-pattern check after determining the aspect ratio, are compared. The results of the image recognition show that the recognition rate of the proposed technique is increased by approximately 39% and that the error-rate is decreased by approximately 15%. We applied this pattern-check technique and conducted a performance evaluation for the image processing speed. Table 4 shows the result of license plate recognition using the pattern check technique in the image of VGA size.

We evaluated the proposed S-ROI and D-ROI techniques, and confirmed the processing speed and recognition rate based on an evaluation of each algorithm. For the evaluation of the ROI techniques, we experimented with pedestrian and license plate recognition algorithms. We confirmed the result of the S-ROI using only lane and lane departure recognition algorithms. We evaluated the processing speed of each algorithm, and measured the average processing time of the algorithms overall. For an evaluation of the hardware-based image processing

TABLE IV
RESULT OF PLATE RECOGNITION USING THE PATTEN CHECK TECHNIQUE

	Non-check	Pattern check
Recognition-rate	45.5%	84.7%
Error-rate	30.6%	15.32

TABLE V
RECOGNITION-RATE AND ERROR-RATE OF RECOGNIZING OBJECTS

	Pedestrian (fps)		License plate (fps)	
Arndale	S-ROI	5.22fps	S-ROI	4.30fps
	D-ROI	6.95fps	D-ROI	16.78fps
Odroid-X2	S-ROI	6.14fps	S-ROI	6.31fps
	D-ROI	8.60fps	D-ROI	24.37fps
Proposed device	S-ROI	8.76fps	S-ROI	6.99fps
	D-ROI	10.48fps	D-ROI	21.81fps
R-rate	S-ROI	70.7%	S-ROI	63.3%
	D-ROI	67.4%	D-ROI	45.7%

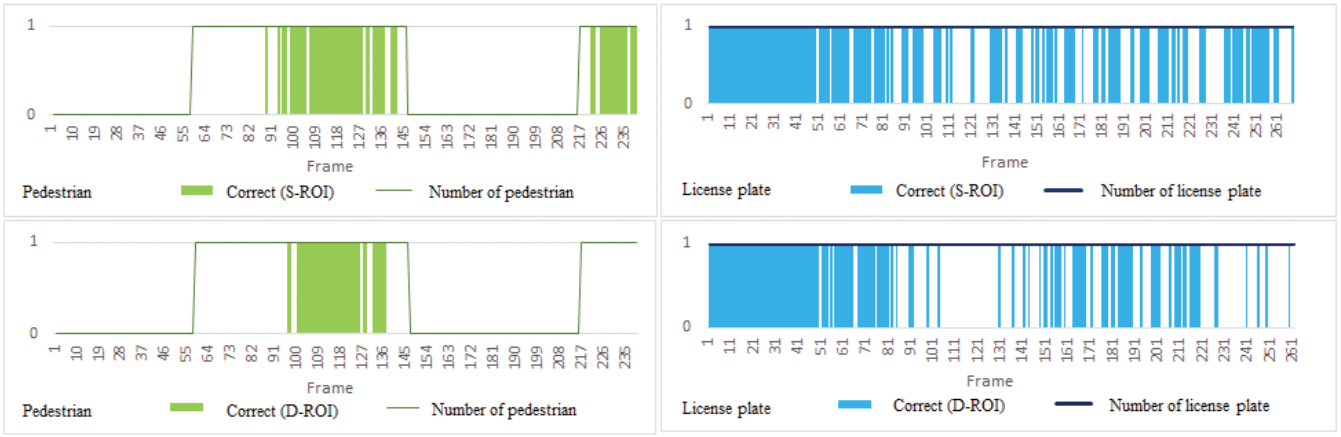


Fig. 9. Recognition-rate of pedestrian and license plate according to each ROI method and each recognition algorithm

performance of the proposed device, we conducted experiments on various development boards such as an Odroid-X2 (Exynos 4412) [13] and Arndale Octa (Exynos 5420) [14].

We evaluated the S-ROI and D-ROI techniques for improving the image processing speed for pedestrian and license plate recognition on each device. While applying the ROI techniques, confirmed the fps and recognition rate. The recognition rate was the same for all devices because they all use the same algorithm. Figure 7 shows a graph of the results for fps when pedestrian recognition was applied. When applying the S-ROI technique, the recognition algorithms search the same reduced region, and thus the fps value is always similar. When applying the D-ROI technique, the recognition algorithms search the same sized S-ROI. If a target object is found, the size of the D-ROI is reduced, and thus the image processing speed is increased. Therefore, the two ROI techniques show the same FPS values when there is no target in an image. If a target object is found, the fps value of the D-ROI is increased.

The results of the license plate recognition in Figure 8 show that a plate is continuously present within the image. Thus, we confirmed that the fps value of the D-ROI is always high. A comparison of the results for each device shows that the proposed device has the highest fps value. Figure 9 shows the results of the recognition rate, which is determined as follows (1). In (1), *TP* denotes the true positive and *FN* represents the false negative. *TN* denotes the true negative and *FP* represents the false positive.

$$Recognition\ rate = TP+FN / TP+FN+TN+FP \quad (1)$$

We confirmed the results of the image recognition processing for each frame. The parameters for each recognition algorithm were set to minimize false positives. For the pedestrian recognition, we confirmed that the recognition rate using S-ROI is 70.7%, and using the D-ROI is 67.4%. For the license plate recognition, we confirmed that the recognition-rate using the S-ROI is 63.3%, and using the D-ROI is 45.7%. As a result, the recognition-rate when using the S-ROI is higher than that when using the D-ROI for the following reason. The D-ROI technique reduces the ROI size when the algorithm recognizes the target in an image. If the recognition is incorrect, the next frame recognition will fail with a high probability. The experimental results are given in Table 5.

We determined the average image processing time required to run all recognition algorithms for each device. In the experiments, the recognition processing times, including for pedestrians, license plates, lanes, and lane departures, were measured. We used three test devices and applied a full-sized frame, S-ROI, and D-ROI. Figure 10 shows the experiment results. The three devices received a WVGA image and applied four recognition algorithms. In the full-sized search, each device showed a rate of approximately 1 FPS. In contrast, the proposed device showed a processing time of 0.19 s. In the experiments, we confirmed an improved image processing speed of 3.0- and 4.8-times the original when using the S-ROI and D-ROI, respectively.

5 Conclusions

For this paper, we designed and implemented a computer-vision based smart vehicular camera for real-time extraction and visual metadata sharing. The proposed device analyzes a recorded image frame from the camera module and extracts the visual metadata, and then extracts the sensor metadata over the sensor modules. We developed the device and main server such that the device transmits metadata to the main server using a mobile cellular network to share information with other vehicles on the road. In addition, we proposed a dynamic ROI technique to minimize the computational load for image processing, and evaluated the performance.

We designed the smart vehicular camera device to include a high-performance embedded processor for real-time image processing, and evaluated the proposed device experimentally. The experiment results confirm that the image processing speed is increased 3.0-fold when the recognition algorithms apply the S-ROI, and 4.8-fold when the D-ROI is applied. However, when the algorithms applied the D-ROI, we confirmed that the recognition rate was

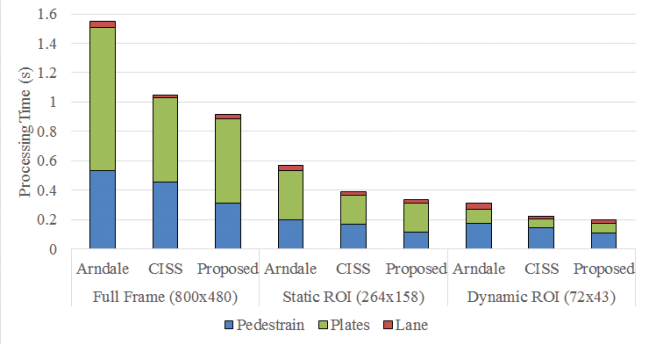


Fig. 10. Processing time of all recognizing operation

reduced. Therefore, the application should use the appropriate ROI technique based on the processing-time and recognition-rate requirements.

The proposed device is executable for an image processing speed of 5.3 fps. However, applications require a higher real-time processing speed for full input frames. Thus, we must improve the image processing speed by using advanced processors and libraries. For an object recognition technique, we have to research how to conduct an analysis under a variety of situations. Finally, we have to solve the problem of a lack in the number of libraries for image recognition processing when using the GPU core in an embedded processor. We have a plan to conduct research to improve the above problems.

6 Acknowledgement

This work was supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project (CISS-2011-0031863). Corresponding author: Yunju Baek (yunju @pusan.ac.kr)

7 References

[1] S.G. Klauer, T.A. Dingus, V.L. Neale, J.D. Sudweks, and D.J Ramsey, "The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 10-Car Naturalistic Driving Study Data," NHTSA Crash Avoidance Research Technical Publications, Report No. DOT HS 810 594, 2006.

[2] Lawrence D. Burns, "A vision of our transport future," Nature Vol. 497, pp.181-182, 2013

[3] Eric Raphael, Raymond Kiefer, Pini Reisman and Gaby Hayon, "Development of a Camera-Based Forward Collision Alert System," SAE World Congress & Exhibition, 2011

[4] http://www.mobileye.com

[5] Yuh-Shyan Chen, Yun-Wei Lin and Sing-Ling Lee, "A Mobicast Routing Protocol in Vehicular Ad-Hoc Networks," in Proceedings of the ACM/Springer Mobile Networks and Applications, Vol. 15, No. 1, pp.20-35, 2010.

[6] IEEE Std 802.11p, IEEE Standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-sepcific requirements, Part 11, Amendent 6: Wireless Access in Vehicular Environments, 2010.

[7] David Geronimo, Antonio M. Lopez, Angel D. Sappa, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," Pattern Analysis and Machine Intelligence, IEEE, Vol. 32. pp. 1239-1258, 2010

[8] Research and Markets: Global Advanced Driver Assistance Systems Market 2012-2016 with General Motor Co., Robert Bosch GmbH, Takata Corp., and Toyota Motor Corp Dominating, Technavio Research report, 2013

[9] Paul Viola and Michael J. Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition, CVPR, Proceedings of the IEEE Computer Society Conference, Vol, 1, 2001

[10] Navneet Dalai and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 886-893, 2005.

[11] David Gero´nimo, Antonio M. Lo´pez, Angel D. Sappa, and Thorsten Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32, pp.1239-1258, 2009

[12] http://www.opencv.org

[13] http://www.hardkernel.com/main/products/prdt_info.php?g_code=G137510300620

[14] http://www.hardkernel.com/main/products/prdt_info.php?g_code=G135270682824

[15] http://www.arndaleboard.org/wiki/index.php/Main_Page

Qualitative Image-Based Localization in a Large Building

Christopher Card, William Hoff
Department of Electrical Engineering and Computer Science
Colorado School of Mines
Golden, USA
ccard@mymail.mines.edu, whoff@mines.edu

***Abstract*—Interest in indoor localization is growing because it is an important component of many applications. Image-based localization, using naturally-occurring features in the environment, is an attractive solution to this problem. A challenge is to be able perform this on a mobile device with limited computing power. Another challenge is that buildings can have locations with a similar appearance, which can confuse an image-based recognition system. Since many applications do not need exact location, we focus on qualitative localization, which is the problem of the problem of determining approximate location by matching a query image to a database of images. We propose a novel approach that uses an efficient hashing scheme to quickly identify candidate locations, then applies a strong geometric constraint to reject matches that have similar appearance. On experiments in a large campus building, we show that this approach can localize a query image with high accuracy and has potential to run in real time on a mobile device.**

***Keywords*-computer vision; indoor localization; image retrieval; vision based navigation.**

I. Introduction

The goal of indoor localization is to determine the location of a mobile device in an indoor environment. Interest in this problem is growing because determining location inside a building is an important component of many applications, such as augmented reality, customer navigation, and behavior and movement tracking.

The problem is very different from outdoor localization because the device no longer has access to a reliable GPS signal. A variety of alternative methods can be used in place of GPS. The most popular approaches require some kind of infrastructure to be present in the building. For example, Tesoriero *et al* [1] places Radio Frequency IDentification (RFID) markers throughout the environment. Another approach is to use Wi-Fi fingerprinting, as done by Hile *et al* [2].

An alternative to RFID and Wi-Fi is to use image-based localization. This is attractive because the vast majority of people already have mobile devices (*e.g.* smart phones) with cameras and the approach is applicable to buildings without RFID or Wi-Fi. A recent survey of optical indoor positioning systems is given in [3].

Image-based localization can use naturally-occurring features in the environment. This has the advantage that no infrastructure is required. One challenge is that doing localization based on naturally-occurring features can be computationally intensive, but we want to be able to run our application on a mobile device with limited computing power. Another challenge is that in a large building, there can be many locations that have a similar appearance, thus potentially confusing an image-based recognition system.

In this paper we describe an approach that can perform localization within a large building using no infrastructure or any special mapping steps. The above challenges are addressed in the following ways: first, an efficient hashing scheme is used to quickly identify candidate locations that match a query image. Next, a strong model based on geometric constraints is employed to identify the correct match, while rejecting matches that have a similar appearance. Finally, a local map in the vicinity of the user is constructed to limit the search for candidate matches. Although the system was not implemented on a mobile device, an initial analysis shows that it has the potential to run in real time on a reasonably capable mobile device.

The remainder of this paper is organized as follows. Section 2 describes previous related work and motivates our approach. Section 3 describes the approach in detail. Section 4 provides an evaluation of the system on a large campus building. Section 5 is the conclusion.

II. Related Work

In this section we focus on methods that perform image-based localization by detecting naturally-occurring features in the environment. The most common way of performing localization using natural features is to match them to a 3D model, or map of the building. The 3D map can be automatically constructed from images taken by standard 2D cameras (such as are present on smart phones) using methods known as "structure from motion" (Sfm). Several groups have performed image-based localization by matching features from a query image to 3D points estimated by Sfm [4],[5]. However, Sfm is computationally expensive and requires accurate camera calibration. In our application we may be using images from smart phone cameras from multiple users with no a priori calibration.

Rather than trying to create a metrically accurate map, a qualitative map can be used. For many tasks, the exact

Figure 1. Examples of some of the database images depicting scenes that have similar features but are captured at different locations.

(a) (b)

Figure 2. (a) and (b) show two different doorways that are different but have many similar features.

pose of the camera does not need to be known; approximate locations are sufficient (*e.g.* within 10 or 20 feet). We call this "qualitative localization" (following the terminology of Kosecka [6]). The localization problem then reduces to the problem of matching a query image to an existing image in the database; if such an image can be found then the user is near the location where the database image was taken. The advantage of this "place recognition" approach is that an expensive and difficult map building process is unnecessary and uncalibrated camera images captured by users can be used.

A standard method for place recognition is to use the Bag of Words (BoW) approach. BoW quantizes feature vectors into visual words thus creating a visual vocabulary [7]. To match a query image to an image in a database, the algorithm simply finds the distribution (histogram) of visual words found in the query image and compares this distribution to those found in the database images. Although this could be used for qualitative localization, BoW can fail when the histograms of visual words are too similar.

In a large building, there can be many locations that have a similar appearance. Walls and floors often have little or no texture and doors look very similar. For example, Figure 1 shows a set of images from a large building on the Colorado School of Mines (CSM) campus. There are many features (such as the corners between doors and the floor) which are present in all the images. Thus, the histograms of words are not very distinctive. This would result in incorrect matches to the database.

The *placement* of features in the image is potentially more distinctive than the histogram of features. For example, the images in Figure 2 (a) and (b) are very similar in terms of the types of features that are present. However, the poster to the left of the door is in a different place in each of these two images. This suggests that geometric constraints can be used to uniquely match images. A geometric constraint which is very general and powerful is the fundamental matrix [8]. The locations of all feature points between two images are related via the same fundamental matrix. Therefore, if a fundamental matrix that relates a sufficiently large number of features between two images can be found, then these images were likely taken of the same scene. In our approach, we fit a fundamental matrix to verify candidate image matches. The approach is similar to that of [9], who also uses the fundamental matrix to verify candidate image matches. Using the fundamental matrix as a model for matching is a much stronger constraint than simply comparing histograms of features and should result in much more reliable matches.

III. Localization Algorithm

Our approach is logically divided into three steps which are discussed in the subsections below: (1) feature detection, (2) feature matching, and (3) verification. This section is concluded with a description of the "local map" method.

A. Feature Detection

ORB [10] was chosen as the feature detection algorithm since it provides robustness to image deformation that is close to SIFT and SURF while providing a computational speedup of an order of magnitude [10]. This makes it ideal for localization in real time on a mobile device. In our experiments, ORB descriptors were computed for a query image in 0.05 seconds. The ORB descriptors for the database images were precomputed.

B. Feature Matching

Given a set of ORB descriptors extracted from a query image, these descriptors then need to be matched to the descriptors from the database images. The simplest technique is Brute Force (BF) matching, which exhaustively compares the query descriptor against each database descriptor to find the closest match in feature space. Although the most accurate method, this has the drawback that as the database increases in size the computational time becomes prohibitively expensive.

To avoid this problem, we store the database descriptors in a hash table. The same hash is applied to a query descriptor; the database descriptors at that location in the hash table

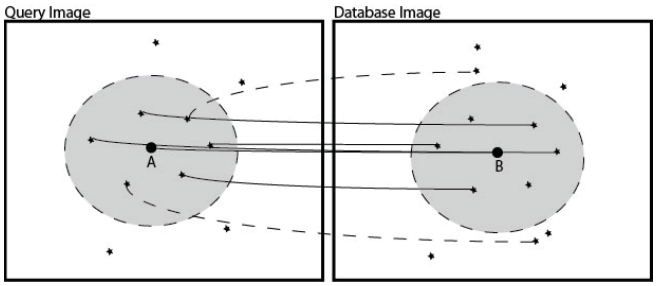


Figure 3. An illustration of spatial consistency voting. This verifies that the kNN of points (A,B) are spatially consistent. Each matched pair of points that is the nearest neighbor to both A and B casts a vote for the match between A and B.

are retrieved. This is a very efficient and fast operation. We use a method called Locality Sensitive Hashing (LSH) [11] which preserves the locality of key points in feature space when generating the hash of the image descriptor. In other words, the difference between hash values is a good approximation of the distance between the points in feature space. This allows finding nearby descriptors in feature space, not just the descriptors at the hash location. This is important because image deformation and noise can cause the descriptors to change. Our algorithm finds the k nearest neighbors for the query point, where $k = 15$.

LSH is extremely fast when matching features against a large database (1,073,903 feature points). In our experiments LSH was able to match against a large database in about 1.47 seconds and it was able to match against a local map (containing 33,000 feature points) in about 0.095 seconds.

The potential matches for each query point, q_i, are then filtered using two steps, as described below:

1) **Ratio test**. The first step of the filter process is to determine if there are multiple feature points from one database image that are nearest neighbors to q_i. If this is the case, the closest feature point from the database image has to be 80% closer to q_i than the second closest feature point; otherwise the match is discarded.
2) **Spatial consistency test**. The next step checks whether each matched pair of points is spatially consistent. The approach of Sivic *et al* [12] is used for this step (see Figure 3). The idea is that neighbors of the query point should have matches that are neighbors of the database point. Here, "neighbors" means that the points are neighbors in image space, not feature space. If the number of spatially consistent neighbors is below a threshold (a threshold of 6 points was used in this work), then the potential match is discarded.

After the two filtering steps are completed, the two database images with the highest number of matches to the query image are selected. These are the candidate matches to the query image. If there is a tie for second place, all images that are tied for second place are kept.

C. Verification

The verification step of the algorithm tests each candidate database image to see if the matching points fit a geometric constraint with the query image. The model used for the geometric relationship is the fundamental matrix [8]. The fundamental matrix models the epipolar geometry between two camera views of the same scene. RANSAC [13] is used to eliminate outliers. A fundamental matrix is found between the query image and every candidate database image. Then the image with the most inliers is found. If the number of inliers exceeds a threshold (described in Section 4), then that database image is determined to be the correct match. If not, then all the candidate images are passed to a secondary processing step.

The secondary processing step rematches all image features in the query image to the candidate set of images, except that it now uses BF matching instead of LSH. The image with the most inliers to a fundamental matrix is found and, if the number exceeds the threshold, it is determined to be the correct image; otherwise, the query image is considered to have no acceptable match to the database.

D. Local Map

If the size of the database can be reduced, this can potentially speed up computation as well as improve the accuracy of matching. To do this, we propose using a "local map" in the vicinity of the user which contains only the database images near the current location of the user. The size of the local map depends on how fast a user can reasonably walk in a given amount of time. As long as the user is within the boundaries of the local map, localization queries can be done by matching to the local map.

Our concept for this is as follows: When a user first runs our system to do localization, the image is sent to a server, which matches the query image to the entire image database. Once the user's approximate location is found, the system sends a local map to the user's mobile device. The user's mobile device then uses the received local map to perform localization. This greatly speeds up processing and improves accuracy which allows the mobile device to perform localization in near real time. When the user approaches the edge of the local map, the server downloads a new local map to the client. Although we did not implement this concept, we did evaluate the potential benefits of using a local map in terms of run time and accuracy, as described in the next section.

IV. Experiments

This section describes the database used and details the methodology applied to test the algorithm. We implemented our algorithm using C++ and the open source software OpenCV. The algorithm was tested on a laptop running Windows 7 with a 2.6GHz processor and 4GB of RAM.

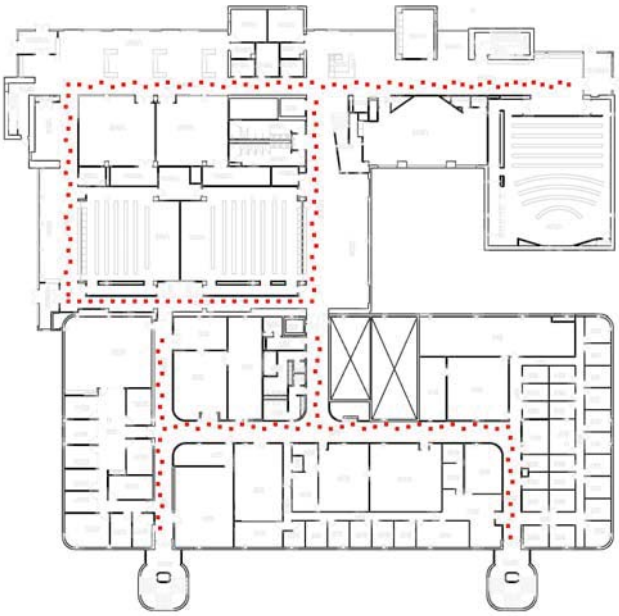


Figure 4. Red dots indicate where images were captured on the 2^{nd} floor of Brown Hall.

A. Database

The database was captured using a Cannon Rebel t2i Single Lens Reflex (SLR) camera with 8 megapixels per image. The database was captured in Brown Hall at CSM which is a large (100,000 square foot) building containing offices, classrooms and laboratories. The database consists of the 1^{st}, 2^{nd}, and 3^{rd} floors of Brown Hall, as these floors contain a representative sample of indoor environments which contain sparse texture and similar structural features. The images were taken at intervals of approximately 5 feet (see Figure 4). At each position multiple images were taken, facing both directions in the hall and additional directions to capture the appearance of nearby characteristic features (*e.g.* doors, side halls). The location where each image was taken was physically measured and recorded. The operation of the system is not dependent upon knowing these locations. These measurements were taken solely to test the system's accuracy.

The database is comprised of 1,382 images with a total of 1,073,903 feature points. Images in the database overlap, meaning that nearby images typically view a portion of the same scene (see Figure 1 for example images).

B. Tests

The following subsections describe the tests used to evaluate the algorithm using the collected database. A match is deemed to be correct if the location of the database image is less than 21 feet from the query image. In our tests, the correct match to a query image was in the database about 92% of the time.

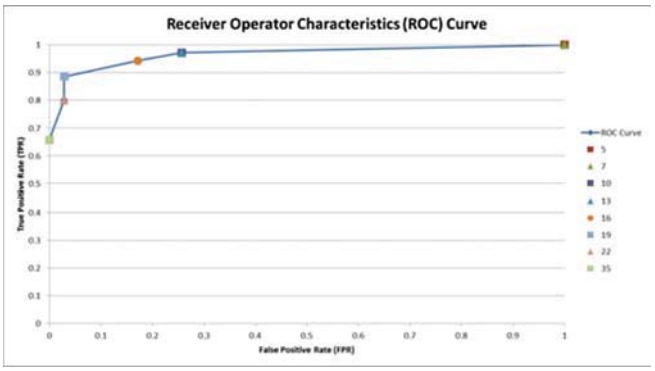


Figure 5. The ROC curve for a test set of 70 images matched against the remainder of the database. The same test set was used for all 8 runs.

1) Parameter Evaluation: One of the most important parameters in the algorithm is the threshold for the number of inliers to a fundamental matrix, which determines if a query image is successfully matched.

To avoid incorrect matches, it is desirable to use a higher threshold for the required number of inliers. This reduces the probability of a false match. However, this also reduces the probability of a true match. Conversely, lowering this threshold makes it more likely that a query image will be successfully matched to the correct database image. However, if a query image actually has no correct match in the database, lowering the threshold also increases the probability that a false match will occur.

To evaluate the effect of changing (*i.e.* tuning) this parameter on the probability of getting a false match, the following study was done. We randomly chose 35 images and removed them from the database and ensured that each of the 35 images had a correct match in the database. Thirty-five other images were captured (using the same camera) from parts of the building that were not in the database. These images have no correct match in the database.

A Receiver Operating Characteristics (ROC) curve was generated. ROC curves are based on a 2x2 confusion matrix, which records the count of the four possible outcomes of running the localization algorithm at each setting of the algorithm parameter. The four possible outcomes are:

- True Positive (TP). The correct match to the query image was in the database and the system found the correct match.
- True Negative (TN). The correct match to the query image was not in the database and the system correctly decided that there was no match.
- False Positive (FP). The correct match to the query image was not in the database, but the system matched it to an image that was not correct.
- False Negative (FN). The correct match to the query image was in the database, but the system was unable to find a match.

Table I
THE RESULTS FOR THE SUBSAMPLE TESTS. EACH PART OF THE TABLE CONTAINS THE SUM FROM 20 SUBSAMPLE TESTS.

Outcome	Number
Query has a match in database and algorithm found a correct match	538
Query has a match in database and algorithm found an incorrect match	14
Query has a match in database and algorithm declared "no match"	27
Query has no match in database and algorithm found an incorrect match	4
Query has no match in database and algorithm declared "no match"	17
Total number of queries	600

The True Positive Rate (TPR) is defined as the ratio of true positives (TP) to the total number of positives (TP+FN). The False Positive Rate (FPR) is defined as the ratio of false positives to the number of total negatives (FP+TN) [14].

The ROC curve is formed by plotting TPR against FPR. The resulting ROC curve is shown in Figure 5. As can be seen, the TPR is fairly high for most parameter settings. For example, using a threshold of 16, the TPR is about 94%, meaning that if the correct match is in the database the system will find it 94% of the time. The FPR for this case is about 17%, meaning that in those cases when the correct match is not in the database, the system finds an incorrect match instead of outputting a "no match" decision. Although this FPR seems high, the number of cases where there is no correct match in the database is small, so this outcome is relatively rare.

2) Subsample Test: To assess the overall accuracy of the algorithm over multiple runs, a subsample test was performed. Twenty test sets were created, where each test set consisted of 30 randomly chosen images from the full database, with no restriction on proximity. For each test set, the 30 images were removed from the database and then were used to query the database. In this experiment a threshold of 16 inliers was used as the decision threshold.

Overall, the algorithm performed well. Combining the results from all 20 subsample tests, the algorithm achieved an accuracy of 92.5% (see Table I). Here accuracy is defined as the fraction of all outcomes that were correct. Specifically, it is the number of outcomes in rows 1 and 5 in the table divided by the total number of trials. These results show that the algorithm can localize a query image with a high degree of confidence. Two examples of TPs are shown in Figure 6. Two examples of the query image having a match in the database but the algorithm found an incorrect match (FP) are shown in Figure 7. The FPs were caused by a set of highly clustered points. An FN is shown in Figure 8 where an insufficient number of inliers were found.

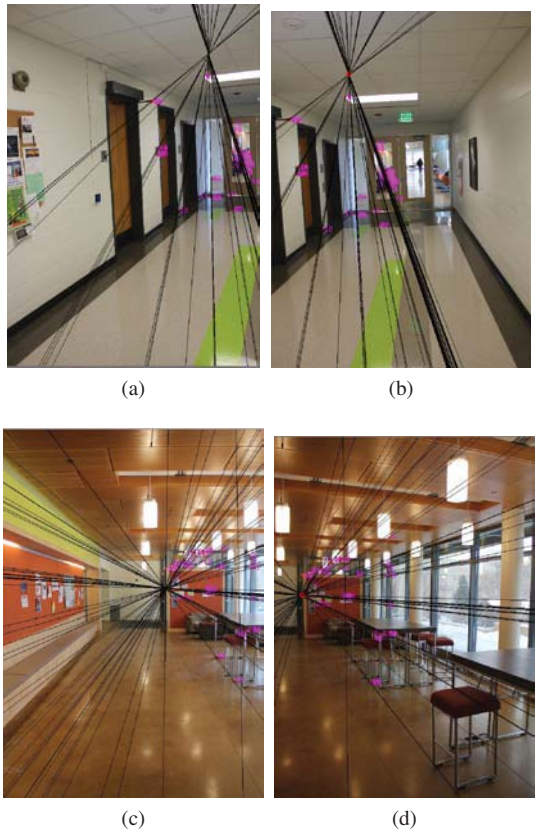
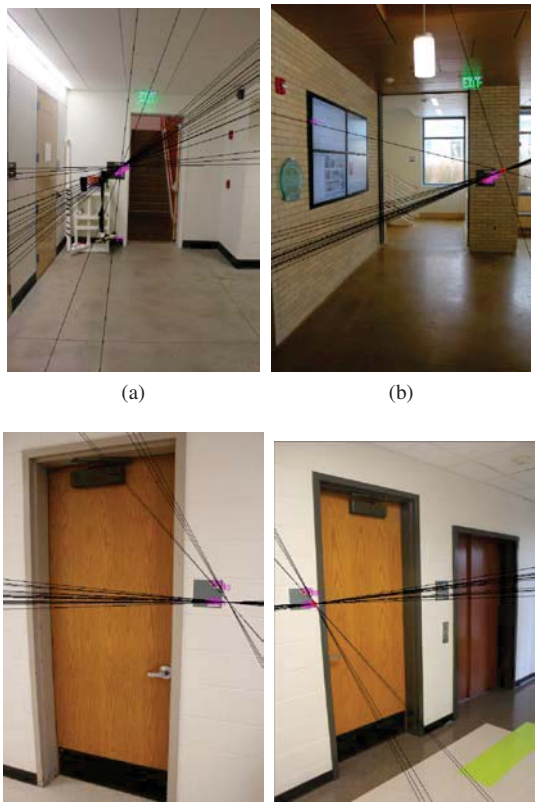


Figure 6. These are two examples of TP matches; (a) and (c) are the retrieved database images to their respective query image (b) and (d). The black lines are the epipolar lines found using the fundamental matrix and the pink numbers are points that are inliers to the fundamental matrix.

The average time to match a single query image to the full database (minus the 30 images for each test set) was 6.22 seconds. While not especially fast, this only needs to be done once, when the user first performs the localization step. After that, the localization steps are performed with a local map that has a much smaller database of images. These steps are much faster, as is described in the next subsection.

This testing procedure was also used to compare the accuracy of using BoW as the indexing method instead of LSH. Following the method of Sivic *et al* [12], a vocabulary of 10,000 words was created from the ORB descriptors extracted from the database images. A query image is then mapped into its constituent visual words, and the distribution of visual words is converted to a "term frequency-inverse document frequency (tf-idf)" vector. The 10 most similar database images are returned as candidates, and the geometric constraint verification step is applied to each of these.

Using BoW as the indexing method resulted in an accuracy of 83.7% which is lower than using LSH, which had 92.5% accuracy. One possible reason for the lower performance of the BoW method is that it computes the distribution of visual words from the entire image. If two



(a) (b) (c) (d)

Figure 7. These are two examples of FP matches; (a) and (c) are the retrieved database images to their respective query image (b) and (d). The black lines are the epipolar lines found using the fundamental matrix and the pink numbers are points that are inliers to the fundamental matrix.

(a) (b)

Figure 8. (a) was the query image used in the test and (b) was the database image. These two images should have resulted in a TP but an FN occurred.

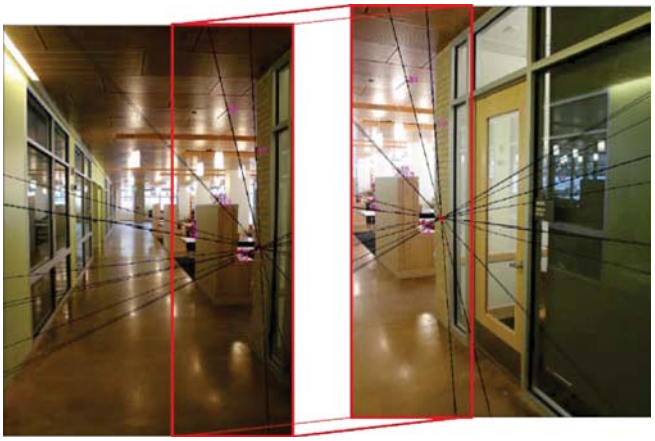


Figure 9. The right is the query image and the left is the database image that it should match. The red rectangle indicates the overlap of the images. LSH correctly matched these two images (the pink points show the matched features), but BoW failed to match these two images.

images have only a small overlapping area, the distributions from the two images can be significantly different. For example, Figure 9 shows a query image and the database image it should match. However, BoW failed to match these two images because they only overlap by 20-30% whereas LSH correctly matched these two images.

3) Local Map Test: Once the first query image is localized by the server, the mobile device receives a local map of images surrounding its current location. The number of images in the local map is chosen so that users will likely remain within this local map for only a short time. Thus, all queries performed in that time frame will most likely correctly match to an image in the local map. The motivation for using a local map is that it will reduce the time to perform a query as well as improve the accuracy of the algorithm.

In our test, 65 images were used to form the local map because that number of images approximates the distance a user who is unfamiliar with a building would travel in about 20 seconds. For query images, 19 test images from the 3^{rd} floor of Brown Hall were captured independently from the database in the same area as the images in the local map.

The results from localization using the local map showed an accuracy of 94.74% (see Table II). The fact that the accuracy of the test is not closer to 100% is because of minor changes to the environment between the time that the database images and the query images were captured (see Figure 10). However, changes like this are to be expected as the indoor environment is not static. If the environment changes the database needs to be updated. This is a problem for this approach as well as the other approaches researched ([2], [15], [16]).

The system took an average of 1.902 seconds to localize a query image. These results show that the algorithm has the potential to run on a mobile device in near real time.

V. Conclusion

In this paper we have presented a novel approach to indoor localization that does not require any additional infrastructure or any special mapping techniques. Using only naturally-occurring features in the environment it was

(a) (b)

Figure 10. The query image (a) incorrectly matched the database image (b). This FP is the result of a change in the environment that caused (a) to have a similar appearance to (b).

Table II
THE RESULTS FOR THE LOCAL MAP TESTS.

Outcome	Number
Query has a match in database and algorithm found a correct match	16
Query has a match in database and algorithm found an incorrect match	1
Query has a match in database and algorithm declared "no match"	0
Query has no match in database and algorithm found an incorrect match	0
Query has no match in database and algorithm declared "no match"	2
Total number of queries	19

demonstrated that our approach can qualitatively localize an image in a large building with a high degree of confidence. The results also show that the use of a local map around the mobile device's known location improves the accuracy of localization. Although the approach was not implemented on a mobile device our analysis shows that it has the potential to run in real time on such a device. Future research could include implementing this approach on a mobile device as well as exploring other ways that the accuracy of the algorithm can improved.

REFERENCES

[1] R. Tesoriero, R. Tebar, J. A. Gallud, M. D. Lozano, and V. M. R. Penichet, "Improving location awareness in indoor spaces using RFID technology," *Expert Systems with Applications*, vol. 37, no. 1, pp. 894–898, 2010.

[2] H. Hile and G. Borriello, "Positioning and Orientation in Indoor Environments Using Camera Phones." *IEEE Computer Graphics and Applications*, vol. 28, no. 4, pp. 32–39, 2008.

[3] R. Mautz and S. Tilch, "Survey of optical indoor positioning systems," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 Intl Conf on*. IEEE, 2011, pp. 1–7.

[4] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf on*. IEEE, 2012, pp. 1043–1050.

[5] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *Computer Vision (ICCV), 2011 IEEE Intl Conf on*. IEEE, 2011, pp. 667–674.

[6] J. Kosecka, L. Zhou, P. Barber, and Z. Duric, "Qualitative image based localization in indoors environments," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conf on*, vol. 2. IEEE, 2003, pp. II–3.

[7] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2010, section 14.4.1 pg 697-701.

[8] Q.-T. Luong and O. D. Faugeras, "The fundamental matrix: Theory, algorithms, and stability analysis," *Intl journal of computer vision*, vol. 17, no. 1, pp. 43–75, 1996.

[9] D. Sinha, M. T. Ahmed, and M. Greenspan, "Image retrieval using landmark indexing for indoor navigation," in *Computer and Robot Vision (CRV), 2014 Canadian Conf on*. IEEE, 2014, pp. 63–70.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Computer Vision (ICCV), 2011 IEEE Intl Conf on*. IEEE, 2011, pp. 2564–2571.

[11] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, 2004, pp. 253–262.

[12] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] J. Egan, *Signal Detection Theory and ROC-analysis*, ser. Academic Press series in cognition and perception. Academic Press, 1975.

[15] M. Werner, M. Kessel, and C. Marouane, "Indoor positioning using smartphone camera," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 Intl Conf on*. IEEE, 2011, pp. 1–6.

[16] H. Kawaji, K. Hatada, T. Yamasaki, and K. Aizawa, "Image-based indoor positioning system: fast image matching using omnidirectional panoramic images," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*. ACM, 2010, pp. 1–4.

MOTION CORRECTED PET SIGNALS COMPRESSING

Krzysztof Malczewski

Department of Electronics and Telecommunications
Poznan University of Technology
Poznan, Poland
kmal@et.put.poznan.pl

ABSTRACT

This article delivers the new compression sensing based super-resolution algorithm for enhancing the image resolution in clinical positron emission tomography (PET) scanners. The concerns of this technique are motion artifacts. The PET measurements are being gathered over a limited period of time. As the patients cannot hold breath during the PET data gathering, spatial blurring and motion artifacts are the typical side effect. These may lead to unreadable scans. It is exposed that the presented algorithm improves PET spatial resolution in cases when Compressed Sensing (CS) sequences are applied. Compressed sensing is able to reconstruct signals from significantly fewer measurements than were traditionally thought necessary. The application of CS to PET has the value for significant scan time reductions, with visible benefits for patients and health care economics. In this work the objective is to combine Super-Resolution image enhancement algorithm with CS framework to achieve high resolution PET output keeping the scans free of motion artifacts. Both methods focus on maximizing image sparsity on known sparse transform domain and minimizing fidelity.

Index Terms— PET/MRI, compressed sensing, super-resolution

1. INTRODUCTION

Recently, the application of CS in MRI and PET has been gaining importance in research interest. Compressed Sensing model was first described in the literature of Information Theory and Approximation [3]. The essence of this technique is it measures a small number of random linear combinations of the signal values–much smaller than the number of signal samples nominally representing it. The crucial is the signal may be reconstructed with sufficient accuracy from these simplified measurements by a nonlinear procedure.

Moreover a nonlinear reconstruction must be done to impose both sparsity of the image representation and consistency with the acquired data. Compressed sensing may be used to multiplex a large number of individual readout sensors to drastically reduce the number of readout channels in a large area PET block detector. The compressed sensing idea can be utilized to treat PET data acquisition as a sparse readout problem and achieve sub-Nyquist rate sampling, where the pixel pitch of all the individual SiPM sensors determines the Nyquist rate. In this way, the sensing matrix is prepared by using discrete elements or wires that uniquely connect pixels to readout channels [5]. Technically, by analyzing the recorded magnitude on several ADC channels, the original pixel values can be recovered even though they have been scrambled through a sensing matrix. In a PET block detector design comprising 128 SiPM pixels arranged in a 16 × 8 array, compressed sensing can provide higher multiplexing ratios (128:16) than Anger logic (128:32) or Cross-strip readout (128:24) patterns while resolving multiple simultaneous hits. Unlike Anger and cross-strip multiplexing, compressed sensing may recover the positions and magnitudes of simultaneous, multiple pixel hits. Interpreting multiple pixel hits can be applied to improve the positioning of events in light-sharing designs, inter-crystal scatter events, or events that pile up in the detector.

Unfortunately, motion artifacts may distort PET images what reduces their utility and reliability. Regrettably PET imaging spatial resolution is limited due to a predetermined detector width [2]. It may be overcome when the number of image samples is increased. Super-resolution (SR) techniques have been used in PET imaging to produce a high-resolution image by combining a set of low-resolution images that have been acquired from different points of view (POV). In this paper, the author proposes a novel technique, which combines super-resolution, motion correction procedures and compressed sampling. The results argued its compellingly in experimental studies. Comparison [5-7] between the SR image and low-resolution images exhibits batter resolution and higher number of details. The presented SR al-

gorithm may replace the present approach in current PET scanners without any hardware modifications.

2. PET MATRIX SENSING

A formal approach for reconstruction could be briefly described the in following way. Represent the reconstructed image by a complex vector *m*, let denote the linear operator that transforms from pixel representation into the chosen representation. Let denote the undersampled Fourier transform, corresponding to one of the k-space undersampling schemes. The reconstructions are obtained by solving the following constrained optimization problem:

$$\text{minimize } \|\psi m\|_1 \qquad (1)$$
$$\text{s.t. } \|F_s m - y\|_2 < \varepsilon$$

where *y* is the measured data from the PET scanner and ε controls the fidelity of the reconstruction to the measured data. The threshold parameter ε is roughly the expected noise level. The l_1 norm means . $\|x\|_1 = \sum_i |x_i|$.

Minimizing the l_1 norm of $\|\psi m\|_1$ promotes sparsity [1,2]. The constraint $\|F_s m - y\|_2 < \varepsilon$ enforces data consistency. Formally, among all solutions that are consistent with the acquired data, we want to find a solution that is compressible by the transform ψ. It is worth mentioning that when finite differencing is used as the sparsifying transform, the objective becomes the well-known total variation (TV) penalty [1-2].

The goal of this work is to focus on the acquisition of 2-D SPECT/PET projections basing on compressive sampling and their reconstruction using a non-linear recovery algorithm.

3. SUPER-RESOLUTION IN PET

Super-Resolution is the problem of generating one or a set of high-resolution images from one or a sequence of low-resolution frames [1]. Most methods have been proposed for super-resolution based on multiple low-resolution images of the same scene, which is called multiple-frame super-resolution.

Accurate measurement and map of structure in living tissues is basically limited by the imaging system attributes. Super-resolution methods allow overcoming limitation of the acquisition devices without any modifications within hardware.

In the algorithm presented below a local-affine adaptive smoothing approach for the regularization in the demons algorithm has been nested into the super-resolution scheme.

4. MOTION CORRECTION

This approach models the dense deformation as a set of local affine transformations, and adaptively smooth out the dense deformation field while preserving the discontinuities along the local affine components using the anisotropic smoothing approach [zrodlo11]. We are assuming that voxels that are close, both spatially and in their intensity value, represent the same structure, and thus have similar affine motion. Hence, the local-affine modeling process relates an affine transform to each voxel, by analyzing its local neighborhood voxels, weighted by their intensity similarity. The coupling of efficient dense deformation estimation with local affine adaptive smoothing yields a better registration algorithm, which is more suitable for the registration of images, affected my internal motion.

The idea of this algorithm is given below:

We are given a patient image I_p. The goal is to find a dense deformation field K_p^r that minimizes its dissimilarity to the reference image I_r. Thirion's demons algorithm [5] computes the deformation field that minimizes the energy:

$$\hat{K}_p^r = \underset{D_p^r}{\operatorname{argmin}} E\left(I_p, I_r, K_p^r\right) + S\left(K_p^r\right) \text{ where}$$

$E\left(I_p, I_r, K_p^r\right)$ represents the dissimilarity measure between the reference and deformed images, and $S\left(K_p^r\right)$ is a regularization term that regulates the smoothness of the follow-on deformation field. The solution is being found by applying the following two successive steps iteratively:

Compute an unconstrained dense deformation field that minimizes the dissimilarity between the reference image and other ones.

Regularize the deformation field by homogeneous isotropic Gaussian smoothing to keep its spatial coherence.

Because the smoothing step may over-smooth the deformation field discontinuities that are associated with independent movements of different organs, we replace the smoothing step (2) with a new anisotropic smoothing filter that is inversely proportional to the differences between the local affine transformations.

Accordingly, the local affine fitting is expressed as a weighted least-squares problem:

$$\hat{A}\left(\vec{x}\right) = \arg\min_A \sum_{\vec{y}\in\Omega_{\vec{x}}} w_{\vec{y}} \cdot \left\| A\left(\vec{y}\right) - D_p^r\left(\vec{y}\right) \right\|^2 \qquad (2)$$

The weights $w_{\vec{y}}$ are defined as:

$$w_y = \exp\left(-\frac{\left(I\left(\vec{x}\right) - I\left(\vec{y}\right)\right)^2}{2\sigma^2} \right) \qquad (3)$$

where σ is a predefined scaling parameter, representing the expected signal intensity homogeneity within same structure. The solution can be found efficiently using Horn's method, or using the SVD method [4].

Local affine domain gradient computation
We use the Frobenius metric between matrices:

$$\|A_1 - A_2\| = \sqrt{\sum_i^N \sum_j^M \left(A_1(i,j) - A_2(i,j)\right)^2} \quad (4)$$

with proper scaling for the translational components of the affine transformation, to define the gradient of the local affine domain. The gradient is then calculated using the finite differences approach.
Local affine adaptive smoothing

Given the local affine domain gradient $\nabla\hat{A}(\vec{x})$, we are using the anisotropic diffusion operator to adaptively smooth the deformation vectors associated with each voxel, while preserving the discontinuities between local affine motions. Following the anisotropic diffusion approach presented in [11], the diffusion operator could be defined as follows:

$$\frac{\partial D_p^r(\vec{x})}{\partial t} = \nabla c(\vec{x}) \cdot \nabla D_p^r + c(\vec{x}) \cdot \nabla D_p^r(\vec{x}) \quad (5)$$

where $c(\vec{x})$ is the diffusion coefficient formulated as:

$$c(\vec{x}) = \exp\left(\frac{\nabla\hat{A}(\vec{x})}{k^2}\right) \quad (6)$$

and *k* is a predefined constant scaling parameter that differentiates between local affine differences that relate to independent organ movements and noise.
In this paper this algorithm has been combined with the iterative framework given below, see figure 2.
The next step is the image alignment procedure, where the motion compensated frames are combined to produce one blurred HR frame by using the L1-norm. Using a regularization-based optimization method is deblurring the so generated HR frame.

5. REGULARIZATION ALGORITHM

In this chapter the cost function optimization method is presented. In details, the goal of this part is to minimize the error between the simulated LR frames and the interpolated observed LR frames. The applied cost function incorporates the consequences of assumed local motion, see figure 2.
The super-resolution algorithms are computationally complex and numerically ill posed problems [5]. The problem of SRR can be expressed as follows:

$$W(x) = \sum_{k=1}^{N} \left\| HF^k X - \hat{Y}^k \right\|_1 + \lambda \|CX\|_1 \quad (7)$$

Where the norm given above is the L1 norm and it describes the cost function measuring error. The $\hat{Y}^k$ means upsampled image frame from the observed sequence Y_k:

$$\hat{Y}^k = Up\left(D^T Y^k\right). \quad (8)$$

Where *Up* is the upscaling operator and *C* is the Laplacian operator , λ is the regularization operator. It is helpful in the ill-posedness.
The super-resolution algorithm may be divided into a couple of steps. The final part is coherent motion image parts fusion. It can be expressed in the following way:

$$\vec{X} = \underset{x}{\operatorname{argmin}} \left\| HX - \hat{Z} \right\|_1 + \lambda \|CX\|_1 \quad (9)$$

where $\vec{X}$ is current *X* high resolution image estimate, $\hat{Z} = H\hat{X}$. Applying the conjugate gradient procedure does technically the minimization, see figure below.

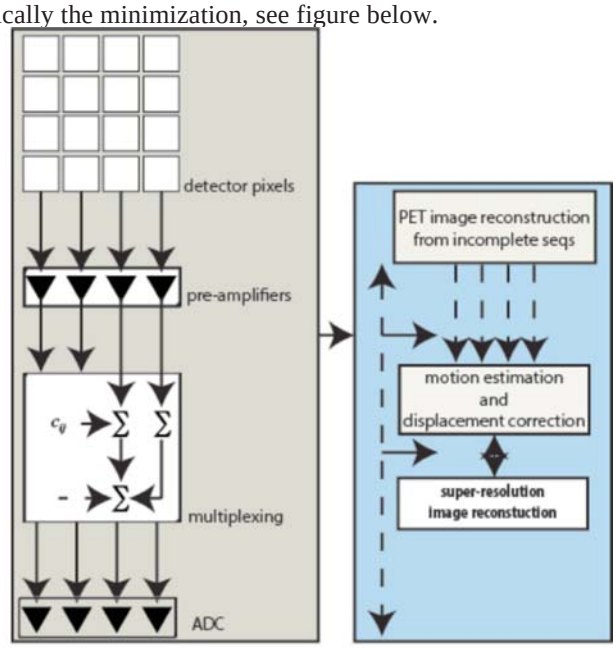


Figure 1. Presented algorithm diagram.

6. DISCUSSIONS AND EXPERIMENT

In this paper, the practicability of implementing super resolution into PET scans has been exposed.

The method has been applied using 18F-FDG PET/CT 1.0 scanner data sets and internal organ shifts as well as transaxial rotational displacement. In this way two different sets were acquired. The first one with all detectors operational (baseline) and once with 8 equidistant detector blocks turned off (partially sampled, 12% detectors have been

turned off. The resulting partially sampled sinogram is after that split into two different components, each sparsely represented in a specific transform domain. An iterative numerical optimization algorithm was then used to recover the PS sinogram based on the solution of a combination of conjugate gradient, underdetermined system of equations and block-coordinated relaxations. In addition, the total variation has been minimized for the first component to direct it into much more convenient a piece-wise smooth model. Finally the two components were added together to achieve the sinogram, which was used to compare with the original PS sinogram. Compressed sensing seems to be the perfect choice for recovering PS PET data. This approach can potentially be used to generate PET images with accurate quantitation while reducing number of detectors/ring.

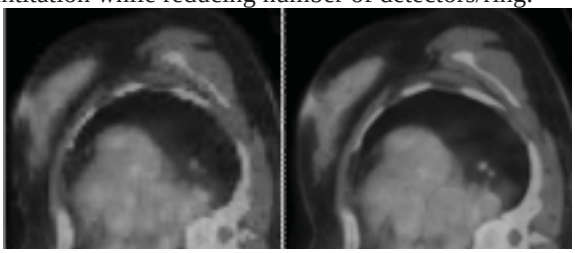


Figure 2. Giant Cell Arteritis Axial fused PET/CT image showing diffuse abnormal FDG uptake in the brachiocephalic and subclavan arteries. From left to right: conventional PET with CS, the super-resolution image with motion correction.

7. CONCLUSION AND FUTURE PROSPECTS

PET/CT and other combined scanners have gained their importance over the last decade. Understanding the nature and purpose of these tools is thus the first step to becoming an important subject of research area. The proposed algorithm may reduce artifacts caused by undersampled data, even in the presence of motion.

This report presents the successful use of a super-resolution algorithm to enhance the resolution of PET images. With an increase in scan time for one FOV, a patient trial showed that the super-resolution technique in the axial direction is feasible in a clinical setting without increasing the radiation dose and with no changes in hardware. As expected, the proposed method improves the spatial resolution, but also enhances noise and artifacts. This effect gets more distinct as the number of super resolution image reconstruction algorithm iterations increases. Preliminary trial results show that the super-resolution approach can be applied to PET imaging, noticeably improving the spatial resolution achievable. During an emission tomography study, induced motion due to patient breathing can lead to artifacts in the reconstructed image. This factor may produce less accurate diagnosis and more important, incorrect radiotherapy planning. The methodology to correct for respiratory motion in the super-resolution image reconstruction step has been developed. It resulted in motion artifacts free scan. Higher resolution PET images provided by a super-resolution algorithm may show a more differentiated anatomical structure, see Figure 2. Considering the success of PET/CT modalities, public expectations for any new combination, such as MR/PET are pretty high. Since MRI does not utilize any ionizing radiation its use is recommended in preference to CT when either modality could yield the same information. Thus, it turns into a perfect anatomical complement to PET. The principal idea behind merging PET and MRI is to combine the functional / metabolic information provided by PET with the high soft-tissue contrast and the functional information offered by MRI. The expectations related to their performance are high, mostly because of the potential for superior tissue contrast inherent in the MR modality, as well as the potential for multiparametric functional imaging in conjunction with PET.

8. REFERENCES

[1] K. Malczewski, „Breaking The Resolution Limit In Medical Imaging Modalities", The 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'12), Worldcomp 2012.

[2] J.A. Kennedy, O. Israel, A. Frenkel, R. Bar-Shalom, and H. Azhari, "Super-resolution in PET imaging" IEEE transactions on medical imaging, 25(2):137{147, 2006.

[3] M. Davenport, "The Fundamentals of Compressive Sensing", IEEE Signal Processing Society Online Tutorial Library, April 12, 2013.

[4] Freiman, M., Voss, S., Warfield, S.: Demons registration with local affine adaptive regularization: application to registration of abdominal structures. In: Proceedings of the 8th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2011. pp. 1219–1222 (2011)

[5] Olcott, P.D., Chinn, G., Levin C.S., "Compressed sensing for the multiplexing of PET detectors", Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE.

[6] J. Dijk, A. W. M. van Eekeren, K. Schutte, D. J. J. de Lange, and L. J. van Vliet, "Super-resolution reconstruction for moving point target detection," Opt. Eng., vol. 47, no. 8, 2008.

[7] T. Q. Pham, L. J. van Vliet, and K. Schutte, "Robust fusion of irregularly sampled data using adaptive normalized convolution," J. Appl. Signal Process., vol. 2006, pp. 1–12, 2006.

Detection of Aberrant Responses in OMR Documents

Raghuveer Kanneganti[1], **Randy Fry**[1], **and Lalit Gupta**[2]
[1] McGraw-Hill Education CTB, Monterey, CA, USA
[2]Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA

Abstract - *Optical Mark Recognition (OMR) is a well-known process of capturing human-marked data in documents. It is an extremely accurate and rapid form of data capture especially when each response can be entered as a single mark and for the same reason this process has been employed in various examinations/tests across the world. It is clearly understood that conducting these tests in an equitable manner is of the utmost importance. Sadly, in the past few years, there have been several cases in which teachers/administrators of elementary and high schools across the United States were identified for fraudulently correcting the answers written by their students in order to improve the success rate of their respective schools. In order to identify this format of cheating, a procedure was developed to autonomously determine if cheating has occurred by detecting the presence of aberrant responses in scanned OMR test books. The challenges introduced by the significant imbalance in the numbers of typical and aberrant bubbles were identified. The aberrant bubble detection problem was formulated as an outlier detection problem. A pool of features is initially selected by examining bubbles that are penciled by a group of individuals and analyzing the differences between them. Several possible outlier detection methods were considered and a feature based procedure in conjunction with a one-class SVM classifier was developed. A multi-criteria rank-of-rank-sum technique was introduced to rank and select a subset of features from a pool of candidate features. Using the data set of 11 individuals, it was shown that a detection accuracy of over 90% is possible.*

Keywords: Feature selection; Optical mark recognition; Texture analysis; Outlier detection; Cheating

1 Introduction

The goal of this research is to develop a procedure to autonomously determine if cheating has occurred during testing by detecting the presence of aberrant responses in scanned Optical Mark Recognition (OMR) test books. OMR is a well-known process of capturing human-marked data in documents. The OMR process is an extremely accurate and rapid form of data capture especially when each response can be entered as a single mark. All other methods of data capture on paper require much more extensive manual or electronic processing [1],[2]. OMR test documents with graphite penciled responses are scanned with an infrared light source because most colored inks used for the background artwork are transparent to this wavelength while carbon absorbs the light. As a result, the background artwork vanishes leaving only the graphite responses and a few other carbon-bearing marks. OMR applications are numerous ranging from consumer surveys to evaluating the performance of students in elementary school and national level examinations such as the Scholastic Aptitude Test (SAT) [3].

It is clearly understood that conducting examinations/tests in an equitable manner is of the utmost importance. Sadly, in the past few years, there have been several cases in which teachers/administrators of elementary and high schools across the United States were identified for fraudulently correcting the answers written by their students in order to improve the success rate of their respective schools [4]. Understandably, there is a dire need in the test scoring industry to develop efficient methods to determine whether cheating has occurred at any level of the testing process.

This research focuses specifically on the detection of aberrant or suspicious responses that are "bubbled" or "circled" in scanned test books which typically may contain over 100 questions. Given the large number of students, it is impractical, if not impossible, for a limited number of investigators to visually examine every test book to determine if aberrant bubbles are present.

Given the characteristics of the detection problem and the inability to design pattern classifiers or compare probability distributions, the approach introduced in the next section formulates aberrant bubble detection as an "outlier" detection problem where aberrant bubbles are considered outliers in a distribution of typical bubbles. By definition, an outlier is an observation that falls outside the overall pattern of a statistical distribution [5]. Outlier detection is applied in numerous applications related to fraud detection [6],[7], intrusion detection [8],[9] and abnormalities in medicine [10],[11] and measurements [12]. Outlier detection methods may be classified as statistical or distance-based methods [12]. The statistical methods assume a probability distribution function for the typical data and outliers are data points that have a low probability of occurring in the assumed distribution [12]. Distance based methods assume that outliers are "far" from the typical data points [13]. The methods may also be

classified as univariate or multivariate depending on whether the data is single-variable or multi-variable, respectively [14].

In order to detect outliers, a one-class Support Vector Machine classifier, which has been shown to be quite effective for outlier detection [15],[16],[17] is employed. Specifically, the LIBSVM implementation [18] is selected. The input to the classifier is a set of features that are ranked by means of a multi-criteria ranking strategy. The final set of features is selected by conducting detection experiments using a data set consisting of bubbles marked by 11 individuals.

2 Outlier Detection Approach

The potential outliers are detected by analyzing only those responses to questions which have erasures of wrong answers and un-erased correct answers. Such corrected responses are referred to as wrong-to-right (WTR) corrections. Currently, the WTR corrections are determined automatically during the scanning of the responses by first detecting erasures based on intensities using infrared light scanners. Furthermore, post-processing queries in order to identify individual test books with large numbers of WTR corrections. Suspicious behavior is flagged when the number of WTR corrections are exceptionally high for a given class or a school district. However, the flagged test books have to be analyzed visually to determine whether none, or some, or all of the WTR corrections were made by an individual other than student. This is a painstakingly slow and tedious process. As mentioned in the introduction, the goal is to automatically detect aberrant WTR corrections. Only those test books with high numbers of detected aberrant responses are analyzed visually to determine if cheating has occurred.

The bubbles in a test book with unusually high WTR corrections can be modelled as the mixture

$$B(x,y) = b(x,y) \cup b_E(x,y) \tag{1}$$

where $\cup$ represents the mixture operation and $b(x,y)$and $b_E(x,y)$ are the bubbles in the un-erased responses and the bubbles in the WTR responses, respectively. The WTR bubbles can, in turn, be modelled as the mixture

$$b_E(x,y) = b_i(x,y) \cup b_a(x,y) \tag{2}$$

where $b_i(x,y)$ and $b_a(x,y)$ are innocent (made by the student) and the aberrant bubbles, respectively. Therefore, the ensemble of bubbles in a text book can be written as

$$B(x,y) = b(x,y) \cup b_i(x,y) \cup b_a(x,y) \tag{3}$$

The goal, therefore, is to detect the bubbles $b_a(x,y)$, if they exist, in $B(x,y)$. Given the above model, the following outlier detection approaches are possible:

1. Blind multiple outlier detection: The bubbles in a flagged test book are pooled into $B(x,y)$and the outliers are detected in the pooled set. It is hoped that the outliers detected will belong to $b_a(x,y)$. The statistical and distance-based outlier detection methods can be used for this case.
2. Targeted multiple outlier detection: The set $B(x,y)$ is divided into the sets $b(x,y)$and $b_E(x,y)$. The bubbles in $b_E(x,y)$ are regarded as potential outliers and collectively tested against the bubbles in $b(x,y)$. Statistical outlier detection methods such as the multivariate two sample Hotelling's T2 test can be used for this case. Although an elegant method, implementation issues are likely to occur due to the lack of enough bubbles in $b_E(x,y)$.
3. Targeted single outlier detection: The set $B(x,y)$ is divided into the sets $b(x,y)$and $b_E(x,y)$. The bubbles in $b_E(x,y)$ are regarded as potential outliers and each bubble in $b_E(x,y)$ is individually tested against the bubbles in $b(x,y)$. Distance-based outlier detections methods can be used for this case.

Flagging Rule for Visual Analysis :

For a given test book, if N_E is the total number of bubbles in the set of WTR corrections and N_A is the number of outliers detected, then, the test book is flagged for visual analysis if

$$\frac{N_A}{N_E} > T_a \tag{4}$$

where T_a a user is defined threshold. For example, setting $T_a = 0.5$ will correspond to the simple majority rule.

3 Feature Selection

Because the direct comparison of bubbles using template matching is not effective, a feature based approach is introduced to represent the bubbles. A pool of features is initially selected by examining bubbles that are penciled by a large group of individuals and analyzing the differences between them. The analysis led to the following choice of candidate features:

Histogram based features: It is observed that the bubbles of different individuals tend to have different intensity distributions resulting from the applied pressure variations and pencil sharpness. If $b(x,y)$ is a bubble and $p(z_i), i = 0,1,\dots,(L-1)$, is its histogram, candidate distinguishing histogram features extracted from the histogram include:

Mean:	$m = \sum_{i=0}^{L-1} z_i p(z_i)$	(5)
Variance:	$\sigma^2(z) = \sum_{i=0}^{L-1} p_i (z_i - m)^2$	(6)
Skew:	$\mu_3(z) = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$	(7)
Kurtosis:	$\mu_4(z) = \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i)$	(8)
MAD:	$median(\lvert b(x,y) - median[b(x,y)]\rvert)$	(9)

Texture based features: The texture or coarseness of the bubbles tend to vary across individuals due to variations in the bubbling style and the pencil sharpness. The variations in the bubbling style are due to the manner in which individuals fill out bubbles. For example, individuals may fill bubbles by scratching across, clock-wise, counter-clockwise, circular, starting from the center, and starting from the outside. The candidate distinguishing texture features include 13 Haralick

features [19] computed from the normalized co-occurance matrix of the response bubble and:

$$\text{Smoothness:}\quad S(t) = 1 - \frac{1}{1+\sigma^2(z)} \tag{10}$$

Shape based features: The bubbles of individuals tend to vary in shape. The shape features, which range from almost circular to elliptical with different orientations, can be derived from the moments given by

$$\text{Moments of order } (i+j)\text{: } M_{ij} = \sum_x \sum_y x^i\, y^j\, b(x,y) \tag{11}$$

Central moments of order :

$$(p+q)\text{:}\quad \mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q b(x,y) \tag{12}$$

$$\text{Centroid:}\quad \{\bar{x},\bar{y}\} = \{M_{10}/M_{00}\,, M_{01}/M_{00}\} \tag{13}$$

Second order central moments:

$$\mu'_{20} = \mu_{20}/\mu_{00} = M_{20}/M_{00} - \bar{x}^2 \tag{14}$$

$$\mu'_{02} = \mu_{02}/\mu_{00} = M_{02}/M_{00} - \bar{y}^2 \tag{15}$$

$$\mu'_{11} = \mu_{11}/\mu_{00} = M_{11}/M_{00} - \bar{x}\bar{y} \tag{16}$$

The candidate shape features include

$$\text{Area:}\quad A = M_{00}, \sum_x \sum_y b(x,y) \tag{17}$$

$$\text{Centroid difference:}\quad C_d = |\bar{x} - \bar{y}| \tag{18}$$

$$\text{Perimeter:}\quad P = \sum_x \sum_y b(\hat{x},\hat{y}),$$

where $\hat{x},\hat{y}$ are boundary elements of $b(x,y)$ (19)

$$\text{Circularity:}\quad S_f = \frac{P^2}{A} \tag{20}$$

$$\text{Area difference:}\quad A_d = \pi r^2 - M_{00}\text{ , where, } r = \frac{L_M}{2} \tag{21}$$

Major axis length: $L_M = \sqrt{(x+y)^2 - f}$ where f is the distance between foci and x and y are the distances of each foci to any point on the ellipse that has the same normalized second central moments as the bubble (22)

$$\text{Minor axis length:}\quad L_m = (x+y) \tag{23}$$

$$\text{Eccentricity:}\quad \varepsilon = \sqrt{1 - \frac{\lambda_2}{\lambda_1}}$$

where,

$$\lambda_1 = \frac{\mu'_{20}+\mu'_{02}}{2} + \frac{\sqrt{4{\mu'_{11}}^2 + (\mu'_{20}-\mu'_{02})^2}}{2},$$

$$\lambda_2 = \frac{\mu'_{20}+\mu'_{02}}{2} - \frac{\sqrt{4{\mu'_{11}}^2 + (\mu'_{20}-\mu'_{02})^2}}{2} \tag{24}$$

$$\text{Orientation:}\quad \Theta = \frac{1}{2}\arctan\left(\frac{2\mu'_{11}}{\mu'_{20}-\mu'_{02}}\right) \tag{25}$$

Convex area: $A_c =$ number of pixels in 'Convex bubble'

$$\text{Equiv. diameter:}\quad D_E = \sqrt{\frac{4(A)}{\pi}} \tag{26}$$

$$\text{Solidity :}\quad S_H = \frac{A}{Convex\ area} \tag{27}$$

$$\text{Extent:}\quad E = \frac{number\ of\ pixels\ in\ bubble}{number\ of\ pixels\ in\ bounding\ box} \tag{28}$$

Contour based features: The contour of a bubble is the boundary of the bubble. A signature (one-dimensional representation) of a bubble can be obtained from the clockwise ordered distance of each contour pixel to the centroid [20]. Candidate features can be extracted from the signatures because they tend to vary in shape. If X=[x_1, x_2, x_3... x_n] is the signature consisting of the ordered distance of each contour pixel to centroid, the candidate features set include:

$$\text{Variance:}\quad \sigma^2(X) = E\left[\left(X - E(X)\right)^2\right] \tag{29}$$

$$\text{Skewness:}\quad \mu_3(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{(n-1)S^3}\text{, where, } S \text{ is the standard deviation of X} \tag{30}$$

$$\text{Kurtosis:}\quad \mu_4(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{(n-1)S^4} \tag{31}$$

Divergence based features: Divergence is a measure of the amount of flux entering or leaving a point. Plots of the divergences of bubbles show differences which can be exploited to generate distinguishing features. Let $D(x,y)$ be the divergence of $b(x,y)$ and $D = [d_1, d_2, \ldots, d_n]$ be the vector of divergence values of the contour of the bubble. The following features can be extracted from D:

$$\text{Mean:}\quad \mu(D) = \frac{1}{n}\sum_{i=1}^n D \tag{32}$$

$$\text{Variance:}\quad \sigma^2(D) = E\left[\left(D - E(D)\right)^2\right] \tag{33}$$

$$\text{Skewness:}\quad \mu_3(D) = \frac{\sum_{i=1}^n (D_i - \bar{D})^3}{(n-1)S^3}\text{ , where, } S \text{ is the standard deviation of D} \tag{34}$$

$$\text{Kurtosis:}\quad \mu_4(D) = \frac{\sum_{i=1}^n (D_i - \bar{D})^4}{(n-1)S^4} \tag{35}$$

The final candidate pool, therefore, consists of 39 features.

4 Feature Ranking

The features can be ranked and selected according to several criteria such as the interclass separation, classification accuracy [21], coefficient of variation [22], Relief [23], and PCA combinations [24]. In general, the rankings obtained from different criteria vary across data sets and it may be difficult to select the best criterion for a given application. Instead of using a single criterion, a selection rule can be formulated from a combination of suitable criteria. In this research, a rank-of-rank-sum approach [24] is used to rank and systematically select a subset of features from the candidate pool using the interclass separation, classification accuracy, and coefficient of variation criteria.

Interclass Separation Criterion :

The interclass separation is a normalized measure of the separation between two distribution means [21]. Given a pool of K features, if $f(k)$ and $g(k)$ are the kth features of the bubbles $b(x,y)$ and $b_E(x,y)$, respectively, the interclass separation between $f(k)$ and $g(k)$ is given by

$$\rho(k) = \frac{\left[\bar{f}(k) - \bar{g}(k)\right]^2}{\sum_{i=1}^n \left[f_i(k) - \bar{f}(k)\right]^2 + \sum_{j=1}^m \left[g_j(k) - \bar{g}(k)\right]^2},$$

$$k = 1,2, \dots, K \quad (36)$$

where, $\bar{f}(k)$ and $\bar{g}(k)$ are the means of *k*th feature in the two classes and $f_i(k), i = 1,2, \dots, n$, and $g_j(k), j = 1,2, \dots, m$, are the ensembles of features $f(k)$ and $g(k)$, respectively.The ranking of the features is given by

$$R_1(k) = Rank[\rho(k)] \quad (37)$$

where the feature with the highest inter-class separation is assigned rank 1(best) and that with the lowest is assigned *K* (worst) because higher values of $\rho(k)$ indicate both higher interclass and lower intra-class separations.

Classification Accuracy Criterion :
The features can be ranked according to their classification accuracies using a suitable classifier. If a univariate Gaussian classifier is used to test an element z_k as belonging to class ω (un-erased or aberrant), the discriminant function for determining the classification accuracy of the *k*th feature is given by

$$g_{k,\omega}[z_k] = -\ln \sigma_\omega^2 - [z_k - \mu_\omega]^2 / 2[\sigma_\omega^2] + \ln P_\omega, \text{ k} = 1,2,...\text{K}, \quad (38)$$

where, μ_ω and σ_ω^2 are the mean and variance of z_k under class ω, respectively, and P_ω is the *a priori* probability of class ω. The test element z_k is assigned to the discriminant function that yields the highest value. That is, the test sample z_k is assigned to a category ω^* given by

$$\omega^* = \arg\max_c \{g_{k,\omega}[\ z_k]\} \quad (39)$$

If $\alpha(k)$ is the classification accuracy of the classifier for feature k, then, the ranking of the features is given by directly ranking $\alpha(k)$. For this case

$$R_2(k) = Rank[\alpha(k)] \quad (40)$$

is a ranking such that the feature with the highest classification accuracy is assigned rank 1 (best) and the lowest is assigned rank *K* (worst).

Coefficient of Variation Criterion :
The coefficient of variation is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation and the mean of a distribution. That is, the coefficient of variation of the *k*th feature is given by

$$\beta(k) = \frac{\sigma(k)}{\mu(k)},\ k = 1,2, \dots, K \quad (41)$$

The ranking of the features is given by

$$R_3(k) = Rank[\beta(k)] \quad (42)$$

where the feature with the smallest coefficient of variation is assigned rank 1 (best) and that with the highest is assigned *K*(worst) because smaller values indicate smaller intra-class variations. Note that $\beta(k)$ is a single-class parameter and is computed only from the un-erased bubbles.

Multi-Criteria Feature Ranking
If the sum of the ranks of feature k across the three criteria is

$$S(k) = \sum_{i=1}^{3} R_i(k) \quad (43)$$

then, the final rank of the *k*th feature is given by

$$R(k) = Rank[S(k)] \quad (44)$$

Where $R(k)$is the rank of the rank sums. In this rank ordering, the features are ranked such that the best feature has the smallest $R(k)$ value and the worst feature has the largest $R(k)$ value.

5 Outlier Detection

The one-class SupportVector Machine (OSVM), which has been shown to be quite effective for outlier detection is employed in this research. Specifically, the LIBSVM implementation described [18] was selected. Given training vectors $x_t \in R^n, i = 1, \dots . l$ without any class information, the primal problem of OSVM is

$$Min_{w,\varepsilon,p} \frac{1}{2}\omega^T\omega - \rho + \frac{1}{vl}\sum_{i=1}^{l} \varepsilon_i \quad (45)$$

subject to

$$\omega^T\varphi(x_i) \geq \rho - \epsilon_i, \epsilon_i \geq 0, i = 1, \dots, l. \quad (46)$$

The dual problem is

$$Min_{\propto} = \frac{1}{2} \propto^T Q \propto \quad (47)$$

subject to

$$0 \leq \propto_i \leq \frac{1}{vl}, i = 1, \dots, l \quad (48)$$

$$e^t \propto = 1 \quad (49)$$

$v \in [0,1]$ is an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors, (ω, ρ) are a weight vector and an offset parameterizing a hyper-plane in the feature space associated with the kernel, $l \in N$is the number of observations. $\varphi$is a feature map $X \to F$, a map into an inner product space F such that the inner product in the image of φ can be computed by evaluating some simple kernel and $\propto$ is a multiplier $\propto \geq 0$. $\varepsilon_i$is called a slack variable which measures the degree of misclassification of x_i. $e = [1, \dots \dots, l]^T$ is the vector of all positive classes, where $Q_{ij} = K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$. The decision function is

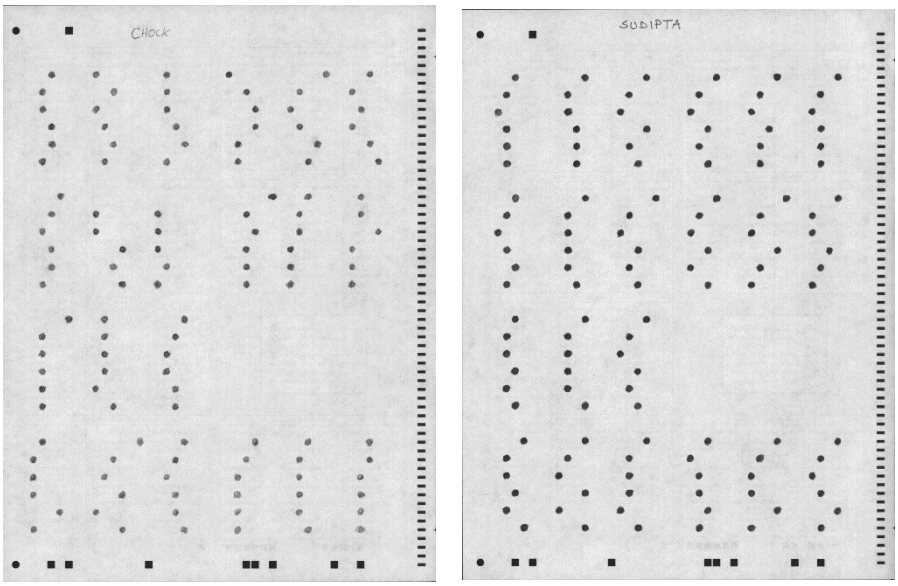


Fig. 1 Example bubbles marked by 2 individuals

$$sgn\left(\sum_{i=1}^{l} \alpha_i K(x_i, x) - \rho\right) \quad (50)$$

Solving the above OSVM implementation in LIBSVM is equivalent to solving

$$Min_{\propto} = \frac{1}{2} \propto^T Q \propto \quad (51)$$

subject to $0 \leq \propto_i \leq \frac{1}{vl}, i = 1, \ldots, l$

$$e^t \propto = 1 \quad (52)$$

6 Experimentation and Performance Evaluation

The data set consisted of bubbles that were penciled by 11 different individuals. Each individual marked 120 bubbles in a typical test book. This set serves as a large artificial data that could be used to rank and select features across a large number of users. Figure 1 shows a few examples of the bubbles of 2 individuals. Figure 2 shows zoomed sections of Figure 1 to facilitate visual comparisons. The features were extracted from the biggest contour *C* that encompasses the bubble based on the Otsu thresholding algorithm [25].

The data of each individual were randomly partitioned into equal-size training and test sets. The training sets were used to rank the 39 candidate features. Using the method outlined in Section 4, the features were ranked in a pairwise fashion for two individuals at a time. The bubbles of one individual in the pair were regarded as the un-erased bubbles and bubbles of the other individual were regarded as the WTR bubbles. The total number of rankings for each feature across the 11 subjects was (11)(10)/2=55 for the interclass and classification accuracy criteria. For the coefficient of variance criteria, each feature had 11 rankings. For each criteria, the sum of the ranks was determined and the rank of the rank-sum was computed. The final ranking of the features across the 11 individuals was obtained from the rank of the rank-sum. The features were normalized to take values in the interval [0,1].

The performance of the SVM outlier classifier was evaluated systematically to determine the *L* coefficients, out of the 39 coefficients, which gave the best detection results. A SVM classifier was developed for the following cases: the highest ranked feature, the two highest ranked features,…, all 39 features. The detection accuracy, defined as the percentage of correctly detected aberrant bubbles, was computed for each case. Figure 3 shows the averaged detection results. The figure clearly shows that a detection accuracy of 90.87% is possible using the top 10 ranked features.

Template Matching :
It was stated in the introduction and in Section 3 that template matching was not effective for detecting aberrant bubbles directly. In order to confirm this, template matching was implemented on the dataset in a pairwise fashion and classification accuracies of 82% and 65% were obtained across the entire set using normalized cross-correlation (r) and mean square error (MSE), respectively. The template for each individual was the average of the bubbles in the respective training set and the template matching threshold was varied as a percentage of inliers in the training set. The normalized cross correlation (r) and mean square error (MSE) were computed as follows:

$$r = \frac{1}{n}\sum_{x,y}\frac{(b_E(x,y)-\overline{b_E})(t(x,y)-\bar{t})}{\sigma_b\sigma_t} \quad (53)$$

$$MSE = \frac{1}{n}\sum_{x,y}(b_E(x,y) - t(x,y))^2 \quad (54)$$

where $t(x,y)$ and $b_E(x,y)$ are the template and target images, n is the number of elements in the images, and σ_b and σ_t are the standard deviations of $b_E(x,y)$ and $t(x,y)$, respectively. The results are shown in Figure 4.

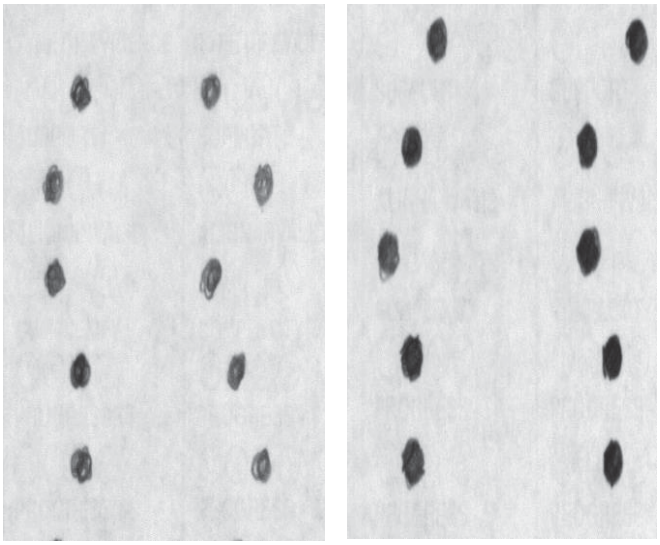


Fig. 2 Zoomed section of bubbles in figure 1

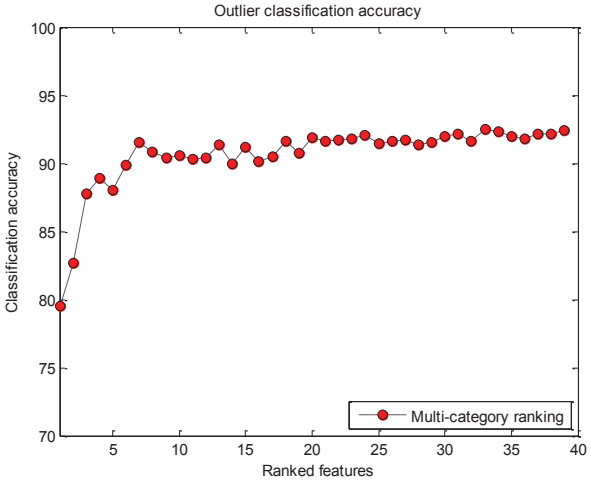


Fig. 3 Detection results using one-class SVM classifier

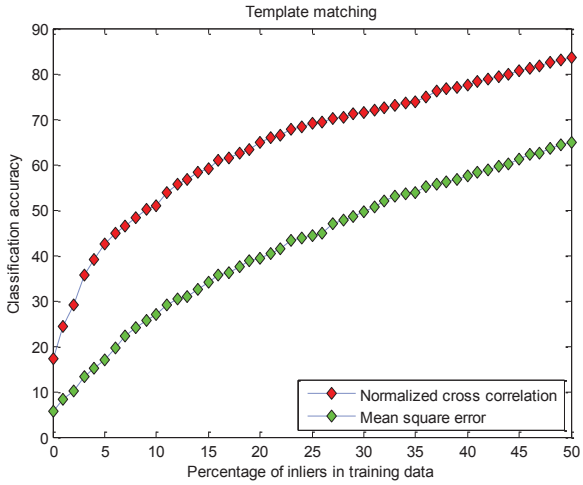


Fig. 4 Template matching results

7 Conclusions

The goal of this research was to develop a procedure to autonomously determine if cheating has occurred by detecting the presence of aberrant responses in scanned Optical Mark Recognition (OMR) test books. The challenges introduced by the significant imbalance in the numbers of typical and aberrant bubbles were identified. The aberrant bubble detection problem was formulated as an outlier detection problem. Several possible outlier detection method were considered and a feature based procedure in conjunction with a one-class SVM classifier was developed. A multi-criteria rank-of-rank-sum technique was introduced to rank and select a subset of features from a pool of candidate features. Using the data set of 11 individuals, it was shown that a detection accuracy of over 90% is possible.

8 Acknowledgements

This research project was supported by CTB/McGraw Hill LLC under a grant entitled "Forensic Erasure Analysis." The authors would like to express their sincere thanks to Dr. Wim van der Linden and Michelle Boyer for their invaluable support during the entire project. The opinions and conclusions contained in this paper are solely those of the author and do not necessarily reflect the policy or position of CTB/McGraw-Hill.

9 References

[1] R. Wynn, "Optical mark recognition," *Data processing,* vol. 26, no. 9, pp. 26-27, 1984.

[2] "http://www.apperson.com/datalink-main/omr-technology," [Online].

[3] "http://en.wikipedia.org/wiki/Optical_mark_recognition," [Online].

[4] "http://www.foxnews.com/us/2012/11/25/teachers-embroiled-in-test-taking-fraud-for-15-years-feds-say/,"

[Online].

[5] D. S. Moore and G. P. McCabe, The introduction to the practice of statistics, New York: W.H. Freeman and Company, 1999.

[6] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science,* vol. 17, no. 3, pp. 235-249, 2002.

[7] Y. Kou, C.-T. Lu, S. S and Y.-P. Huang, "Survey of fraud detection techniques," in *IEEE Intenational Conference on Networking, Sensing and Control*, 2004.

[8] P. Winter, E. Hermann and M. Zeilinger, "Inductive Intrusion Detection in Flow-Based Network Data Using One-Class Support Vector Machines," in *4th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2011.

[9] J. J. Davis and A. J. Clark, "Data preprocessing for anomaly based network intrusion detection: A review," *Computers & Security,* vol. 30, no. 6-7, pp. 353-375, 2011.

[10] M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper and G. Clermont, "Outlier detection for patient monitoring and alerting," *Journal of Biomedical Informatics,* vol. 46, no. 1, pp. 47-55, 2013.

[11] Y. Yang, L. Ji and J. Wu, "Outlier detection in heart rate signal using activity information," in *10th World congress on intelligent control and automation*, 2012.

[12] V. J. Hodge and J. Austin, "A survey of Outlier Detection Methodologies," *Artificial Intelligence Review,* vol. 22, no. 2, pp. 85-126, 2004.

[13] K. Zhang, M. Hutter and H. Jin, "A newlocal distance - based outlier detection approach for scattered real-world data," *Advances in knowledge, discovery and data mining, lecturer notes in computer science,* vol. 5476, pp. 813-822, 2009.

[14] B. V and L. T, Outliers in statistical data, 3 ed., John Wiley & Sons, 1994.

[15] L. M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," *Journal of Machine Learning Research,* vol. 2, no. 2001, pp. 139-154, 2001.

[16] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. Williams, J. Shawe-Taylor and M. Brammer, "Patient classification as an outlier detection problem: An application of the one-class support vector machine," *NeuroImage,* vol. 58, no. 3, pp. 793-804, 2011.

[17] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation,* vol. 13, pp. 1443-1471, 2001.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," in *ACM transactions on intelligent systems and technology*, 2011.

[19] R. M. Haralick, S. K and D. Itshak, "Textural features for image classification," *IEEE transactions on systems, man and cybernetics,* Vols. SMC-3, no. 6, pp. 610-621, 1973.

[20] L. Gupta and M. D. Srinath, "Contour sequence moments for the classification of closed planar shapes," *Pattern recognition,* vol. 20, no. 3, pp. 267-272, 1987.

[21] L. Gupta, B. Chung, M.D. Srinath, D. L. Molfese and H. Kook, "Multi-channel fusion models for the parametric classification of differential brain activity," *IEEE transactions on biomedical engineering,* vol. 52, no. 11, pp. 1869-1881, 2005.

[22] K. D and U. P, "Class specific feature selection for identity validation using dynamic signatures," *Biometrics and Biostatistics,* 2013.

[23] R.-S. M and K. I, "Theoritical and empirical analysis of ReliefF and RReliefF," *Machine learning journal,* vol. 53, pp. 23-69, 2003.

[24] L. Gupta, S. Kota, S. Murali, D. L. Molfese and R. Vaidyanathan, "A feature ranking strategy to facilitate multivariate signal classification," *IEEE transactions on systems, man and cybernetics,* vol. 40, no. 1, pp. 98-108, 2010.

[25] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man and cybernetics,* vol. 01, pp. 62-66, 1979.

Automatic Estimation of Optimal Resected Liver Regions Considering Practical Surgical Conditions

Masanori Hariyama[1], **Takeaki Suzuki**[1], **Keisuke Maeda**[1], **Mitsugi Shimoda**[2], **Keiichi Kubota**[2]
[1]Graduate School of Information Sciences, Tohoku University, Japan
[2] Second Department of Surgery, Dokkyo Medical University, Japan

Abstract— *This article presents an automatic approach to estimate optimal resected-liver regions for oncologic surgery planning considering practical surgical conditions. Since the liver has complex vessel structure, it is difficult for human to find optimal resected liver regions. We provide efficient approaches for two types of practical problems: (1) Finding an optimal resected region under cut volume limitation, and (2) Finding an optimal cut surface based on a 3-dimensional(3-D) parabola model. For both problems, a tumor domination ratio is efficiently used to find all portal vessels related to tumors. The experimental results demonstrate that the resected liver regions of the proposed approach are much smaller than those of the conventional manual approach in most cases.*

Keywords: Medical imaging, 3D simulation analysis, anatomic hepatectomy, local thickness

1. Introduction

3D simulation plays an important role in surgical planning for hepatectomy since the liver has a complex structure, that is, some different vessels are arranged complexly as shown in Fig. 1. The estimation of regions perfused by the portal vein is especially one important task in anatomical hepatectomy, since the hepatocellular carcinoma(HCC) tends to metastasize via the portal vein [1], [2]. Figure 2 explains how the perfused region is estimated. The tumor affects the nearest parts of the portal vein; from the nearest parts, the HCC metastasize in the downstream direction via the portal vein. The overestimate of the perfused region tends to prevent the metastases, but it increases the possibility of the post-operative liver failure. Therefore, the volume of perfused region should be minimized while including all the perfused regions of the portal vein feeding the tumor.[2]

In preoperative planning, surgeons search for the combination of the cut points on the portal vein in a trial-and-error way until the tumor is covered by the perfused regions of the portal cut points. However, it is difficult and time-consuming for surgeons to find optimal or near-optimal combination since the portals vein has complex structure, that is, it has many branches. Surgeons select simple combinations, and this results in overestimation of the perfused region.

In order to solve this problem, an approach using a tumor domination ratio have been proposed[3],[3]. The tumor donation ratio reflects how much volume of the tumor a part of the portal vein feeds, and allows to pick up all the tumor-related parts of the portal vein.

However, the resected region resulting from the tumor donation ratio (called "100%" resected region) can not be used to practical surgeries in most cases. The patients suffering from HCC tends to have much lower liver functionality than normal persons; their upper limits of the resected-region volumes are much smaller than those of normal persons. As a result, 100% resected region must be shrunk so as to meet the upper limit of the resected-region volume.

Another problem of the 100% resected region that the cut-points may not be recognized by surgeons in real surgeries since surgeons cannot recognize veins thiner than 6mm. This problem occurs when tumors exists near the liver surface since the veins get thiner around liver surface.

In the following, we provide efficient approaches to solve these two practical problems:

(1) Finding an optimal resected region under cut volume limitation for the former problem
(2) Finding an optimal cut surface based on a 3-dimensional(3-D) parabola model for the latter problem

2. Estimating an 100% resected region using the tumor donation ratio[3],[4]

This section briefly explains 100% resected region and how to get it since the method given in Sections 3 and 4 is based on the 100% resected region. The region perfumed by a point on the portal vein is estimated by using a Voronoi diagram[2]. Figure 3 (a) shows a Voronoi diagram. A Voronoi diagram is a way of dividing space into a number of regions. A set of points (called seeds) is specified beforehand and for each seed there will be a corresponding region consisting of all points closer to that seed than to any other. The regions are called Voronoi cells. A seed and Voronoi cell in a Voronoi diagram correspond to a point on a portal vein and a region perfumed by the portal point as shown in Fig.3 .

The tumor domination ratio TDR of a portal point P is defined as follows.

$$TDR = \frac{\text{[Volume of the region perfused by } P \text{ in the tumor]}}{\text{[Volume of the tumor]}} \times 100 \ \ [\%], \quad (1)$$

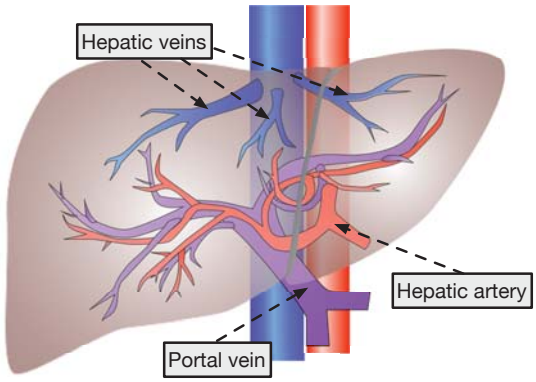


Fig. 1: Vascular system of the liver.

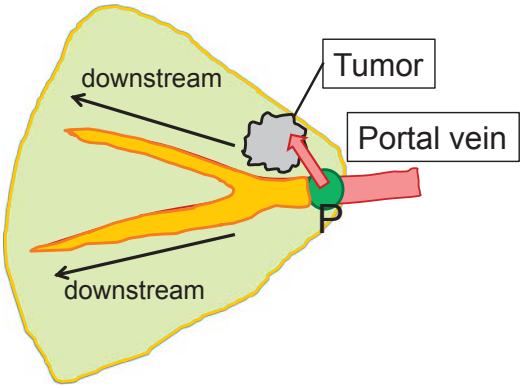


Fig. 2: Region perfused by the portal point P and its downstream.

where the denominator and the enumerator are illustrated in Fig.3 (c). The tumor donation ratio reflects how much the portal-vein point feed the tumor and is affected by the tumor.

Given a tumor location, let us compute an 100% resected region by using the tumor donation ratio, where the 100% resected region means a minimum region including all sub-regions perfused by tumor-related portal points. All tumor-related portal points are found by using the tumor donation ratio; all the portal point with TDR larger than 0 are picked up as tumor-related portal points. Finally, the 100% resected region is given as the union of all the regions perfused by the tumor-related portal points.

This procedure to find an 100% resected region is extended so as to consider more practical conditions such as condition limiting cut points to branch points with pre-set thickness of veins[3].

3. Finding an optimal resected region under a cut-volume limitation

Although the 100% resected region is ideal from the view of preventing the cancer from metastasizing, it may not be acceptable for most patients because of their low liver functionality. In pre-operative planning, surgeons usually determine the upper limit of the cut volume depending the liver functionality of each patient. Hence, from the practical view, it is important to finding an optimal resected region under a cut-volume limitation.

In the following, we describe the formulation. Let V_{max} be the upper limit of the total cut volume for a patient. This parameter is given by surgeons based on the result of the pre-operative tests for a patient. Let TDR_{low} be the lower limit of the tumor donation ratio; we do not consider the cut points whose tumor donation ratio is less than TDR_{low}. Parameter TDR_{low} is empirically set to around 5% so as to the optimization result is acceptable for most surgeons. The reason why we use TDR_{low} is that, without this parameter, the search procedure sometimes select the cut points with large cut volumes and small tumor donation ratios, and this optimization result is unacceptable for most surgeons. As cut points, we select portal-vein branch points that satisfies the following conditions:

- branch points,
- the thickness is larger than the pre-set thickness,
- the tumor donation ratio is larger than TDR_{low}.

Let $CP = \{c_i | 1 \leq i \leq N\}$ the set of the cut points, where N is the total number of the cut points.

The optimization problem can be formulated as follows:

$$minimize \quad \sum_{i}^{N} T_i \tag{2}$$

under the condition

$$\sum_{i}^{N} V_i \leq V_{max}. \tag{3}$$

Figures 4 and 5 shows the optimization results for two samples. The left figure (a) shows the 100% resected region; the right figure (b) shows the proposed resected region in this paper. These results demonstrate that our proposed method can reduce the total cut volume, and provides almost same results as those by surgeons automatically.

4. Finding an optimal cut surface based on a 3-D parabola model

This section describes another optimization problem where the tumor exists around the liver surface. When the tumor exists around the liver surface, surgeons usually plan to scoop the tumor out from the liver. The shape of the scooped region can be well approximated by a 3-dimensional(3-D) parabola as shown in Figure **??**. In order to prevent the cancer from metastasizing, the scooped region should include the 100% resected region while the its volume is minimized. A arbitrary 3-D parabola has 8 parameters as

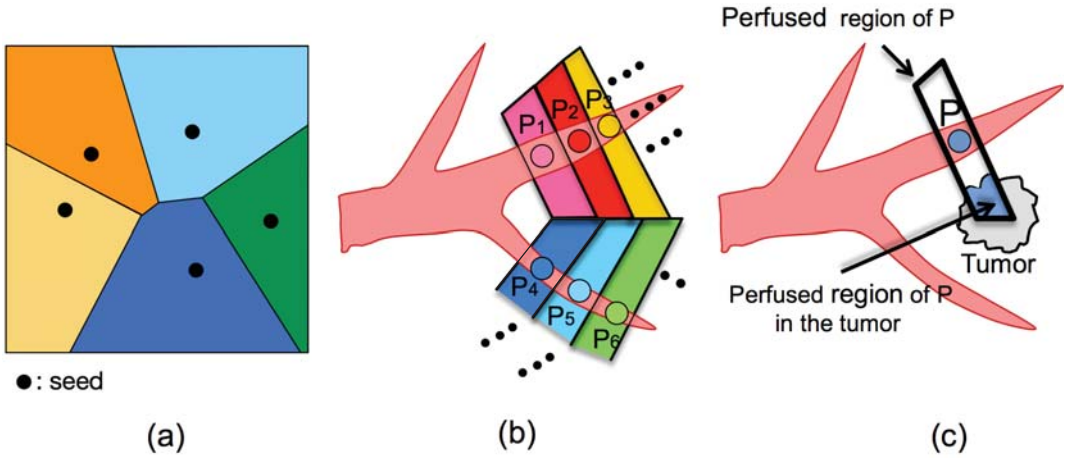


Fig. 3: Estimating perfused regions by portal-vein points based on Voronoi diagram.

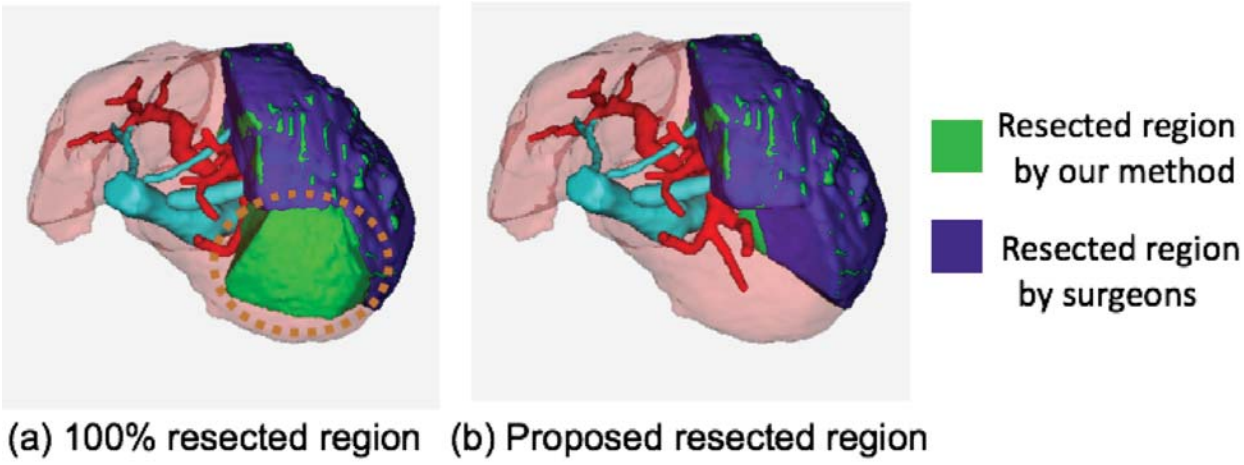


Fig. 4: Result for sample 1.

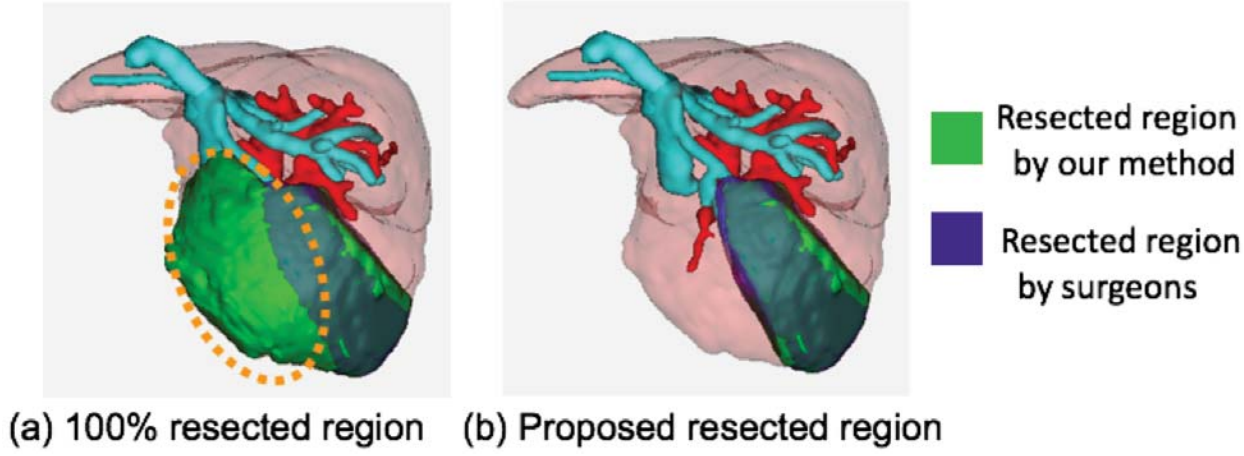


Fig. 5: Result for sample 2.

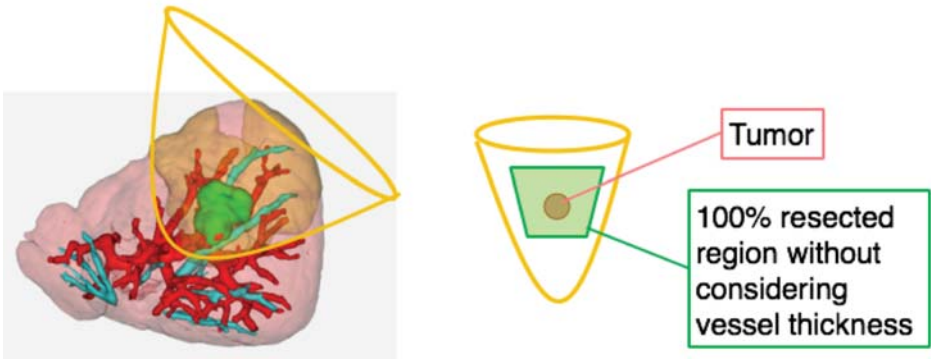
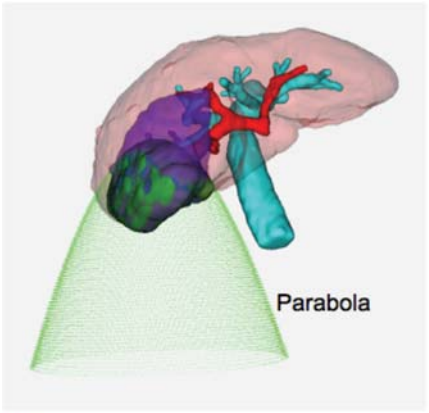


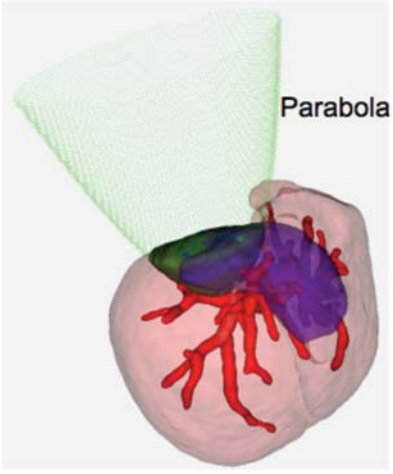
Fig. 6: Scooping a tumor out in the parabola shape.



100% resected region	75.5[cc]
Resected region by parabola-baesd optimization	123.8[cc]
Resected region by surgeon	189.9[cc]

Resected region planned by surgeons manually
Resected region by parabola-based optimization

Fig. 7: Result 1.



100% resected region	11.5[cc]
Resected region by parabola-baesd optimization	66.2[cc]
Resected region by surgeon	116.4[cc]

Resected region planned by surgeons manually
Resected region by parabola-based optimization

Fig. 8: Result 2.

follows:

Shape parameters(Curvatures): A_x, A_y

Rotation angles in 3-D space: roll, pitch, yaw

Translation distances in 3-D space: T_x, T_y, T_z

When the tumor shape, the tumor location, and the 100% resected region are given, the 8 parameters of a parabola are optimized in such a way that the scooped volume is minimized while the parabola still includes the 100% resected region. Note the 100% resected region without considering the thickness of vessels is used here, and is obtained through the processing described in Section 2. In search process, some efficient search techniques are used since the search has a lot of time consuming procedures such as 3-D inclusion check. Figures **??** and **??** show the optimization results. In each result, the cut volume using parabola approximation is reduced to around half of the manual planning by surgeons. Note that the 100% resected region used here cannot be used cut volume in real surgery although its volume is smaller than that of the volume using parabola approximation; this is because the 100% resected region does not consider vessel thickness to obtain as small a resected region as possible; the vessel thickness used to compute the 100% resected region is too thin for surgeons to recognize in real surgery. From this observation, we can say that our approach provides good results from a practical view point.

5. Conclusion

This paper have presented practical approaches to estimate an optimal resected region for practical surgical conditions. The key to success to minimize the cut volume while preventing a cancer from metastasizing is the use of the tumor donation ratio that reflects how much portal-vein point affects the cancer tumor. As future work, cut volume estimation considering hepatic veins is remaining since, the blood congestion occurs which reduces the liver functionality when the hepatic vein is cut.

References

[1] M. Makuuchi, H. Hasegawa, S. Yamazaki, “Ultrasonically guided subsegmentectomy“, Surg. Gynecol. Obstet., Vol.161,pp.346-50 (1985).

[2] T. Takamoto, T. Hashimoto, S. Ogata, K. Inoue, Y. Maruyama, A. Miyazaki, M. Makuuchi, “Planning of anatomical liver segmentectomy and subsegmentechtomy with 3-dimensional simulation software”, The American Journal of Surgery Vol. 206, Issue 4, pp. 530-538 (2003).

[3] Masanori Hariyama, Moe Okada, Mitsugi Shimoda, Keiichi Kubota, "Estimation of Resected Liver Regions Using a Tumor Domination Ratio", Proc.International Conference on Image Processing, Computer Vision, and Pattern Recognition(IPCV), pp.52-56, (2014).

[4] Masanori Hariyama and Mitsugi Shimoda, "Automatic estimation of a resected liver region using a tumor domination ratio", in "Emerging Trends in Image Processing, Computer Vision, and Pattern Recognition", Chapter 23(pp.369–378), Morgan Kaufmann Publishers

Automated Distortion Defect Inspection of Curved Car Mirrors Using Computer Vision

Hong-Dar Lin, Kuan-Shen Hsieh
Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung, 41349, Taiwan

Abstract – *Currently, the curved mirrors have been widely used in vehicle rearview mirrors and security mirrors on the driving roads and make drivers have better fields of views and driving information. In the production process of curved car mirrors, mirrors with reflected distortion defects result from the unstable temperature changes of ovens and inappropriate control of over-flow fusion process. It is not easy to measure the magnitudes of distortion defects on curved car mirrors. This study proposes a novel approach based on small-shift process control schemes to inspect reflected distortion defects on curved car mirrors. We first detect the intersection points of the standard reflected pattern, then measure the distances of the intersection points from the origin (center point), and calculate the distance deviations of the corresponding intersection points between the defective and normal images. Finally, we apply the cumulative sum (CUSUM) and EWMA control methods to judge the existence of the distortion defects based on the accumulative deviation distances.*

Keywords: Industrial inspection; curved car mirrors; distortion defects; computer vision system; small-shift control scheme.

1 Introduction

Comparing with plane car mirrors, curved car mirrors have the characteristics of higher reflectance and wider field of view. Currently, the curved mirrors have been widely used in vehicle rearview mirrors and security mirrors on the driving roads and make drivers have better driving information. Since the reflected distortion defects and surface defects directly affect the display quality of car mirrors, the detection of the kinds of defects is very important for car mirror manufacturers. In the production process of curved car mirrors, mirrors with reflected distortion defects result from the unstable temperature changes of ovens and inappropriate control of over-flow fusion process. Since the distortion defects do not have regular shapes and clear boundaries, it is not easy to measure the magnitudes of distortion defects on curved mirrors. Furthermore, the curved mirrors with the property of higher reflection increase the difficulty of discrimination of the distortion defects on car mirrors. Therefore, this research aims at exploring the automated visual inspection of reflected distortion defects of the curved car mirrors.

The defective car mirrors with distortion defects providing shape-distorted scene information may lead car drivers making wrong decisions when driving. Figure 1 shows the normal and defective images of curved car mirror surfaces with reflection of street scene. The object shapes reflected in the defective image are significantly distorted. The mirror distortion defects may make reflected objects look irregularly, out of focus, and blurry in the defective images. These distorted images may result in making wrong judgment by car drivers and lead to dangerous car accidents.

(a) (b)

Figure 1. The curved car mirror images with reflection of street scene: (a) a normal image; (b) a defective image with distortion defect.

Inspection difficulties of surface defects are existing in manufacturing process. Surface defects affect not only the appearance of industrial parts but also their functionality, efficiency and stability. The most common detection methods for surface defects are human visual inspections. Human inspection is vulnerable to wrong judgments owing to inspectors' subjectivity and eye fatigues. Furthermore, difficulties also exist in precisely inspecting distortion defects by machine vision systems because when product images are being captured, the region of a distortion defect could expand, shrink or even disappear due to uneven illumination of the environment, different view angles of the inspectors, shapes of reflected patterns, and so on.

Current automated computer vision system (off-line and sampling) uses a concentric circle pattern reflected on mirrors to acquire images and quantize distortion magnitude for selection. It is hard to precisely inspect the mirror distortion flaws by current machine vision systems due to high reflection. The property of higher reflection on curved mirrors increases the difficulty of discrimination of the distortion defects on car

mirrors. In this research, the testing samples with length 18.1 cm, width 10.71 cm, and thickness 0.2 cm, were randomly selected from manufacturing process of car mirrors. Figure 2 shows the dimension of the testing sample and a testing sample with high reflection on mirror surface.

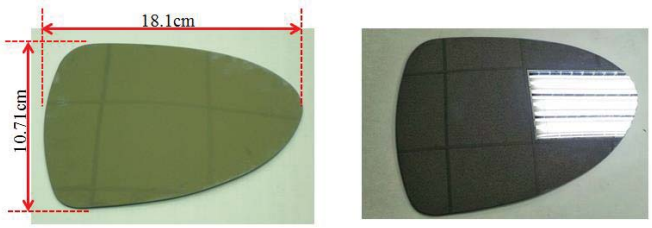


Figure 2. Dimension of the testing sample and a testing sample with high reflection.

This study proposes a vision system with a trapezoidal mask for image acquisition and applies cumulative sum control schemes to inspect distortion defects on curved car mirrors. To quantify the deformation (degree of distortion) of a car mirror, a inspection standard pattern (checkerboard grids) is used to reflect the pattern on a testing car mirror for image acquisition. The reflected pattern image of a defective mirror with distortion is compared with that of a normal mirror for quantifying the deformation and locating the distortion defects.

2 Automated defect inspections

Automated visual inspection of surface flaws has become a critical task for manufacturers who strive to improve product quality and production efficiency [1-3]. Chiou [4] presented an intelligent method for automatic selection of a proper image segmentation method upon detecting a particular flaw type in roll-to-roll web inspection. The results show a significant reduction in misclassification rate from about 44% to 13.96%. Perng et al. [5] developed a fast and robust machine vision system for wire bonding inspection. A new lighting environment was devised which will highlight the slope of the bonding wire and suppress the background from being extracted. Adamo et al. [6] proposed a low-cost inspection system based on the Canny edge detection for online defects assessment in satin glass. Liu et al. [7] presented the method based on watershed transform methods to segment the possible defective regions and extract features of bottle wall by rules. Then wavelet transform are used to exact features of bottle finish from images.

Many researches explored the defect detection of glass related products. Li and Tsai [8] proposed a wavelet-based discriminant measure for defect inspection in multi-crystalline solar wafer images with inhomogeneous texture. The proposed method performs effectively for detecting fingerprint, contaminant, and saw-mark defects in solar wafer surfaces. Lin and Tsai [9] presented a Fourier transform-based approach to inspect surface defects of capacitive touch panels. A multi-crisscross filter is designed to filter out the frequency components of the principal band regions. In the restored image, the defective region will be clearly retained. Chiu and Lin [10] applied block discrete cosine transform, Hotelling's T-squared statistic, and grey clustering technique for the automatic detection of visual blemishes in curved surfaces of LED lenses.

Regarding the distortion correction techniques, Duan and Wu [11] proposed a new method for distortion correction in the barrel distortion of wide-angle lens. The cubic B-spline interpolation function was adopted to interpolate the surface and the bi-linear interpolation was used to reconstruct the gray level of pixels. Simulation results show that the method can make a good correction of the coordinate position and gray value. Zhang et al. [12] presented a distortion-correction technique that can automatically calculate correction parameters, without precise knowledge of horizontal and vertical orientation. The method is applicable to any camera-distortion correction situation. Based on a least-squares estimation, the proposed algorithm considers line fits in both field-of-view directions and global consistency that gives the optimal image center and expansion coefficients. Ngo and Asari [13] presented an architecture design for real-time correction of nonlinear distortion in wide-viewing angle camera images. The architecture is designed based on the method of back mapping the pixels in the corrected image space to the distorted image space and performing linear interpolation of four neighboring pixel intensities. The distortion correction coefficients are obtained by the least-squares estimation technique. Smith and Smith [14] proposed a methodology for improving the accuracy of machine vision calibration through applying regression analysis and neural network modelling. The regression analysis has been employed for assisting with the data collation and organization needed for implementation of neural network training. The neural network was developed for modelling the error in the measured location of image features such as a matrix of dots.

Most of the distortion related works focus on the distortion correction of optical lenses. Most of the automated inspection systems of glass and mirrors mainly detect surface defects and the distortion defect is not included. It is difficult to precisely detect reflected distortion defects embedded on surface of curved car mirrors with high reflection. Currently, there are very few literatures on inspection of mirror distortion defects using automated visual inspection system. In this research, a small-shift control scheme based vision system is proposed to detect reflected distortion defects on curved mirrors.

3 Proposed method

To quantify the deformation (degree of distortion) of a car mirror, this research proposes a inspection standard pattern (checkerboard grids) to reflect the pattern on a testing car mirror for image acquisition. The reflected pattern image of a defective mirror with distortion is compared with that of a

normal mirror for quantifying the deformation and locating the distortion defects. Firstly, we detect the intersection points of the inspection standard pattern, then measure the distances of the intersection points from the origin, and calculate the distance deviations of the corresponding intersection points between the defective and normal images. Finally, we apply the small-shift control schemes to judge the existence of the distortion defects based on the detection of the slight changes of the distance deviations.

3.1 Image acquisition

To clearly capture images with proper reflection for further process, this study proposes a vision system with a trapezoidal mask for image acquisition shown in Figure 3. The testing sample is put in the bottom of the mask and the standard pattern is attached on the top inside the mask. Figure 4 shows the three-view drawings of the trapezoid mask and the specifications of the standard concentric circle pattern. To acquire the images with proper reflected intensity, the control of lighting environment is very important. Figure 5 demonstrates the image acquisition with trapezoidal mask and light sources: (a) the captured image without light sources; (b) the captured image with proper light sources.

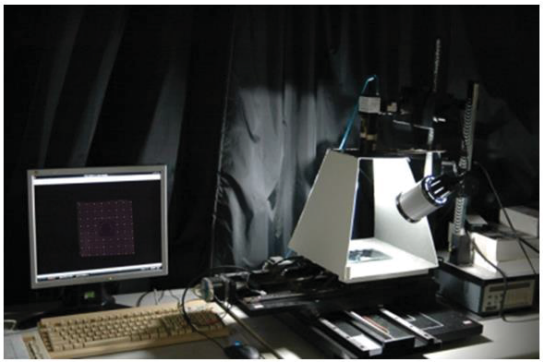


Figure 3. The proposed vision system with a trapezoidal mask for image acquisition.

3.2 Image process procedures

The captured testing image will be processed in several steps. Figure 6 shows the results and differences performed the proposed approach for detecting distortion defects in curved car mirrors. Figure 6(a) and (b) present the captured testing image and the corresponding gray level image using the checkerboard pattern. Figure 6(c) depicts the binary image that the Otsu method applied to do segmentation. Figure 6(d) describes the feature extraction of the feature points in the checkerboard pattern and the cumulative sum charts of the small-shift detection method. And, Figure 6(e) is the resulting image that show the detected distortion defects in red by the proposed detection method. The results reveal that the slight distortion defects in curved mirror surface are correctly separated in the binary image, regardless of insignificant distortion differences.

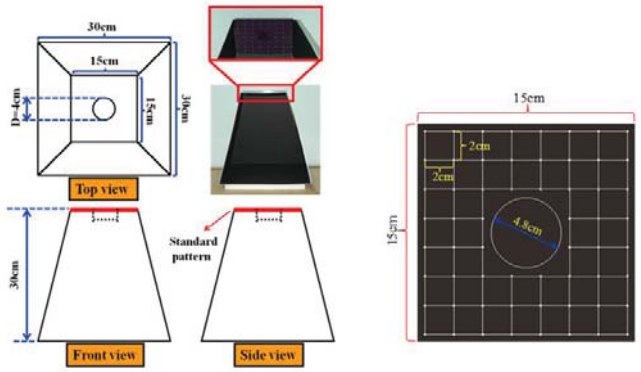


Figure 4. (a) Three-view drawings of the trapezoid mask; (b) Specifications of the standard checkerboard pattern.

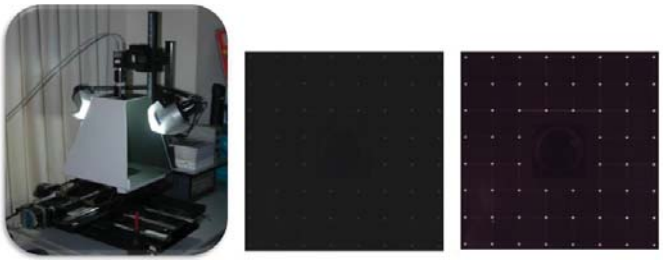


Figure 5. Image acquisition with trapezoidal mask and light sources: (a) the captured image without light sources; (b) the captured image with proper light sources.

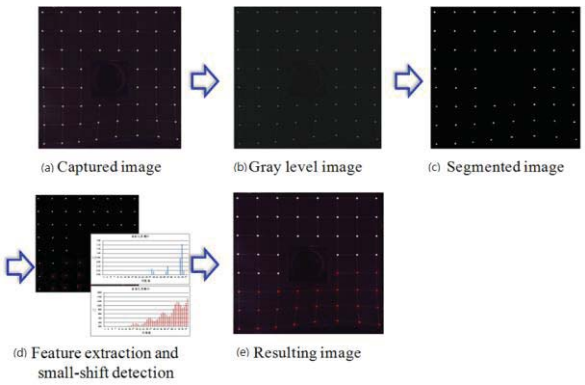


Figure 6. Procedures of the image process flow by the proposed method.

3.3 Standard concentric circle pattern

A standard concentric circle pattern includes 6 concentric circles in this study. Figure 7(a) shows the definition of the feature points in the concentric circle pattern and Figure 7(b) illustrates the coordinates of 8 intersection points on the innermost concentric circle. For each of the 6 concentric circles, 8 intersection points are $I_{i,j}$ with coordinates $(x_{i,j}, y_{i,j})$ and the feature values are the distances $d_{i,j}$ between the intersection points and center point of the concentric circles. The center point $O(x, y)$ is determined by the 8 intersection points of the innermost circle:

$$O(x,y)=(\frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n}x_{i,j},\frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n}y_{i,j}) \quad (1)$$

where $m=1,n=8$.

The feature values $d_{i,j}$ are the distances calculated from $O(x,y)$ and $I_{i,j}(x_{i,j},y_{i,j})$, correspondingly:

$$d_{i,j}=\sqrt{(x_{i,j}-x)^2+(y_{i,j}-y)^2} \quad (2)$$

The distances of a testing image will be compared with those of a normal image to measure the deviations of the corresponding distances for detecting the distortion defects.

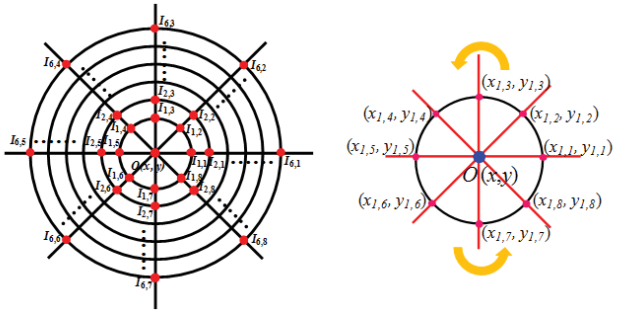


Figure 7. (a) Definition of the feature points in the concentric circle pattern; (b) The coordinates of 8 intersection points on the innermost concentric circle.

3.4 Standard checkerboard grid pattern

A standard checkerboard grid pattern includes 3 concentric squares in this study. Figure 8(a) shows the definition of feature points in the checkerboard pattern and Figure 8(b) illustrates the coordinates of 12 intersection points on the innermost concentric square. For the 3 concentric squares, 60 intersection points are $I_{i,j}$ with coordinates $(x_{i,j},y_{i,j})$ and feature values are distances $d_{i,j}$ between the intersection points and center point $O(x,y)$ of the concentric squares. The center point $O(x,y)$ is determined by 12 intersection points of the innermost square:

$$O(x,y)=(\frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n}x_{i,j},\frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n}y_{i,j}) \quad (3)$$

where $m=1,n=12$.

The feature values $d_{i,j}$ are the distances calculated from $O(x,y)$ and $I_{i,j}(x_{i,j},y_{i,j})$, correspondingly:

$$d_{i,j}(x_{i,j},y_{i,j})=\max(|x-x_{i,j}|,|y-y_{i,j}|) \quad (4)$$

Similarly, the distances of a testing image will be compared with those of a normal image to measure the deviations of the corresponding distances for detecting the distortion defects.

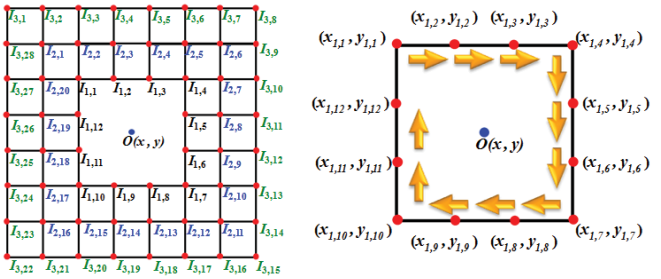


Figure 8. (a) Definition of feature points in the checkerboard pattern; (b) The coordinates of 12 intersection points on the innermost concentric square.

3.5 Small shift control schemes - CUSUM methods

We measure the distances of the intersection points from the origin, and calculate the distance deviations $\Delta d_{i,j}(\Delta d_{i,j}=d_{i,j}-\overline{d_{i,j}})$ of the corresponding intersection points between the testing image $(d_{i,j})$ and normal image $(\overline{d_{i,j}})$. The small-shift control schemes is applied to detect the slight changes of the distance deviations for detecting distortion defects.

3.5.1 Tabular CUSUM method

To detect slight changes in the distance deviations, this research proposes the CUSUM algorithm, which is commonly used in statistical process control to detect the slight shift or deviation from the normal production process [15, 16]. Generally, the CUSUM method processes data, that are smooth in the beginning periods and that deviate slightly in the later periods. The cusum scheme works by accumulating derivations from μ_0 that are above target with one statistic C_s^+ and accumulating derivations from μ_0 that are below target with another statistic C_s^-. The statistics C_s^+ and C_s^- are called one-sided upper and lower CUSUMs (cumulative sum), respectively. They are computed as follows:

$$C_i^+=\max\lfloor 0,\Delta d_{i,j}-(\mu_0+K)+C_{i-1}^+\rfloor \quad (5)$$

$$C_i^-=\max\lfloor 0,(\mu_0-K)-\Delta d_{i,j}+C_{i-1}^-\rfloor \quad (6)$$

where $C_0^+=C_0^-=0, K=(\delta/2)\sigma$.

In Eqs. (5) and (6), K is usually called the reference value, and it is often chosen about halfway between the target μ_0 and the out-of-control value of the mean μ_1 that we are interested in detecting quickly. Thus, if the shift is expressed in standard deviation units as $\mu_1=\mu_0+\delta\sigma$, then K is half the magnitude of the shift.

$$K = \frac{\delta}{2}\sigma = \frac{|\mu_1 - \mu_0|}{2} \Rightarrow \delta\sigma = |\mu_1 - \mu_0| \Rightarrow \delta = \frac{|\mu_1 - \mu_0|}{\sigma} \quad (7)$$

Note that C_s^+ and C_s^- accumulate deviations from the target value μ_0 that are greater than K, with both quantities reset to zero on becoming negative. When either C_s^+ or C_s^- exceeds the decision interval H, the sample set is considered to be out-of-control. A reasonable value for H is five times the standard deviation σ [17].

3.5.2 Standardized CUSUM method

Two advantages of the standardized Cusum scheme, the choices of the parameters k and h do not depend on standard deviation. The other is the standardized Cusum scheme leads naturally to a cusum for controlling variability [17]. The standardized Cusums are defined as:

$$y_i = \frac{x_i - \mu_0}{\sigma} \quad (8)$$

$$C_i^+ = \max\left[0, y_i - k + C_{i-1}^+\right],\ C_i^- = \max\left[0, -k - y_i + C_{i-1}^-\right] \quad (9)$$

where the initial values $C_i^+ = C_i^- = 0$, $i = 0$.

3.6 Small shift control scheme - EWMA method

The exponentially weighted moving average (EWMA) control method is also a good alternative in detecting small shifts [17, 18]. The exponentially weighted moving average Z_i is defined as:

$$Z_i = \lambda x_i + (1-\lambda) Z_{i-1} \quad (10)$$

where $0 < \lambda \leqq 1$ is a constant and the starting value is the process target $Z_0 = \mu_0$. The values of the parameter λ smoothing constant or called weight in the interval 0.05~0.25 work well in practice. A good rule of thumb is to use smaller value of λ to detect smaller shifts. The control limits for the EWMA control method are as follows:

$$UCL_i = \overline{X} + L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda^{2i})\right]} \quad (11)$$

$$LCL_i = \overline{X} - L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda^{2i})\right]} \quad (12)$$

The design parameters of the chart are the multiple of sigma used in the control limits (L) and the value of λ. The performance of the EWMA control scheme is approximately equivalent to that of the CUSUM method, and in some ways it is easier to set up and operate.

4 Experiments and analyses

To evaluate performance of the proposed approaches, experiments were conducted on real curved car mirrors, provided by a car mirror manufacturing company. All samples were randomly selected from manufacturing process of car mirrors. Testing images (386) of the curved car mirrors, of which 136 have no defects and 250 have various reflected distortion defects, were tested. Each image of the surface has a size of 256 × 256 pixels and a gray level of 8 bits. The mirror distortion defect detection algorithm is edited in Matlab language and executed on the 7th version of the MATLAB interactive environment (data analysis, algorithm development, and model creations and applications). The system is implemented on a personal computer with CPU Core 2 Duo 2.33 GHz and 2GB D-RAM.

The higher the performance evaluation indices, $(1-\alpha)$ and $(1-\beta)$, the more accurate the detection results. Statistical type I error α suggests the probability of producing false alarms, i.e. detecting normal regions as distortion defects. Statistical type II error β implies the probability of producing missing alarms, which fail to alarm real distortion defects. Area of normal region detected as distortion defects is divided by the area of actual normal region to obtain type I error, and the area of undetected distortion defects by the area of actual distortion defects to obtain type II error. Correct classification rate (CR) is defined as: $CR = (N_{cc} + N_{dd}) / N_{total}$ where N_{cc} is the pixel number of normal textures detected as normal areas, N_{dd} is the pixel number of real distortion defects detected as defective regions, and N_{total} is the total pixel number of a testing image.

As the decision threshold value changes, so do its false alarm rate (α) and detection rate $(1-\beta)$, both of which are used to describe the performance of a test according to hypothesis testing theory [19]. When various decision thresholds are used, their pairs of false alarm rates and detection rates are plotted as points on a Receiver Operating Characteristic (ROC) curve. The upper-left corner indicates a 100% detection rate and a 0% false alarm rate. The more the ROC curve approaches the upper-left corner, the better the test performs. In industrial practice, a more than 90% detection rate and a less than 10% false alarm rate are a good rule of thumb for performance evaluation of a vision system.

The choices of the parameters k and h determine the control limits of the cumulative sum schemes. Table 1 shows the parameter settings and performance evaluation of the two cumulative sum schemes. Figure 9 demonstrates the ROC curve of the proposed CUSUM methods with different parameter settings of k and h values, respectively. It shows the defect detection performance of the standardized CUSUM method with parameter settings (k, h) values of (1.5, 5), (1.5,

5.2), or (1.75, 4) is better than those of the Tabular CUSUM method with parameter settings (k,h) values of (2.25, 4.6) or (2.25, 4.8). Similarly, the selections of the parameters λ and L decide the control limits of the EWMA method. Figure 10 shows the ROC curve of the proposed EWMA method with different parameter settings of λ and L values, respectively. It indicates the defect detection performance of the EWMA method with parameter settings (λ,L) values of (0.8, 4) has the best detection result with false alarm rate 4.41% and defect detection rate 98%. Accordingly, an appropriate approach and good parameter settings, with its ROC curve closest to the upper-left corner, outperforms the other methods. This implies that the more accurate parameter settings of the small-shift detection approaches are selected, the better the defect detection results will have.

Table 1. Parameter settings and performance evaluation of the two cumulative sum schemes.

Performance Evaluation	Tabular CUSUM	Standardized CUSUM
α	5.88%	5.88%
(1-β)	95.60%	98.00%
Parameter	k=2.25 h=4.6、4.8	k=1.5、h=5、5.2& k=1.75、h=4

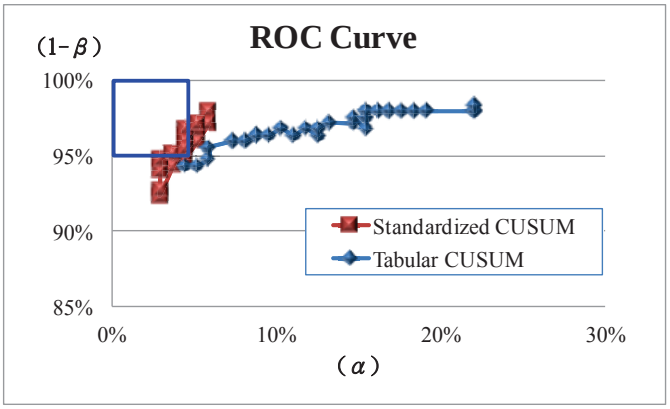


Figure 9. ROC curves of CUSUM methods.

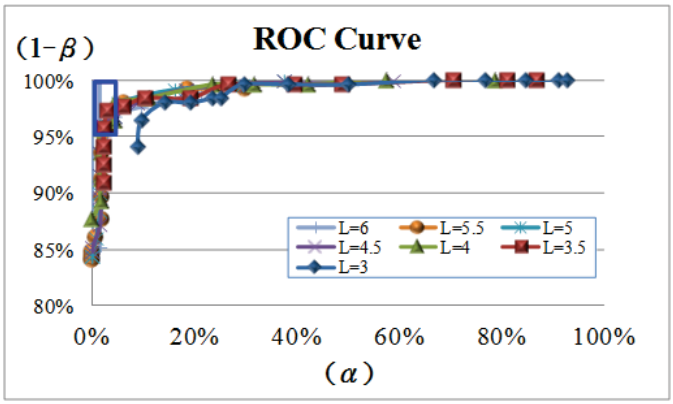


Figure 10. ROC curves of EWMA method.

The current visual inspection method uses the concentric circle pattern as the inspection standard pattern to measure the magnitudes of distortion defects on curved car mirrors. For the concentric circle pattern, we calculate the distortion rate ε %:

$$\varepsilon\% = \frac{\left|d_{i,j} - \overline{d_i}\right|}{\overline{d_i}} \times 100\% \tag{13}$$

where d_{ij} is the distance between the intersection point $I(i,j)$ and the center point $O(x,y)$, $\overline{d_i}$ is the average of the distances of the 8 intersection points on the same concentric circle i.

$$\overline{d_i} = \frac{d_{i,1} + d_{i,2} + \cdots + d_{i,8}}{8} \tag{14}$$

For a normal curved mirror, the distortion rate ε % $\leqq$ 3.8%. And, for a normal plane mirror, the distortion rate ε % $\leqq$ 1.7%. If the distortion rate of a testing curved mirror image is more than 3.8%, we can conclude that some distortion defects exist in the image. Table 2 shows the parameter settings and performance evaluation of the current visual inspection method. The current method achieves the best detection result with false alarm rate 4.41% and defect detection rate 98% when the threshold of distortion rat 1.15 is applied.

Table 2. Parameter settings and performance evaluation of the current visual inspection method.

Control limits	1	1.05	1.1	**1.15**	1.2
α	8.09%	8.09%	5.15%	**4.41%**	2.94%
1-β	93.20%	92.40%	90.80%	**90.80%**	88.40%
Control limits	1.25	1.3	1.35	2.5	3.8
α	2.21%	2.21%	2.21%	0.00%	0.00%
1-β	84.40%	82.40%	81.60%	39.20%	31.60%

To compare the performance of the mirror distortion defect detection, Table 3 summarizes the detection results of our experiments. Three small-shift detection approaches and two traditional techniques are evaluated against the results by professional inspectors. The average defect detection rates $(1-\beta)$ of all testing samples by the five methods are, respectively, 95.6% (Tabular CUSUM method), 98.0% (Standardized CUSUM method), 98.0% (EWMA method), 95.55% (Shewhart method) [17], and 90.8% (current method). However, the two small-shift CUSUM methods have slightly higher false alarm rates (α), 5.88% (Tabular CUSUM method) and 5.88% (Standardized CUSUM method). On the contrary, the other small-shift detection approach has rather lower false alarm rate, 4.41% (EWMA method). The proposed EWMA method has higher correct classification rates (CR) than do the other methods applied to distortion defect detection of curved car mirror images. The average

computation time for processing an image of 256 × 256 pixels is as follows: 2.10 seconds by Tabular CUSUM method, 2.13 seconds by Standardized CUSUM method, 1.97 seconds by EWMA method, and 1.10 seconds by the current method. Hence, the proposed small-shift EWMA method can overcomes the difficulties of detecting distortion defects on curved car mirror and excels in its ability of correctly discriminating slight distortion defects from normal regions.

Table 3. Summarized comparison table of distortion defect detection of curved car mirror for five different methods.

	Tabular CUSUM (K=2.25/ H=4.6 4.8)	**Standardized CUSUM** (k=1.75/ h=4)	**EWMA** (λ=4 L=0.8)	**Shewhart** (L=4)	**Current visual inspection system (Distortion rate)**
(α)	5.88%	5.88%	**4.41%**	3.20%	4.41%
(1-β)	95.60%	98.00%	**98.00%**	95.55%	90.8%
Time (Sec.)	2.1005	2.1324	**1.9724**	1.4642	1.1006

5 Conclusions

This study proposes a spatial domain approach based on small-shift control schemes to inspect reflected distortion defects on curved car mirrors. To quantify the deformation of a car mirror, a standard checkerboard pattern is designed to reflect the pattern on a testing car mirror for image acquisition. The reflected pattern image of a defective mirror with distortion is compared with that of a normal mirror for quantifying the deformation and locating the distortion defects by small-shift control schemes. Experimental results show that the proposed EWMA control scheme achieves a high 98.00% probability of correctly discriminating distortion defects and a low 4.41% probability of erroneously detecting normal images as defective ones on curved car mirror images. The further research is to extend the proposed method to judge the severity levels of the surface distortion defects (e.g. very serious, serious, moderately serious, minor, etc.) and apply the proposed method to inspect transparent glass with different surface distortion defects.

6 Acknowledgment

This work was partially supported by the National Science Council (NSC) of Taiwan, under Grant No. MOST 103-2221-E-324-036.

7 References

[1]. H. D. Lin, D. C. Ho, "Detection of pinhole defects on chips and wafers using DCT enhancement in computer vision systems," *International Journal of Advanced Manufacturing Technology*, 34(5-6), 567-583 (2007).

[2]. Rafael Vilar, Juan Zapata , Ramo´n Ruiz, "An automatic system of classification of weld defects in radiographic images", *NDT&E International*, 42, 467-476 (2009).

[3]. W. K. Wonga, C. W. M. Yuen , D. D. Fan , L. K. Chan , E. H. K. Fung, "Stitching defect detection and classification using wavelet transform and BP neural network", *Expert Systems with Applications*, 36, 3845-3856 (2009).

[4]. Y. C. Chiou, "Intelligent segmentation method for real-time defect inspection system," *Computers in Industry*, 646-658 (2010).

[5]. D. B. Perng, C. C. Chou, S. M. Lee, "Design and development of a new machine vision wire bonding inspection system", *International Journal of Advanced Manufacturing Technology*, 34, 323-334 (2006).

[6]. W. C. Li, D. M. Tsai, "Wavelet-based defect detection in solar wafer images with inhomogeneous texture", *Pattern Recognition*, 45, 742-756 (2012).

[7]. F. Adamo, F. Attivissimo, A. Di. Nisio, M. Savino, "A low-cost inspection system for online defects assessment in satin glass," *Measurement*, 42, 1304-1311 (2009).

[8]. H. Liu, Y. Wang, and F. Duan, "Glass bottle inspector based on machine vision," *International Journal of Computer Systems Science and Engineering*, **3(3)**, 162-167 (2008).

[9]. H. D. Lin, H. H. Tsai, "Automated quality inspection of surface defects on touch panels," *Journal of the Chinese Institute of Industrial Engineers*, **29(5)**, 291-302 (2012).

[10]. Y. P. Chiu, H. D. Lin, "An innovative blemish detection system for curved LED lenses," *Expert Systems with Applications*, 40(2), 471-479 (2013).

[11]. M. L. Duan, K. X. Wu, "New method of correcting barrel distortion on lattice model," *Journal of Computer Applications*, 1113-1115 (2012).

[12]. C. Zhang, J. P. Helferty, G. McLennan, W. E. Higgins, "Nonlinear distortion correction in endoscopic video images", *IEEE ICIP-2000*, 34, 439-442 (2000).

[13]. H. T. Ngo, V. K. Asari, "A pipelined architecture for real-time correction of barrel distortion in wide-angle camera images", *IEEE Transactions on Circuits and Systems for Video Technology*, 15, 436-444 (2005).

[14]. L. N. Smith, M. L. Smith, "Automatic machine vision calibration using statistical and neural network methods", *Image and Vision Computing*, 23, 887-899 (2005).

[15]. E. S. Page, "Cumulative sum charts," *Technometrics*, 3, 1-9 (1961).

[16]. A. F. Bissell, "CUSUM techniques for quality control," *Applies Statistics*, 18, 1-30 (1969).

[17]. D. C. Montgomery, *Introduction to Statistical Quality Control* (6e), John Wiley & Sons, Inc. (2009).

[18]. F. J. Yu, Y. Y. Yang, M. J. Wang, Z. Wu, "Using EWMA control schemes for monitoring wafer quality in negative", *Microelectronics Reliability*, 51, 400-405 (2011).

[19]. D. C. Montgomery, G. C. Runger, *Applied Statistics and Probability for Engineers*, 2nd ed., New York: John Wiley & Sons, 296-304 (1999).

[20]. R. C. Gonzalez, R. E. Woods, *Digital Image Processing* (3/e), Pearson Education, Upper Saddle River, New Jersey (2008).

Continuous RBM Based Deep Neural Network for Wind Speed Forecasting in Hong Kong

Yanxing Hu
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: csyhu@comp.polyu.edu.hk

Jane You
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: csyjia@comp.polyu.edu.hk

James N.K. Liu
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: jame.liu@polyu.edu.hk

Pak Wai Chan
Hong Kong Observatory
134A Nathan Road, Kowloon, Hong
Email: pwchan@hko.gov.hk

***Abstract*—The wind speed forecasting in Hong Kong is more difficult than in other places in the same latitude for two reasons: the great affect from the urbanization of Hong Kong in the long term, and the very high wind speeds brought by the tropical cyclones. Therefore, prediction model with higher learning ability is in need for the wind speed forecast in Hong Kong. In this paper, we try to employ the Deep Neural Network (DNN) to solve the time series problem of wind speed forecasting in Hong Kong since it is believed that Neural Network (NN) with deep architectures can provide higher learning ability than shallow NN model. Especially, in our paper, we use the continuous Restricted Boltzmann Machine (CRBM) to build the network architecture of the DNN. The CRBM is the continuous valued version of the classical binary valued Restricted Boltzmann Machine (RBM). Compared with the Stacked Auto-Encoder (SAE) model applied in our previous study, this CRBM model is more generative, and therefore more suitable for simulating the data in wind speed domain.**

In our research, we employ the DNN to process the massive wind speed data involving millions of hourly records provided by The Hong Kong Observatory (HKO)[1]. The results show that the applied approach is able to provide a better features space for computational models in wind speed data domain, and this approach is also a new potential tool for the feature fusion of continuous valued time series problems.

***Keywords*—Deep Neural Network, Continuous Restricted Boltzmann Machine, Wind Speed Forecasting, Feature Representation**

I. Introduction

Wind speed forecasting has great significance not only in atmospheric related area but also in every aspect of people's life [1]. e.g., in the wind energy industry the forecasting of wind speed can guide the selection of the site position [2]; Engineers frequently utilize information based on wind speed forecasts in the design and construction of large wind-resistantstructures such as bridges, high-rise buildings, and off-shore oil platforms [3]; even in financial markets, wind speed forecasts also play a critical role as weather derivatives and the need to manage weather-related risks, including wind risk and grows[4]. Therefore, the academical and practical value of efficient wind speed forecast approach is obvious.

Currently, there are mainly two families of approaches employed on wind speed forecasting problem: using the numerical models and using the Computational Intelligent(CI) models. Different from numerical models that are too dependent on the psychical restrictive conditions[5], the advantage of using the CI models is that the CI models can "learn" the disciplines from the historical information itself in a statistical manner. One of the mainstream ideas of using the CI models on forecasting is to apply the Neural Networks (NNs) to deal with the given time series data. NNs can recognize the hidden patterns or relationships from the historical observations, meanwhile, additional advantages of the NN approach over the numerical models include data error tolerance, ease of adaptability to online measurements,etc. [6].

On the other hand, in the very recent years, theories about NNs and learning systems have experienced a fast development. More specifically, the applications of DNN or Deep Learning (DL) make breakthroughs in many difference areas [7]. DNN represents a series of multi-layer architecture NNs that training with the greedy layer-wise unsupervised pre-training algorithms[8], [9]. Albeit controversial, this family of NNs have won great success in some fields including Computer Vision, Speech Recognition, Natural Linguistic Programming and Bioinformation Processing. By applying the greedy layer-wise unsupervised pre-training mechanism, DNN can reconstruct the raw data set, in other words, DNN can "learn" features from the original data with a learning system mechanism instead of selecting features manually that we did traditionally[10]. And the intelligent models, like classifiers or regressors usually can obtain higher accuracy and better generalization with the learned features.

As its name suggested, DNN is a kind of NNs that structured by multiple layers. The word "Deep" indicates that such NN contains more layers than the "shallow" ones,

[1] http://www.hko.gov.hk/contente.htm

which mainly includes the most widely used three-layer (single hidden layer) Feed Forward NNs in the past 30 years. Actually, multi-layer NN is not a new conception, some earlier studies have been conducted since 1990s [11], [12], but the successful implementation of multi-layer NNs was not realized until the provision of the novel training mechanism by Hinton in 2006 that a so-called Layer-wise unsupervised Pre-training mechanism is employed to solve the training difficulties efficiently [8]. Via the Layer-wise unsupervised Pre-training mechanism, a DNN represents the raw data set projected from the original feature space into a learned feature space layer by layer in the training process. In each layer, the unsupervised training may provide a kind of regularization to the data set and minimize the variance .

Although theoretically, a shallow NN with three layers trained with Back-Propagation(BP) training algorithm has been proved that can approximate any nonlinear functions with arbitrary precision [13], once the number of hidden neurons is limited, the learning ability of a shallow NN may not be enough and poor generalization may be expected when using an insufficiently deep architecture for representing some functions. The significance of "deep" is that compared with a simple and shallow model, NN with deep architecture can provide a higher learning ability: functions that can be compactly represented by a deep architecture might be required to handle an exponential number of computational elements (parameters) to be represented by a deep architecture. More precisely, functions that can be compactly represented by a depth k architecture might require an exponential number of computational elements to be represented in a depth $k-1$ architecture [9]. Therefore, by adding the number of layers in the network architecture, DNN can provide higher learning ability with less hidden neurons in each layer, this advantage may be more useful for the big data cases. In general, compared with shallow NNs, the DNN model can learn from the massive raw data and map the raw data into a new feature space, classifiers or regressors thus may have chances to obtain higher accuracy and better generalization.

The main work of this paper is an extension of our previous work published in WORLDCOMP'14 last year [14]. In this work, we are continually exploring the potential of DNN in time series problems, especially in weather forecasting domain. In previous research, we noted that for time series problem, a good representation of original feature space may be helpful for the applied model to get better performance [15]. Meanwhile, in time series problems, the correlations among features are obvious but not easy to be identified. If we can analyze the correlations and have the features represented, the prediction accuracy is expected to be improved, and the DNN is a reasonable and suitable tool to analyze the time series features. Moreover, in [14], we have shown that the Stacked Auto Encoder(SAE) [16] can provide positive results on our weather data sets. However, the SAE is considered as a discriminative approach, and for time serise problem, we hope to use a more generative method to build the DNN in order to get the prior knowledge from the data sequence in the model training process. Therefore, in this paper, we applied the Continuous Restricted Boltzmann Machine (CRBM) to build the DNN to deal with the wind speed forecasting problem in Hong Kong. In detail, in the experiment, the CRBM model based DNN is employed to predict the wind speed in the next few hours. The massive data involving millions of weather records employed in this study is provided by The Hong Kong Observatory (HKO).

The contribution and significance of our investigation demonstrate that: we give a further investigation to show that NNs with deep architectures can improve the prediction accuracy in weather forecasting domain; moreover, the modified version of the Restricted Boltzmann Machine(RBM), CRBM, is empolyed in our paper to show that the RBM with continuous stochastic units can partly solve the limitation of classical RBM; more importantly, in our work, we focus on the wind speed forecasting of Hong Kong. For wind speed forecasting in Hong Kong case, we have some special challenges: (i) for wind data at Hong Kong, there is urbanization effect over the long term; (ii) there are tropical cyclones in Hong Kong bringing high speed winds, which are difficult to predict [17], the results of our experiment demonstrate that our model can learn the wind speed change trends better than the previous models.

II. The Wind Speed Prediction Problem in Hong Kong

Unlike data sets in other domain, weather data has some particularities. Specifically, there is season-to-season, and year-to-year variability in the trend of weather data. The cycle could be multi-month, multi-season or multi-year, and the main difficulty of investigations on weather data is to capture all the possible cycles. Hong Kong is characterized by a long coastline and numerous islands for such a relatively small territory.The mesoscale weather system of Hong Kong is quite different from other places since it is heavily affected by rainstorms and tropical cyclones[18], moreover, the high building density may also affect the weather condition of Hong Kong. Therefore, finding the disciplines and capturing the possible cycles of wind speed change in Hong Kong is more difficult than other places in sub-tropical regions.

The changes of wind speed may greatly impact Hong Kong people's daily life, for example, the government's plan of wind power generation system is greatly depending on the long-term wind speed prediction [17], or, the short term forecasting may affect the operation of airport and harbor in Kong Kong. Therefore researchers on Hong Kong put great efforts on wind speed forecasting. Many significant investigations including artificial intelligence technologies have been accepted as appropriate means for wind speed forecasting and reported encouraging results since 1980s [19], [20].

Among many different intelligent models, univariate time series regression is the most fundamental and most widely applied one in wind speed forecasting, especially short-term predictions. In this paper, we also concentrate on employing DNN to represent the feature space for univariate time series model. Generally speaking, for a certain variable, the objective of univariate time series regression is to find the relationship between its status in a certain future time point and its status in a series of past time points, and estimate its future status via:

$$v_t = f(v_{t-1}, v_{t-2}, \ldots, v_{t-n}) \tag{1}$$

The function f, can be obtained by employing different intelligence models such as Linear Regression, Generalized

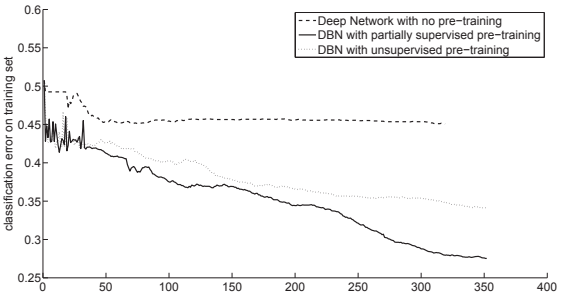


Fig. 1. Training classification error vs training iteration on DNNs, which shows the optimization difficulty for DNNs and the advantage of pre-training methods.

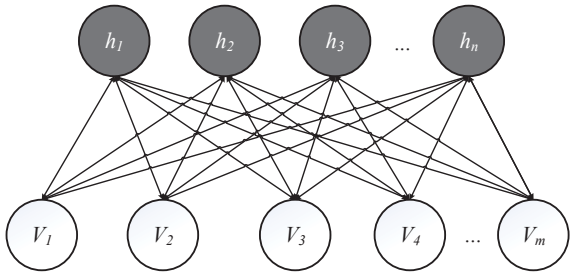


Fig. 2. The typical architecture of a classical RBM model with two layers, m neurons in the visible layer and n neurons in the hidden layer, all neurons a binary-valued and no connection between any two neurons in the same layer.

Linear Model, Auto Regressive Integrated Moving Average Mode (ARIMA), etc.

In our investigation, we target on the wind speed data in the next few hours. We will input the raw data sets into our DNN model, the input n-dimension vector is composed of the status in $(t-1)th$, $(t-2)th$, ..., $(t-n)th$ time points, we try to use the DNN to represent these statuses, and employ a regressor to estimate the status in the tth time point. We hope the seasonal cycles can be captured via massive volume of data by the superior learning ability of DNN.

III. Greedy Layer-wise Unsupervised Pre-training and Layer Model Selection in Deep Learning

Although the idea of Multi-layer(Deep) NN has been proposed for more than twenty years, it wasn't widely used until 2006 since Hinton solved the training difficulties efficiently in [8].

The essential challenge in training deep architectures is to deal with the strong dependencies that exist during training between the parameters across layers [21]. Multi-layer NN has more parameters than NN with shallow architectures. Moreover, in a multi-layer NN, due to the non-convexity of the complex model, the optimization with traditional BP training approach may fall in a local minimum rather than global minimum. This may bring poor generalization to the model.

This problem wasn't well solved until Hinton et al. introduced Deep Belief Network (DBN) that greedily trained up one layer with a Restricted Boltzmann Machine (RBM) at a time in 2006 [8]. Shortly after, strategies for building deep architectures from related variants were proposed by Bengio [22] and Ranzato[23]. They solved the training problem of deep NN in two phases: in the first phase, unsupervised pre-training, all layers are initialized using this layer-wise unsupervised learning signal; in the second phase, fine-tuning, a global training criterion (a prediction error, using labels in the case of a supervised task) is minimized. Such training approach is called the Greedy Layer-wise Unsupervised Pre-training. Fig.1 [21] shows the comparison among different training methods for NNs with deep architectures.

The advantage of learning features from data via a unsupervised approach is that the plentiful unlabeled data can be utilized and that potentially better features than hand-crafted features can be learned. This advantage reduce the need for expertise of the data and often the learned feature space may provide a better regularization effect on the raw data so that can improve the accuracy of the applied model[1], [24]. There are a number of NN architectures categorized into the family of Greedy Layer-wise Unsupervised Pre-training approaches, for example, the Auto Encoder and the Sparse Auto Encoder that obtaining the connection weights of the hidden layer by learning an approximation of the input variables; the RBM, which models the static data via an energy function and the joint distribution for a given visible and the hidden vector; the Convolutional Neural Networks(CNN), which learns the features via a convolutional kernel in each layer, etc. We cannot say which unsupervised pre-training model is definitely better than others since each of the models have its own properties. The choices of model and how the data should be presented to the model are highly dependent on the properties of the data sets [24]. In our previous paper [14], we empolyed the SAE model to do the weather forecasting in the short term. However, compared with SAE, which is a discriminative model, the RBM model is a generative model. A generative model means that it can generate observable data given a hidden representation and this ability is mostly used for generating synthetic data of future time steps [25], [24]. Thus for time series problems, a generative model based DNN is reasonably expected to be able to provide a better performance than the stacked Auto Encoder.

IV. The Restricted Boltzmann Machine and the continuous Restricted Boltzmann Machine

The RBM is a two-layer networking with one visible layer and one hidden layer. Fig.2 gives an illustration of RBM architecture. As shown in Fig.2, the standard type of RBM has binary-valued (Boolean/Bernoulli) m hidden and n visible neurons, and consists of a matrix of weights $W = (w_{i,j})$ (size $m \times n$) associated with the connection between hidden neurons h_j and visible neuron v_i, as well as bias weights (offsets) a_i for the visible units and b_j for the hidden units. The word "restricted" means that there is no connection between any two neurons in the same layer.

Given these, the energy function of a configuration (pair of boolean vectors) (v, h) is defined as:

$$E(v,h) = -\sum_i a_i v_i - \sum_j a_j v_j - \sum_i \sum_j v_i w_{ij} h_j \qquad (2)$$

Since the neurons is binary-valued, the probabilities of the states of the visible and hidden neurons can be obtained via the sigmoid function:

$$p_{vi} = p(vi = 1) = \frac{1}{1 + exp(-\sum_i w_{ij} h_j)} \tag{3}$$

and

$$p_{hj} = p(hj = 1) = \frac{1}{1 + exp(-\sum_i w_{ij} v_i)} \tag{4}$$

respectively.

In RBM, the probability distributions over hidden and/or visible vectors are defined in terms of the energy function in Eq.(1):

$$P(v) = \frac{1}{Z} \sum_h e^{E(v,h)} \tag{5}$$

The RBMs are trained to maximize the product of probabilities assigned to some training set V:

$$\arg\max_W \prod_{v \in V} P(v) \tag{6}$$

In the traning process of the RBM model, the Minimising Contrastive Divergence (MCD) training rule for an RBM replaces the computationally expensive relaxation search of the Boltzmann Machine [26] with a single step of Gibbs sampling [27]. In each iteration, we update the w_{ij} according MCD rule by:

$$\Delta w_{ij} = \varepsilon(v \cdot h^T - \hat{v} \cdot \hat{h}^T) \tag{7}$$

where $\hat{v}$,$\hat{h}$ is the reconstructed states of the node in the last iteration.

As discussed above, we choose the layer component of DNN according to the type of the data sets and the property of the model. From the brief description of the RBM, we can see that the neurons in this model are binary value, this is why we chose Auto-Encoder rather than RBM approach in our previous work to forecast the wind speed data that was continuous-valued. However, according to [24], the RBM is a more generative model than the Auto Encoder, that means, from the aspetacts of model properties, the RBM model maybe a better choice in wind speed forecasting application. Therefore, in this paper, for using the RBM to process the continuous valued wind speed data, a CRBM is employed to build the DNN architectures.

The CRBM is introduced by H. Chen and A.F. Murray in [28]. The continuous stochastic neurons are employed to take the places of the binary-value neurons by adding a zero-mean Gaussian noise to the input of a sampled sigmoid neuron. The binary-value neurons in CRBM have the form:

$$s_j = \varphi_j \cdot (\sum_i w_{ij} s_i + \sigma \cdot N_j(0,1)) \tag{8}$$

with

$$\varphi_j(x_j) = \theta_L + (\theta_H - \theta_L) \cdot \frac{1}{1 + exp(a_j x_j)} \tag{9}$$

where $N_j(O,1)$ represents a Gaussian random variable with zero mean and unit variance. The constant σ and $N_j(O,1)$ thus constitute a noise input component $n_j = \sigma \cdot N_j(O,1)$ according to a probability distribution

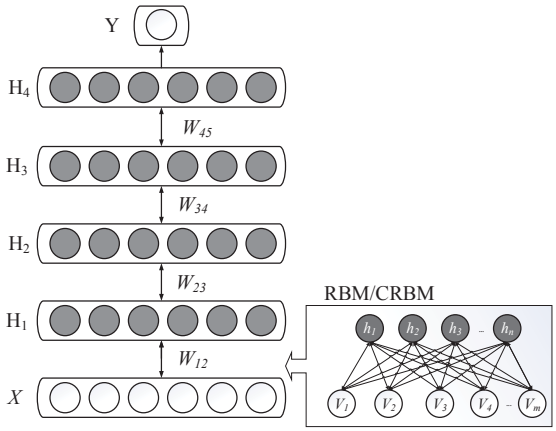


Fig. 3. A 4-hidden-layer DNN with RBM model, by which each layer is greedily pre-trained with an unsupervised RBM model to learn a nonlinear transformation of its input (the output of the previous layer) that captures the main variations in its input by a MCD training methods.

$$p(n_j) = \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-n_j^2}{2\sigma^2}) \tag{10}$$

The parameters θ_L, θ_H and a_j control the asymptotes and slop of the sigmoid function in the neurons. By this way, the nature and extent of the neurons stochastic behavior is simulated [29]. Such behaviour is similar to the noisy units in [30], where the variance of the added noise is tuned [28].

From [28],the energy function of CRBM is analogous to that of the continuous Hopfield model:

$$E_{CRBM} = -\frac{1}{2} \sum_{i \neq j} w_{ij} s_i s_j + \sum_j \frac{1}{a_j} \int_0^{s_j} \varphi^{-1}(s) ds \tag{11}$$

By using the MCD rule, in each iteration, parameters in CRBM model can be updated via:

$$\Delta w_{ij} = \varepsilon_w (s_i \cdot s_j^T - \hat{s_i} \cdot \hat{s_j}^T) \tag{12}$$

and

$$\Delta a_j = \frac{\varepsilon_a}{a_j^2} (s_j \cdot s_j^T - \hat{s_j} \cdot \hat{s_j}^T) \tag{13}$$

Consequently, we need to combine the CRBMs layer by layer with a stacked structure to build the DNN. We follow the method introduced in [8], In each layer, we use a CRBM to train the connection wight in this layer, and then have these layers combined together. Specifically, in the training process of each layer, as shown in Fig.1, the input vectors need to pass through the two layers, meanwhile, the vectors in hidden layers are representations of the input vectors and can be used to reconstruct the input vectors. Thus, in each layer of the DNN, the input of the current layer is the output of the previous layer, then we train the input data via a CRBM, and use the transformed vectors as the output of the current layer. Fig.3 shows the detailed mechanism of CRBM based DNN. We can see that through a DNN, the raw data can be represented into new feature spaces layer by layer, in other words, DNN can learn features from the original data sets. And consequently,

we need to employ a proper regression approach to compute the output with the learned features.

V. Experimental Results and Analysis

In this section, we will describe the experiment and give the results and discussions

A. Wind Speed Data Collection and Pre-processing

The HKO has provided great support to our investigation. Based on our collaboration with HKO, a massive volume of high quality real weather data could be applied in our experiment. The time range of the historical wind speed data sets is almost 30-year long, which covers the period from January, 1, 1983 to December, 31, 2012. The total number of records is more than 230,000. Please note that our data set contains massive records that cover data all the year round of the 30 years in Hong Kong, by this way, we hope the model can catch the urbanization effect change over the long term and learn the rules of the daily, monthly and yearly cycles as well as the seasonal rules of the tropical cyclones in Hong Kong [31].

The wind speed data provided by the HKO has two dimensions: the polar coordinate for the wind direction (measured with degree angle) and the speed (measured with meters per second), moreover, for a certain time points, the direction of the air motion is not stable, i.e. the wind direction at that time point is not fixed. such condition is denoted as "variable" in the raw data. Therefore, according to the requirement of our algorithm, we have to do some pre-processing on the data sets: since the wind speed data (in a fixed horizontal plane) is a vector quantity that has two dimensions in the polar coordinate (as Fig.4), i.e. the angle to show its direction and the speed to measure the magnitude in this direction: the polar coordinate and the speed [32]. However, since our model is focused on single variable time series problems, we have to transform the data set to satisfy the model's requirement. According to the physical significance of the two dimensions, we denote the angle as θ and the speed as v to obtain:

$$v^0 = cos\theta \cdot v \tag{14}$$

where v^0 is the vector components of the wind speed in 0 degree angle direction (as Fig.5). Thus, what we actually simulate is the time series of the speed components of the air motion in 0 degree angle direction. Moreover, there are about 3% of wind speed data with the direction valued as "variable", for such condition, we consider it as a missing value in the data set and use the average value of the wind direction in its previous time point and its next time point to replace the value "variable".

B. Experiment Configuration

In our experiment, the whole data set is divided into two parts, the training set contains the samples of the first 27 years, and the testing set contains the samples of the last 3 years. Thus the ratio of the sizes between the training set and the testing set is 9:1.

To learn the complex effect of the seasonal and yearly cycle of the wind speed change in Hong Kong, we don't input

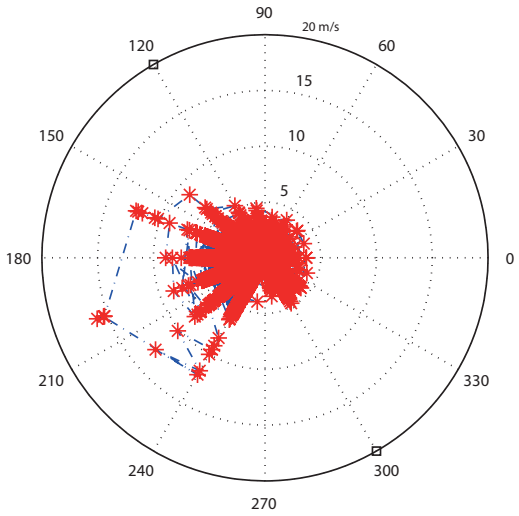


Fig. 4. The distribution of wind speed data in polar coordinate

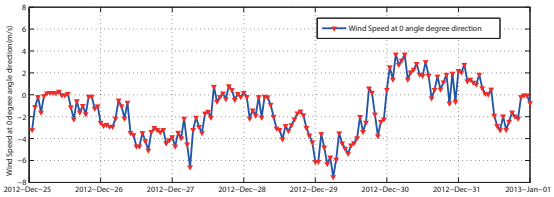


Fig. 5. The distribution of wind speed at a fixed diection

training data to the model randomly as we did in [14] last year. In this investigation, we use shift windows to organize the input model, and there are 7-day data contained in each window. The windows are input into the model according to the time sequence.

In our experiment, we build a four-layer network and employed it to predict the wind speed in Hong Kong. The large size of the data set can avoid the overfitting problem of the complex model. Actually, there is a feature reconstruction of the data sets in each layer, and we hope to obtain a better feature space after 3 feature reconstructions so that the output layer can provide a higher accuracy in the finally obtained feature space. In the top layer, in our model, we choose the Support Vector Regression (SVR) with the Gaussian kernel to give the forecasting output [33]. The parameter configuration of the whole model is given in Table I.

TABLE I. The parameter configuration of the networking

Parameter	Value
Number of neurons in hidden layer 1	168
Number of neurons in hidden layer 2	96
Number of neurons in hidden layer 3	84
Learning rate	0.001
Max Iteration	1000
Parameters in SVR	Default as LibSVM [33]

TABLE II. THE COMPARISON OF WIND SPEED PREDICTION BY THE FOUR MODELS

Model	NMSE	DS	R^2
Single Layer ANN	0.4547	0.694	0.791
CRBM DNN with SVR	0.2213	0.727	0.921
SAE DNN with SVR	0.2395	0.830	0.901
Classical SVR	0.2947	0.741	0.871

C. Experiment Results

In this paper, to evaluate the performance of the CRBM based DNN, other three models are also applied to predict the wind speed in Hong Kong, and the results are compared. Specifically, the four models are the single layer Artificial Neural Networking(ANN), the Classical SVR, the SAE DNN followed with an SVR and the proposed model. From the results comparison, We hope to study the advantages and disadvantages of the SVR and NN models; also, the results will show that whether a feature representation is helpful for improving the accuracy of wind speed prediction; and more importantly; ,the performances of the SVR in feature spaces obtained via the SAEs and via CRBMs are also compared. Table II gives the comparison of the results on three major criteria, and the performance of the four models is respectively shown in Fig 6, Fig 7, Fig 8 and Fig 9.

From the results, we can observe that, all of the four models can catch the main trends of the wind speed change in Hong Kong, but the performances of the four models are not in the same level. The single layer ANN provides the worst results: the single-layer ANN model only has R^2 value less than 0.8; and also provides relatively poorer performance in other two criteria. However, we believe that if we can add more hidden neurons in ANN, the performance will be better, but the computational cost will also be higher. From the results of other three models, we can see that the performance of SVR can be improved (0.03 as the least improvement on R^2 value) by using the DNN feature representation, moreover, compared with the SAE model, DNN with CRBM can provide a 3% improvement of accuracy on weather data prediction. These results demonstrate that as a generative model, CRBM is more suitable than SAE for the time series problem, e.g., wind speed forecasting.

VI. CONCLUSION, LIMITATION AND FUTURE WORK

The wind speed forecasting in Hong Kong is more difficult than that of other places in the same latitude for two reasons: the great affect from the urbanization of Hong Kong in the long term, and the very high speeds of winds brought by the tropical cyclones. In our investigation, we modified the model that applied in our previous paper [14], using the continuous valued RBM model to build the architecture of the DNN instead of the SAE that we used before. The RBM model is more generative than the SAE models and more suitable for time series problem, and we applied the continuous version of the RBM so that the model can be employed to process the wind speed data.

We use massive volume of wind speed data in Hong Kong to test our model. The comparison results are positive: the CRBM based DNN model can learn a better feature space from the raw wind speed data so that the SVR can obtain higher accuracy in this learned feature space. The network can provide lower NMSE by using the CRBM than using the SAE.

The main future work of our investigation is that, we will try to employ the CRBM model on more difficult weather data, such as rain fall data set; and moreover, we will continue exploring the theoretical principle of computational intelligence, especially, we will try to give the mathematical explanation of the DNN.

ACKNOWLEDGMENT

The authors wish to thank the financial support from the Hong Kong Polytechnic University by its Central Research Grants G-YK53 and G-YN39.

REFERENCES

[1] Q. Cao, B. T. Ewing, and M. A. Thompson, "Forecasting wind speed with recurrent neural networks," *European Journal of Operational Research*, vol. 221, no. 1, pp. 148–154, 2012.

[2] T. J. Considine, C. Jablonowski, B. Posner, and C. H. Bishop, "The value of hurricane forecasts to oil and gas producers in the gulf of mexico," *Journal of Applied Meteorology*, vol. 43, no. 9, pp. 1270–1281, 2004.

[3] R. Horonjeff, F. McKelvey, W. Sproule, and S. Young, *Planning and design of airports*. Granite Hill Publishers, 2010.

[4] A. Khanduri and G. Morrow, "Vulnerability of buildings to windstorms and insurance loss estimation," *Journal of wind engineering and industrial aerodynamics*, vol. 91, no. 4, pp. 455–467, 2003.

[5] G. Li and J. Shi, "On comparing three artificial neural networks for wind speed forecasting," *Applied Energy*, vol. 87, no. 7, pp. 2313–2320, 2010.

[6] A. More and M. Deo, "Forecasting wind with neural networks," *Marine structures*, vol. 16, no. 1, pp. 35–49, 2003.

[7] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning." *Journal of Machine Learning Research-Proceedings Track*, vol. 27, pp. 17–36, 2012.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[9] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc., 2009, vol. 2, no. 1.

[10] X. Wang and Q. He, "Enhancing generalization capability of svm classifiers with feature weight adjustment," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2004, pp. 1037–1043.

[11] J. Schmidhuber, "Curious model-building control systems," in *Neural Networks, 1991. 1991 IEEE International Joint Conference on*. IEEE, 1991, pp. 1458–1463.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[14] J. N. Liu, Y. Hu, J. J. You, and P. W. Chan, "Deep neural network based feature representation for weather forecasting."

[15] J. N. Liu and Y. Hu, "Application of feature-weighted support vector regression using grey correlation degree to stock price forecasting," *Neural Computing and Applications*, pp. 1–10, 2013.

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 9999, pp. 3371–3408, 2010.

[17] H. Yang, L. Lu, and J. Burnett, "Weather data and probability analysis of hybrid photovoltaic–wind power generation systems in hong kong," *Renewable Energy*, vol. 28, no. 11, pp. 1813–1824, 2003.

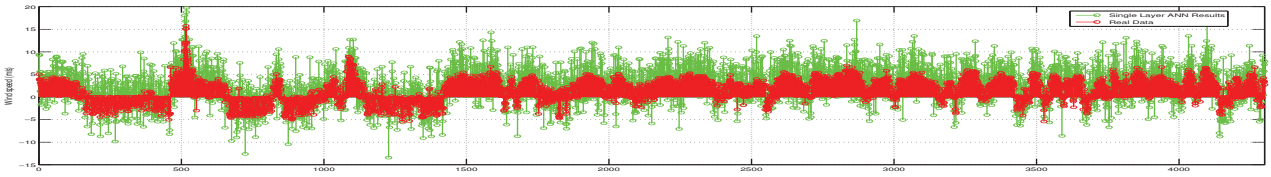


Fig. 6. The prediction results of a singly layer ANN

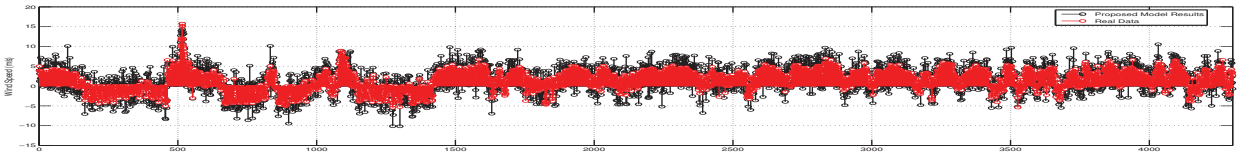


Fig. 7. The prediction results of the proposed model

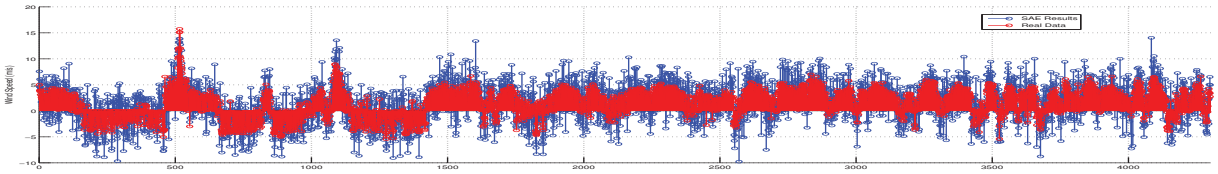


Fig. 8. The predictioin results of the SAE model

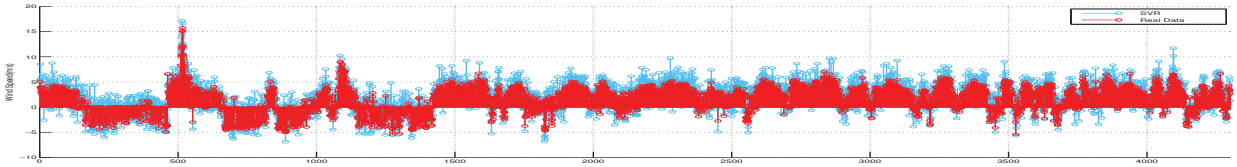


Fig. 9. The prediction results of the SVR model

[18] P. Li and E. S. Lai, "Short-range quantitative precipitation forecasting in hong kong," *Journal of Hydrology*, vol. 288, no. 1, pp. 189–209, 2004.

[19] S.-M. Chen and J.-R. Hwang, "Temperature prediction using fuzzy time series," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 30, no. 2, pp. 263–275, 2000.

[20] K. Kwong, M. H. Wong, J. N. Liu, and P. Chan, "An artificial neural network with chaotic oscillator for wind shear alerting," *Journal of Atmospheric and Oceanic Technology*, vol. 29, no. 10, pp. 1518–1531, 2012.

[21] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[22] S. M. Miller, Y. Geng, R. Z. Zheng, and A. Dewald, "Presentation of complex medical information: Interaction between concept maps and spatial ability on deep learning," *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, vol. 2, no. 1, pp. 42–53, 2012.

[23] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. Lecun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[24] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[25] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in *Advances in Neural Information Processing Systems*, 2013, pp. 899–907.

[26] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in boltzmann machines," *MIT Press, Cambridge, Mass*, vol. 1, pp. 282–317, 1986.

[27] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[28] H. Chen and A. F. Murray, "Continuous restricted boltzmann machine with an implementable training algorithm," in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 150, no. 3. IET, 2003, pp. 153–158.

[29] H. Chen and A. Murray, "A continuous restricted boltzmann machine with a hardware-amenable learning algorithm," in *Artificial Neural NetworksICANN 2002*. Springer, 2002, pp. 358–363.

[30] B. J. Frey, "Continuous sigmoidal belief networks trained using slice sampling," *Advances in Neural Information Processing Systems*, pp. 452–458, 1997.

[31] J. C. Chan, J.-e. Shi, and C.-m. Lam, "Seasonal forecasting of tropical cyclone activity over thewestern north pacific and the south china sea," *Weather and forecasting*, vol. 13, no. 4, pp. 997–1004, 1998.

[32] R. A. Pielke, *Mesoscale meteorological modeling*. Academic press, 2002.

[33] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

Context Aware Exemplar-Based Image Inpainting using Adaptive Image Division

A. Ms. Jainab Bano[1], B. Durga Toshniwal[2]
[1]Computer Science and Engineering Department, Indian Institute of Technology, Roorkee, Uttarakhand, India
[2]Computer Science and Engineering Department, Indian Institute of Technology, Roorkee, Uttarakhand, India

Abstract -*Image inpainting is the process to reconstruct the missing or corrupted regions of the images. This paper presents context aware exemplar based image inpainting using adaptive image division technique. In previous inpainting techniques, complete image is searched for filling the missing region. In the proposed method, adaptive image division technique is used to reduce the search space. A new priority term is also introduced for defining the order of searching for best-first patch. Experimental results on benchmark images demonstrate the improvement and significant acceleration of related exemplar based method.*

Keywords: Inpainting, exemplar-based, context-aware, adaptive image division.

1 Introduction

Image inpainting is a process in which missing or corrupted regions of image i.e. target region are filled by using already filled information in the image. The output of this process is a complete image that is visually plausible to human eyes. The region that need to be filled could be synthetic i.e. user could remove a region and make it as target region to be filled or it could be natural such as scratch on image or text that need to remove from an image. Inpainting algorithms have various applications in image processing such as restoration of old images, photograph editing, to reconstruct missing blocks in image encoding and transmission etc. There are two broad categories of inpainting algorithms that have been proposed in literature: diffusion based methods and exemplar based methods.

Diffusion based methods [1], [14], [15] fill target region by propagating the surrounding information of target region into it. Partial differential equations are used to complete structures such as lines and contours hitting the target region. These methods give good results when target regions are thin and long. But they introduce blur when replicating texture inside the target region. This is because they only use neighboring information of the target region to fill the region.

Exemplar based methods [2-9], [11], [17], [18] inpaint the target region patch-by-patch. The basic idea is to consider a target patch in each iteration and search for most similar patch to the target patch in undamaged part of the image. Then copy the unknown pixel values into target patch from the most similar patch found earlier. These methods have some similarity with patch based texture synthesis but they also focus on structure propagation inside the target region by using various techniques such as: by using specific filling order [2], or by human guided filling [10] or by separating the structure and texture components of the image [11].

Exemplar based methods can be categorized into "greedy" [2-5], multiple candidate [9], [17, 18] and global [6-8], [11]. The "greedy" methods choose only one best match for each patch to be filled therefore they do not ensure global coherence. Multiple candidate methods fill the missing region using weighted average [9] or a sparse combination [17] of multiple candidate patches at each location. Global methods [6-8] model global image context with a Markov random field (MRF). Global methods ensure the global coherence but they have large number of labels to search. T. Ruzič et.al [5-7] proposed context-aware global MRF-based inpainting method that uses contextual descriptors to guide and improvethe inpainting process. This technique gives very impressive results with improve efficiency. T. Ruzič et.al [6-7] also proposed context-aware patch selection technique by dividing the images into blocks. Two types of image division techniques are used: fixed block division [7] and adaptive image division [6]. Adaptive image division [6] is performed by top-down splitting procedure based on contextual descriptors. T. Ruzič et.al has used adaptive image division technique for global exemplar based method. The same division technique could also be used to improve the efficiency of any exemplar based method.

The proposed technique draws idea form exemplar based inpainting as given by Criminisi [2] and image division technique used in [5]. Additional prioritized search is also used based on contextual information. The purpose of introducing priority term in proposed algorithm is to prioritize searching to most contextual similar block first. Search space for each target patch is reduced because of this division technique and also the search is guided by contextual information of the target patch that improves the quality of inpainted images.

Section 2 describes the problem and motivation behind the work. Section 3 contains flow diagram of proposed algorithm and description of proposed algorithm. Experiments performed on benchmark images and results obtained by proposed algorithm and other algorithm with comparisons are given in Section 4. Conclusion and future work are given in Section 5.

2 Motivation

The problem is to reconstruct a missing region Ω in an image I. Criminisi et at. In 2004 proposed exemplar based inpainting algorithm [2] to solve this problem. The algorithm fills missing region Ω iteratively by taking sample from source region Φ = I-Ω. Texture generates correctly because it fills patch by patch and structure generates because it uses specific filling order that is determined by priority term $P(p_x)= C(p_x)D(p_x)$. Priority is the product of confidence term $C(p_x)$ and data term $D(p_x)$. Confidence term represent the confidence in pixel values of patch ψ_p. Data term $D(p_x)$ reconstruct linear structures crossing the target region Ω. Data term computes using equation 1:

$$D(p) = \frac{|\nabla I_p^{\perp} \cdot n_p|}{\alpha} \qquad (1)$$

In equation 1 n_p is a normal vector and α is normalizing factor. Priority of each patch on the contour of target region is computed and highest priority patch is selected for target patch then complete image is searched for best matched patch in the source region and this matched patch use to fill the unfilled part of the target patch.
In this technique complete image need to be search for each target patch. A technique is needed to reduce the search space based on the content of each target patch to make the algorithm efficient in computation time.

3 Proposed Method

The focus of proposed method is on reduction of search space for exemplar based technique as discussed in Section 2. It divides the image into several parts of different sizes based on their contextual similarity. The search space for each target patch is the union of contextually similar image regions. Reduction in search space reduces the total time taken by the proposed method. Figure 1 shows the flow diagram of proposed algorithm. Following is the step by step description of proposed algorithm:

Contextually adaptive image division

The first step of the proposed algorithm is to divide the image into adaptive size blocks based on their contextual similarity. Contextual desctiptor represent the context of the block. Distance between the contextual descriptors of two blocks measures the similarity between them. This distance is compared with some threshold to take decision whether or not to divide the block further. Following is the detailed descriptions of contextual descriptor choose and the image division technique used:

Contextual descriptor:

Contextual descriptors are the characterization of spatial content and textures within blocks. There are many ways to extract texture feature, such as computing co-occurrence matrices, using local binary patterns, multi-channel filtering etc. In the proposed algorithm, gabor filters are used with additional color features included in them. The idea of these desctiptor is taken from [7]. We have used these desctiptor because their simple implementation.

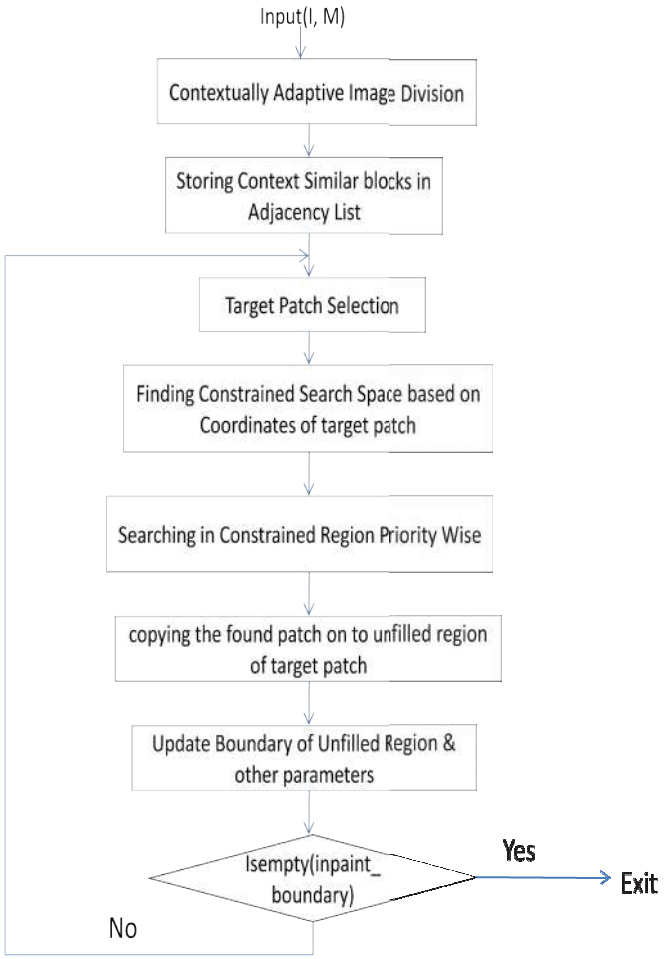


Figure 1: Flow diagram of proposed algorithm

Adaptive Image Division:

Adaptive image division technique given by T. Ruzič et.al in (2015) [6] is used.The process starts by dividing the image into four equal size blocks and all parent blocks assign flag value h. Each block is further divided based on the flag until no further division could possible. Contextual descriptors of horizontal or vertical divisions of the block are computed and contextual dissimilarity is calculated by sum of squared differences (SSD). If SSD is greater than threshold then it is divided in opposite direction and compare with threshold for division. If no division is perform then the algorithm move to next undivided block. The outputs of this technique are the blocks of different sizes with contextually similar content. N is the number of blocks found in this step.

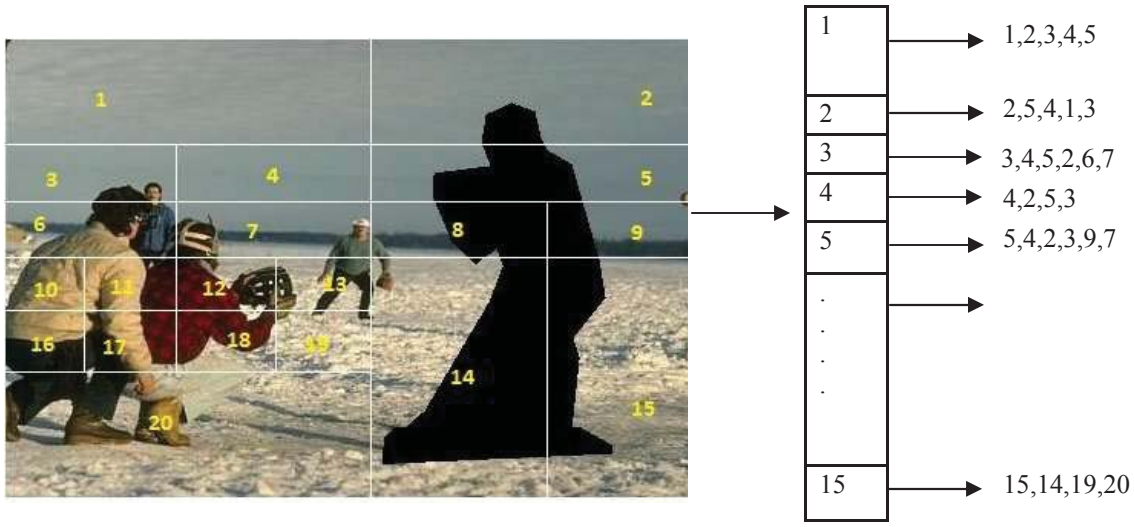


Figure 2: Example of storing context similar blocks in decreasing order of priority

Storing context similar blocks in adjacency list

Adjacency list is used for storing the blocks found in previous step. The list is indexed by block number and its each entryhas a vector that stores the indexes of contextually similar blocks to this block. Indexes in this vector store according to the priority assign based on the contextual similarity of the block as define in equation 1.

$$P_j = \frac{d_{\max}(c_i, c_j)}{d(c_i, c_j)}, \forall i \in \{1, ..., N\} \quad \text{…….. (2)}$$

Here, i is the index of adjacency list and c_i is contextual descriptor of block indexed by i. d is the similarity metric. cj is the contextual descriptor of blocks indexed by j and j varies by indexes stored into adjacency list at index i. For example in Figure 2 a test image and its various blocks stored in adjacency list priority wise are shown:

Target Patch Selection

Target patch selection is performed by computing the priority of each boundary patches given by equation 1 as described in Section 2.1. Highest priority patch is consider as target patch for that iteration.

Finding constrained search space based on coordinates of target patch

In the basic exemplar-based technique [2] search space for each patch is the whole known region of the image, but in the proposed method search space is constrained by using only contextually similar blocks as the search space. For example: In Figure 2 all the blocks that include the end points of the target points that are block no. 2 and 5, and their contextually similar blocks (shown by similar color in Figure 2) that are block no. 1, 4 and 7 are included in the constrain search space of the target patch. New search space $\Phi_c \subseteq \Phi$ for the target patch is subset of the search space used by basic technique.

Searching in constrain region priority wise

This is the main step of proposed algorithm. Contextual information is used to reduce the search and priority given in equation 1 is used to get the most contextually similar patch in case of two similar patches based on known region of the target patch. This improves the visual quality of result. By using the adjacency list searching is done priority wise. The best match patch is found in the searching algorithm. As shown in Figure 2 blocks are searched in decreasing order of their similarity with the target patch.

Copying the found patch on to unfilled region of target patch

The patch that have found in previous step will be pasted in this step on to the target patch.

Update boundary of unfilled region and other parameters

Now, boundary of target region changed. So it will be updated to new boundary i.e. newly filled pixels will be included into source region. Confidence term of newly filled pixels need to update. It will be same as confidence of this target patch as given by equation 2 similar to criminisi method [2]. If target region is filled i.e. contour of target region is zero than terminate the algorithm otherwise repeat step 3 until thiscondition is false. The output of the algorithm is a complete image without that is visually plausible to human eyes.

4 Experiments & Results

It is assumed that shadows are part of the image and the focus of our problem is to only inpaint the user specified target region. Shadow removal is not part of this problem. Proposed method has tested on different natural images. The parameters used in the algorithm are: Number of orientation and scales of gabor filters are six and three respectively. Patch size and threshold were varied and the optimal ones were chosen. Figure 3 shows the results obtained by proposed algorithm and of original exemplar-based technique [2]. Table 1 shows the computation time taken by both the algorithm in increasing order.

In Figure 3 the output obtained by criminisi [2] method has visual artifacts in the inpainted result but in the output of proposed algorithm it is not there. The output is visually consistent and execution time is also less by 1223.3sec by proposed algorithm as shown in Table 1. Results on other images that give same results in proposed algorithm and method [2] are shown in Figure 3. For all the images same patch size 9*9 is used and 0.15 threshold value is used. The difference between proposed method and technique given by Criminisi in [2] is that proposed method takes less computation time as compare to exemplar-based method because search space is less for each target patch in this method. Graph 1 show that there is significant time difference in both the algorithms.

Figure 3: (a) Origional Image, (b) Mask Image, (c) Inpainted image by Criminisi method[2], (d) Inpainted image by proposed method. Snowbaseball Image (image courtesy Le Meur [8])

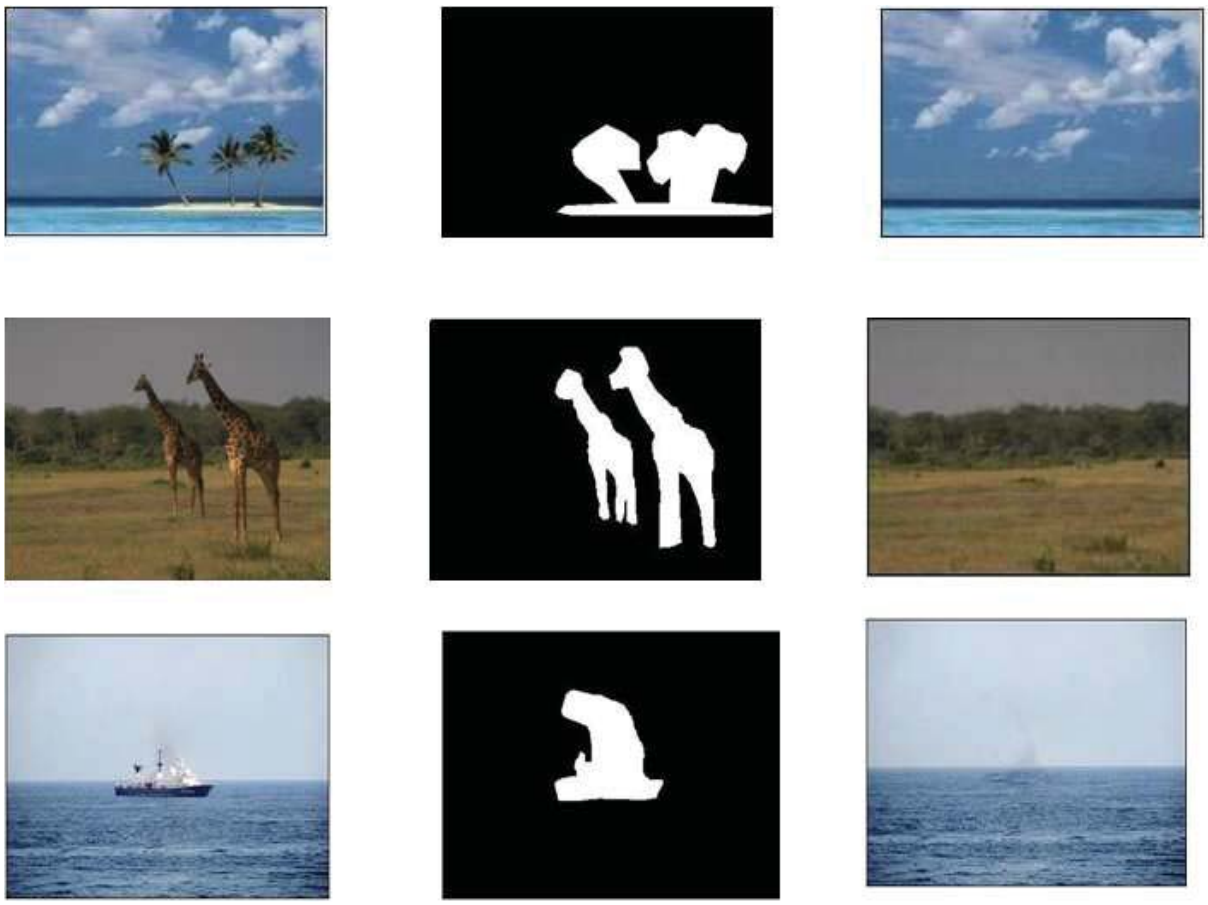
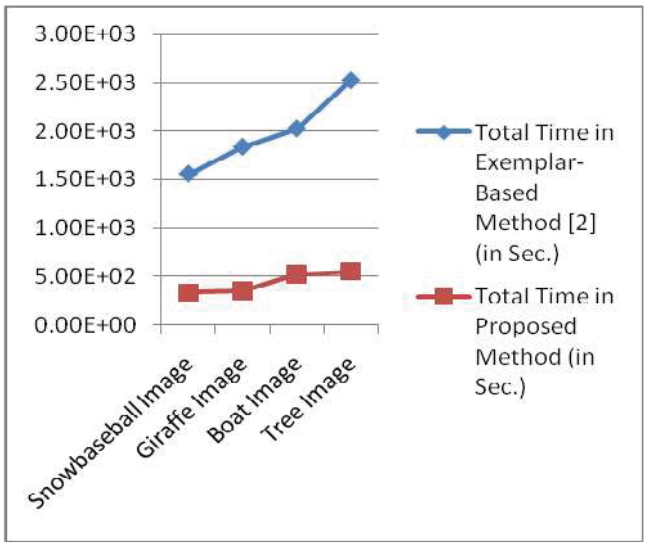


Figure 4: Inpainting results. From left to right: original image, missing region in white color, results of proposed method.(Image reference: top to bottom: tree image[2], giraffe image[1], boat image[4]))

Input Image	Time Taken By Criminisi Algorithm[1](in sec)	Time Taken By Proposed Algorithm(in sec)
Snowbaseball Image	1.5568e+03	333.4936
Giraffe Image	1.8361e+03	350.5825
Boat image	2.0268e+03	521.7715
Tree Image	2.5265e+03	550.6847

Table 1: comparison of computation time between exemplar-based technique given by criminisi[2] and proposed algorithm



Graph 1: Total computation time taken by exemplar-based method [2] and proposed method on different images as given in Table 1.

5 Conclusion

The basic motivation behind the proposed method is to exploit the contextual information to reduce the search space and hence computational time of related exemplar based inpainting technique. The limitation of exemplar based method as proposed by Criminisi et.al is that its search step takes too much time because it searches in the complete image for each patch that needs to be filled. Instead of complete image, searching should be guided by the context of the patch to be filled. This is achieved in the proposed method by dividing the image into contextually similar regions and using only the similar regions to search for best-match patch. Therefore total computation time taken by proposed method is reduced and contextually guided and priorities search gives the visually improved results as compare to state of art exemplar-based method. Experimental results have shown that efficiency in time of proposed method increases with respect to image size as compared to exemplar-based method.

Future work consists of parallelization of the proposed method CAE for searching in different blocks as well as within same block. This will further accelerate the searching and hence the inpainting process, resulting in more fast inpainting algorithm. Other future work would be to use improved data term in proposed algorithm for improved results. Different searching techniques such as *kd-tree* etc. could be used to improve the searching time.

6 Acknowledgement

We thank Olivier Le Meur for providing the benchmark images for test.

7 References

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. "Image inpainting";SIGGRAPH '00, New Orleans, USA, pp. 417–424, 2000.

[2] A. Criminisi, P. Perez, and K. Toyama. "Region filling and object removal by exemplar-based image inpainting"; IEEE Trans. on Image Proc., vol. 13, no. 9, pp. 1200–1212, Sept. 2004.

[3] T. Ruzi˘ c, B. Cornelis, L. Plati´ sa, A. Pi˘ zurica, A. Dooms, W. Philips, M. Martens, M. De Mey, and I. Daubechies, "Virtual restoration of the Ghent altarpiece using crack detection and inpainting,"; ACIVS '11, pp. 417–428, 2011.

[4] C.-W. Fang and J. Lien. "Rapid image completion system using multiresolution patch-based directional and nondirectional approaches,"; IEEE Trans. on Image Proc., vol. 18, pp. 2769–2779, Dec 2009.

[5] T. Ruzi˘ c and A. Pi´ zurica. "Texture and color descriptors as a tool for context-aware patch-based image inpainting,"; SPIE Electronic Imaging '12, vol. 8295, pp. 82 951,Feb. 2012.

[6] T. Ruzi˘ c, A. Pi´ zurica, and W. Philips. "Markov random field based image inpainting with context-aware label selection," ;ICIP '12, pp. 1733–1736, 2012.

[7] T. Ruzi˘ c and A. Pi´ zurica. "Context-aware patch-based image inpainting using Markov random field modeling"; IEEE Trans. on Image Proc., Jan. 2015.

[8] N. Komodakis and G. Tziritas. "Image completion using efficient belief propagation via priority scheduling and dynamic pruning,"; IEEE Trans. on Image Proc., vol. 16, no. 11, pp. 2649–2661, Nov. 2007.

[9] O. Le Meur, J. Gautier, and C. Guillemot. "Examplar-based inpainting based on local geometry,";ICIP '11, pp. 3462–3465, 2011.

[10] Z. Xu and J. Sun. "Image inpainting by patch propagation using patch sparsity,"; IEEE Trans. on Image Proc., vol. 19, no. 15, pp. 1153–1165, May 2010.

[11] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum. "Image completion with structure propagation,";ACM Trans. On Graph.,vol. 24, pp. 861–868, July 2005.

[12] M. Bertalmio, L. A. Vese, G. Sapiro, and S. Osher. "Simultaneous structure and texture image inpainting*,";* IEEE Trans. on Image Proc., vol. 12, no. 8, pp. 882–889, Aug 2003.

[13]He, Kaiming, and Jian Sun. "Statistics of patch offsets for image completion."; Computer Vision- ECCV 2012.

[14] T. Chan and J. Shen. "Mathematical models for local deterministic in-paintings";Technical Report CAM TR 00-11, UCLA, March 2000.

[15] T. Chan and J. Shen. "Non-texture in-painting by curvature-driven diffusions (cdd)"; Technical Report CAM TR 00-35, UCLA, March 2000.

[16] David Tschumperl and Richard Deriche."Vector-valued image regularization with pde's : A common framework for different applications"; IEEE Transactionson Pattern Analysis and Machine Intelligence, 27(4):506–517, 2005.

[17] Z. Xu and J. Sun. "Image inpainting by patch propagation using patch sparsity,";IEEE Trans. on Image Proc., vol. 19, no. 15, pp. 1153–1165, May 2010.

[18] A. Wong and J. Orchard. "A nonlocal-means approach to exemplar-based inpainting,"; ICIP '08, pp. 2600–2603, 2008.

[19] Y. Wexler, E. Shechtman, and M. Irani. "Space-time completion of video";IEEE Trans. on Pattern Anal. and Mach. Intel., vol. 29, no. 3, pp. 463–476, Mar. 2007.

[20] Jiaya Jia and Chi keung Tang. "Inference of segmented color and texture description by tensor voting"; IEEE Transactions Pattern Analysis and Machine Intelligence (PAMI), 26(6):771–786, June 2004.

[21] Yining Deng and b. s. Manjunath. "Unsupervised segmentation of color-texture regions in images and video";IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), 23(8):800–810, 2001.

[22] G. Medioni, Mi-Suen Lee and Chi-Keung Tang. "A Computational FrameworkFor Segmentation And Grouping"; Elsevier, 2000.

[23] M. Oliviera, B. Bowen, R. McKenna, and Y.-S. Chang."Fast digital image inpainting"; Proc. Of Intl. Conf. on Visualization, Imaging and Image Processing (VIIP), page 261-266,2001.

[24] A. Telea. "An image in-painting technique based on the fast marching method"; Journal of Graphics Tools, 9, 2004.

[26] Dong wookcho and Tien D. Bui. "Image In painting Using Wavelet-Based Inter and Intra-Scale Depedency"; IEEE Transactions on Image Processing, 2008.

SPARSELY SAMPLED MRI IMAGE RECONSTRUCTION ALGORITHM MEETS LIPSCHITZ BOUNDS

Krzysztof Malczewski

Department of Electronics and Telecommunications
Poznan University of Technology
Poznan, Poland
kmal@et.put.poznan.pl

ABSTRACT

In the paper a new method for reducing data acquisition time of Magnetic Resonance Imaging (MRI) is presented. This is a SPARSE-SENCE type technique, i.e. it combines parallel imaging and Compressive Sensing (CS) to achieve this aim. Unfortunately, CS methods suffers from local noise, highly limited image resolution, and are extremely sensitive to various types of motion, which usually decreases temporal sparsity and may lead to a temporal blurring of the reconstructed images. The main novelty of the algorithm presented in this paper is a highly efficient method of intestinal motion corrected high-resolution image reconstruction. It is shown that indeed, the approach leads to important improvement of image quality when motion artifacts affect the acquisition.

Index Terms— MRI, compressed sensing.

1. INTRODUCTION

In regular Magnetic Resonance Imaging (MRI) capture of any type of image (T1, FLAIR, DWI, etc) lasts at least 30 minutes. That is why any method leading to shortening of MRI acquisition times is welcomed, as it diminishes patient discomfort, and reduces costs by increasing patient throughput. In previous studies acceleration of MRI imaging (spiral, radial, balanced steady-state free precession, view sharing, PROPELLER) has been done mainly by parallel imaging techniques [4]. It usually leads to acceleration rates 2 to 3. A relatively new group of solutions to this problem are based on application of Compressed Sensing (CS) [1] to MRI [2]: Initially research conducted by Tao suggested that only low acceleration in retrospectively undersampled carotid PC cine MRI data, using temporal fast Fourier transform (FFT) as the sparsifying transform, is achievable. Then, experiment conducted by Velikina resulted in quite high acceleration using second temporal difference as the sparsifying transform. Recently, a combination of Compressed Sensing (CS) and parallel imaging, i.e. k-t SPARSE-SENSE has been presented. Unfortunately, all these approaches are highly sensitive to respiratory motion, which decreases temporal sparsity and implies temporal blurring in the reconstructed images. Because of slowly changing state of magnetization MRI seems to be a perfect candidate for applying CS, section 2. Temporal variation in signal seems to be limited to blood vessel regions, and after applying an appropriate orthogonal transform the resulting image data are sparse. Unfortunately, additional independent motion of organs and other tissues is a common phenomenon during MR imaging. Internal organs movement non-rigid estimation is a particularly challenging task due to the presence of many objects, influenced by the movement of adjacent structures, and is leading to independent deformations. Other important subject of interest in compressed sensing area is its robustness to measurement noise for signals that are normally recoverable without it. It has been proven in [7] that for each vector x that is estimated by $\ell 1$ minimization it exists a Lipschitz bound relating the $\ell 1$-reconstruction error to the measurement error (or noise) for a given sensing matrix.

In this paper the goal is to reconstruct a High Resolution (HR) MR image, from a CS sequence. The majority of previous SR papers have considered only global, relative displacements between set of low-resolution images [4]. The super-resolution method for images containing tissues motion is described.

This paper also describes a method to accelerate PC cine MRI using a combination of k-t SPARSE [2] and SENSE that utilizes joint sparsity among all component coil datasets (k-t SPARSE-SENSE), section 3. Motion analysis and removal is done by local affine adaptive regularization based approach, section 4. The algorithm preserves deformation field discontinuities caused by independent motion of organs and other tissues. The algorithm fits local affine transformation to each voxel, identifies the discontinuities be-

tween the locally fitted affine transformations and adaptively smoothes out the warp field, while preserving its local affine discontinuities [5]. The regularization algorithm is shown in section 5. Presented in section 6 results of experiments show that the new approach indeed significantly reduces problems with motion blur while providing high-quality results characteristic of CS methods.

2. COMPRESSED SENSING IN MRI AND ITS PRACTICAL DIFFICULTIES

Compressed Sensing model was first described in the literature of information theory and approximation as an abstract mathematical idea [1]. The essence of this technique consists in measuring a small number of random linear combinations of signal values–much smaller than the number of signal samples nominally representing it. What is crucial, it should be possible to reconstruct the signal with sufficient accuracy from these simplified measurements (named sparse representation of the signal) by a nonlinear procedure. The nonlinear reconstruction is done under the assumption of both image representation sparsity and consistency with acquired data. In MRI this model reduces to a special case of CS, where sampled linear combinations are directly Fourier coefficients (collecting k-space samples). Then, CS is claimed to be able to reconstruct MRI image from a small subset of k-space rather than an entire k-space grid. The two fundamental factors affecting high performance in CS MRI do the k-space undersampling and image sparsity cause incoherent aliasing artifacts. Though random sampling is an inspiring and instructive idea, sampling a truly random subset of k-space is generally impractical, as sampling trajectories must satisfy hardware and physiological constraints. Hence, sampling trajectories must follow smooth lines and curves. Furthermore, assumed uniform random distribution of samples in spatial frequency does not take into account energy distribution of MR images in k-space, which is far from uniform. Most energy in MRI is concentrated close to the center of k-space and rapidly decays towards edges. Therefore, CS patterns in MRI should have variable density sampling with denser sampling near the k-space center.

To reduce the ghosting artifacts associated with free breathing and bowel peristalsis motion correction should be done. The methods may work with any kind of regularizer. Nevertheless, tweaking up associated with them convergence parameters is an obstacle in their use. Additionally, majorize-minimize algorithms based on a single Lipschitz constant have been reported to be slow in shift-variant applications such as SENSE-type MR image reconstruction, as the associated Lipschitz constants are loose bounds for the shift-variant behavior. So, it is needed to fill the gap between the Lipschitz constant and the shift-variant aspects of SENSE-type MR imaging [6]. Mathematically, the SENSE MRI can be formulated as a $\ell 1$-regularized optimization problem [3]. As the $\ell 1$ term is no differentiable, such problems are hard to minimize using standard gradient methods. Some approaches turn the problem into a different area where fast minimization techniques can be applied. Variable splitting algorithms that constrain optimization problem and then the solution to the original $\ell 1$-regularized one is using Lagrangian formalism [6]. The methods struggle with defining constraint penalty parameter that affects convergence speed. Alternative scenario is the use of majorize-minimize methods such as fast iterative soft thresholding (FISTA) [7]. These algorithms used to converge at a rate that depends on the Lipschitz constant. The constant forms the upper bound for eigenvalues of the Hessian of data fit term. Its value is close to the maximum of squared absolute values sum of the MRI coils sensitivity. The Lipschitz constant turned out to be very loose for low signal regions that occur at the center of the object in SENSE MRI.

The looseness of the Lipschitz bound can be interpreted as resulting from spatial variability of tighter bounds based on coil sensitivity profiles. The method involves finding a diagonal majorizer in the range of the regularizing matrix. Authors of [6] have shown that for several regularizers of interest the diagonal upper bounds are easy to compute and give large accelerations if compared to FISTA with the Lipschitz constant.

3. MINIMIZATION ALGORITHM

Theoretically, a sparse signal x can be optimally approximated by minimizing a convex cost function with $\ell 1$ regularization [7]. The $\ell 1$-minimization procedure for parallel MR image reconstruction can be expressed as follows:

$$\hat{x} = \underset{x \in M}{\operatorname{argmin}} \left\{ f(x) + \beta R(x) \right\} \tag{1}$$

$$f(x) = \frac{1}{2} \left\| y - Ax \right\|_2^2, \qquad R(x) = \left\| R(x) \right\|_1,$$

$A = FS$, where $F \in C^{D \times CN}$ is a block-diagonal matrix with each block having the same down-sampled DFT operator and $S \in C^{CN \times N}$ is a block-column matrix with diagonal blocks, C denotes the number of sensitivity coils, D is the number of data points, and N denotes the number of pixels to be estimated. The mask M eliminates elements outside it. $f(x)$ is the data fit term and $R(x)$ the regularizer. As $A = FS$, we see that A has shift-variant nature, a property that is mentioned in the algorithm [7]. R refers to a sparsifying transform. The β is chosen to balance trade-off between data fit term and the regularizer. Although solving (1) allows one to obtain high-quality estimates of x with less data, (1) is typically difficult to minimize. Instead most methods minimize a different cost function related to (1). The function should be easy to minimize, but still offering information relevant to the solution of (1). Two approaches to defining and minimizing related problems are majorize-minimize

procedures and variable splitting methods. Also "corner rounding" has been proposed for dealing with the non-differentiability of the ℓ_1 regularizer [7], but this has been found to yield algorithms slower than those from variable splitting class. The method applied here [7] is of the majorize-minimize class, nevertheless it differs from previous ones as appropriate coupling of *A* and *R* structures is considered.

4. MOTION ARTIFACTS CORRECTION

Objective of this part of algorithm is to develop a respiratory motion correction method for joint compressed sensing and parallel imaging acceleration of MRI. In the conducted experiment fully sampled low-resolution coil sensitivity reference data has been acquired. The image reconstruction was accomplished in two-steps using the k-t SPARSE-SENSE algorithm [2] with temporal FFT as sparsifying transform, see Figure below.

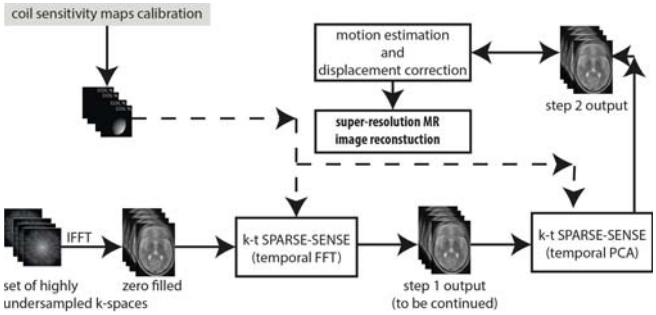


Figure 1. Presented algorithm flowchart.

In the algorithm a local-affine adaptive smoothing approach for the regularization in the demons algorithm [6] has been blend with the compressed sensing framework. This approach models the dense deformation as a set of local affine transformations, and adaptively smooth out the dense deformation field while preserving the discontinuities along the local affine components using the anisotropic smoothing approach [5]. We are assuming that voxels that are close, both spatially and in their intensity value, represent the same structure, and thus have similar affine motion. Hence, the local-affine modeling process relates an affine transform to each voxel, by analyzing its local neighborhood voxels, weighted by their intensity similarity. The coupling of efficient dense deformation estimation [6] with local affine adaptive smoothing yields a better registration algorithm, which is more suitable for the registration of images, affected my internal motion. The idea of this algorithm is given below.

We are given a patient image I_p. The goal is to find a dense deformation field K_p^r that minimizes its dissimilarity to the reference image I_r. Thirion's demons algorithm [5] computes the deformation field that minimizes the energy:

$$\hat{K}_p^r = \underset{D_p^r}{\operatorname{argmin}} E\left(I_p, I_r, K_p^r\right) + S\left(K_p^r\right)$$

where $E\left(I_p, I_r, K_p^r\right)$ represents the dissimilarity measure between the reference and deformed images, and $S\left(K_p^r\right)$ is a regularization term that regulates the smoothness of the folow-on deformation field. The solution is being found by applying the following two successive steps iteratively:

1. Compute an unconstrained dense deformation field that minimizes the dissimilarity between the reference image and other ones.
2. Regularize the deformation field by homogeneous isotropic Gaussian smoothing to keep its spatial coherence. Because the smoothing step may over-smooth the deformation field discontinuities that are associated with independent movements of different organs, we replace the smoothing step 2 with a new anisotropic smoothing filter that is inversely proportional to the differences between the local affine transformations. The smoothing step consists of the following components:

a) Apply an affine transformation $A_f\left(\vec{x}\right)$ to each voxel $\vec{x}$ based on its neighborhood $\Omega_{\vec{x}}$.

b) Calculate gradient of the affine transformations domain.

c) Apply adaptive smoothing [6] to the deformation field, based on the affine transformations gradient.

The next step is the image alignment procedure, where the motion compensated frames are combined to produce one blurred HR frame by using the L1-norm. Using a regularization-based optimization method is deblurring the so generated HR frame.

5. REGULARIZATION ALGORITHM

In this chapter the cost function optimization method is presented. In details, the goal of this part is to minimize the error between the simulated LR frames and the interpolated observed LR frames. The applied cost function incorporates the consequences of assumed local motion, see figure 1.

The super-resolution algorithms are computationally complex and numerically ill posed problems [8]. The problem of SRR can be expressed as follows:

$$W\left(x\right) = \sum_{k=1}^{N} \left\| HF^k X - \hat{Y}^k \right\|_1 + \lambda \left\| CX \right\|_1 \qquad (2)$$

Where the norm given above is the L1 norm and it describes the cost function measuring error. The $\hat{Y}^k$ means upsampled image frame from the observed sequence Y_k:

$$\hat{Y}^k = Up\left(D^T Y^k\right). \quad (3)$$

Where *Up* is the upscaling operator and *C* is the Laplacian operator , λ is the regularization operator. It is helpful in the ill-posedness.

The super-resolution algorithm may be divided into a couple of steps. The final part is coherent motion image parts fusion. It can be expressed in the following way:

$$\vec{X} = \underset{x}{\operatorname{argmin}} \left\| HX - \hat{Z} \right\|_1 + \lambda \left\| CX \right\|_1 \quad (4)$$

where $\vec{X}$ is current *X* high resolution image estimate, $\hat{Z} = H\hat{X}$. Applying the conjugate gradient procedure does technically the minimization, see figure below.

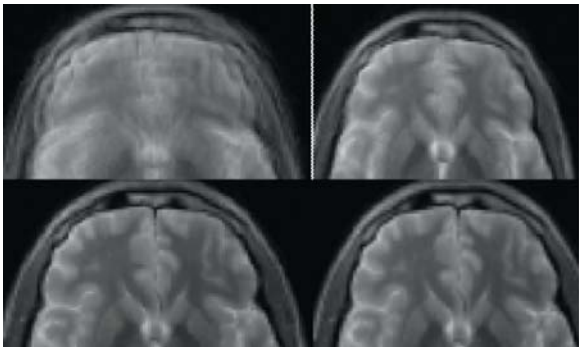


Figure 2. From left to right. Upper row: downsampled and no motion correction applied, downsampled and motion correction applied. Lower row: motion corrected regular sampling scheme (with no downsampling applied), super-resolution CS with motion compensation (proposed algorithm).

6. RESULTS

The experiment has been conducted for two different types of input data. All the initial experiments were conducted on a 1.5T MR Signa Excite scanner sequences. All the CS reconstructions have been implemented in Matlab. Furthermore, two different linear schemes were applied for comparison; zero-filling with density compensation (ZF-w/dc) and reconstruction from a Nyquist sampled low-resolution (LR) acquisition. The LR acquisition has been obtained from centric-ordered data with the same number of data samples as other undersampled sets, see figure 2. The goal of simulation was to examine performance of the compressively sensed super-resolution image reconstruction compared to the LR and Zero Filling with density compensation methods. The further objective was to present superiority of variable density random undersampling over uniform one. Hence, sets of randomly undersampled data with uniform density as well as that with variable density have been constructed from "full" irregularly sampled k-space trajectories. In this test a T2-weighted multislice k-space data of a brain has been analyzed. Figure 2 shows simulation results. While each CS exhibits a decrease in SNR because of the incoherent interference, the uniform density undersampling interference is much more visible and more "structured" than that for variable density. Don't forget that CS leads to acquisition acceleration when compared to the regular k-space sampling pattern. This framework part aims at combining the super-resolution and compressed-sensing in MRI scanners.

The second subsection goal is to illustrate inverse problems of compressed sensing for MRI on phantom data. In particular, the author shows reconstruction examples of the Shepp-Logan phantom from sparse projections, with 25 and 12 radial lines in FFT-domain as welll as reconstruction from limited-angle projections, with a reduced subset of 60 projections within a 90 degrees aperture. Technically semi-PROPELLER k-spaces have been acquiring by compressive-sensing native PROPELLER blades. The low-resolution acquisition has been included in centric-ordered data with the same number of data samples as the undersampled sets, see figures 3 and 5.

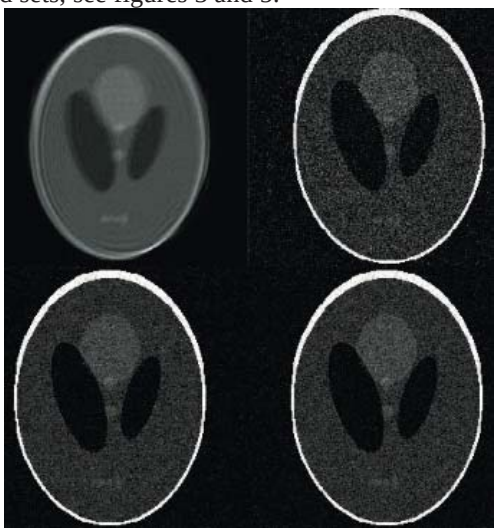


Figure 3. From the upper left to the lower right: the uncorrected image after motion consisting of a modeled shift occurring over the course of the entire acquisition time and reconstruction result and the reconstruction results of the Shepp-Logan phantom. The sampling rates are 25, 40, and 60 percent from left to right, respectively.

Figure 4. The Shepp-Logan phantom results comparison. From the left: the PROPELLER sampling matrix reconstruction output, the proposed algorithm result with enhanced resolution. The lower row exposes detailed images.

7. CONCLUSION

The new SPARSE-SENSING MRI algorithm for super-resolution image reconstruction has been presented. It has been demonstrated that the sparsity of MR images can be exploited to significantly reduce scan time as well improve the resolution of MR imagery. Experiments have shown that MR image resolution has been enhanced keeping motion artifacts at low level. The new technique and more robust segmentation may find applications in many medical applications. Affine motion correction significantly increased sparsity in the temporal Fourier domain, which is due to better alignment among frames. Application of the motion correction decreased temporal blurring, and indeed, presented images quality has been significantly improved. The sparsity in Magnetic Resonance Imaging (MRI) is applied to significantly undersample k-space. Blending Compressed sensing and super-resolution in MRI may become an essential medical imaging tool with an inherently slow data acquisition process. Combining CS, SR in MRI modalities offers potentially significant scanning time reductions, with benefits for patients and health care economical factors

8. REFERENCES

[1] Gribonval, R., and M. Nielsen, "Sparse representation in unions of bases", IEEE Trans. Inf. Theory, vol. 49, no. 12, pp. 3320-3325, (2003).

[2] Wang, Haifeng, "Accelerating MRI Data Acquisition Using Parallel Imaging and Compressed Sensing", Biomedical Engineering and Bioengineering Commons, 2012.

[3] Wainwright, M.J., "Sharp thresholds for high-dimensional and noisy recovery of sparsity", Proceedings of 44th Allerton Conference on Communications, Control and Computing, Monticello, IL, (2006).

[4] Malczewski K., "Breaking the Resolution Limit in Medical Image Modalities," Proceedings of The 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition, World-Comp 2012, USA, (2012).

[5] Freiman, M., Voss, S., Warfield, S.: "Demons registration with local affine adaptive regularization: application to registration of abdominal structures" In: Proceedings of the 8th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2011. pp. 1219–1222 (2011)

[6] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems", SIAM J. on Imaging Sciences (2009)

[7] Muckley M., Fessler J., "Fast Parallel MR Image Reconstruction via B1-based, Adaptive Restart, Iterative Soft Thresholding Algorithms (BARISTA)", IEEE Trans Med Imaging 2014 Oct 14. Epub 2014 Oct 14.

[8] Olcott, P.D., Chinn, G., Levin C.S., "Compressed sensing for the multiplexing of PET detectors", Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE.

A Novel Trajectories Classification Approach for different types of ships using a Polynomial Function and ANFIS

M. Elwakdy[1], **M. El-Bendary**[2], **and M. Eltokhy**[3]
[1]Department of Electronic Technology, Helwan University, El-Amyria, Cairo, Egypt
[2]Department of Electronic Technology / Helwan University, El-Amyria, Cairo, Egypt
[3] Department of Electronic Technology / Helwan University, El-Amyria, Cairo, Egypt

Abstract - *In this paper, a Trajectories Classification Algorithm (TCA) is presented. The points of the tanker ship and fishing boat were collected in the same environment. Each trajectory of the tanker ship and fishing boat is partitioned to many segments to extract the features from each trajectory by using the polynomial function. The features extraction is used as input of the subtractive clustering to put the data in a group of clusters. Also it is used as an input of the neural network in ANFIS.*

The features extraction of each trajectory is represented with the membership functions and group of the Fuzzy If-then rules. The Initial Fuzzy Inference System (IFIS) is trained with artificial neural network to get the Final FIS. The performance of the TCA using a polynomial function and ANFIS is evaluated by different trajectories. The proposed TCA is tested using different trajectories obtaining a high classification accuracy 99.5%.

Keywords: Trajectories Classification; polynomial function; Subtractive clustering; ANFIS.

1 Introduction

Pattern recognition is a section of machine learning which aims to classify the trajectories of different kinds of objects. Trajectories' classification is important research topic for predicting the type of moving objects (tanker ships and fishing boats) based on their trajectories and other features [1-3]. Trajectories' classification is one of the hot topics that scientists are interested to work on. This study aims to discriminate between two different types of ships [4]: Tanker ship and fishing boat based on their trajectory data by using ANFIS. In the classification task, we work on predicting with the type of the object (fishing boat or tanker ship) that a trajectory belongs to. There are different features that are extracted from the tanker ship and fishing boat where the fishing boat moves in bending trajectories (no limitation of the movement of the fishing boat) and the tanker ship which moves in specified trajectories (simple trajectories) at sea. This presented innovative method that can be used widely in the future for detecting any abnormal ship movements through remote sensing, which can be an important assistance for security and coastal safety in general.

Adaptive Neuro-Fuzzy Inference System is used for the trajectories' classification. ANFIS is not offered for all the Fuzzy Inference System options [5]. Fuzzy logic was proposed by Lotfi Zadeh [6]. Fuzzy logic concepts are contributed as an effective tool in automatic decision making systems such as pattern classification systems. In late 1980s, Neural Networks (NNs) and Fuzzy Logic (FL) technologies developed as an effective tool in many fields [7-11].

ANFIS is divided into two main groups: hybrid of neural network and fuzzy logic which combine to enhance the prediction capabilities so ANFIS merge both neural networks and fuzzy logic principles to take advantage of both of them in a single framework [5]. Artificial neural network performance depends on the size of training samples [12]. This means that the data must be prepared well before the training stage by artificial neural network. When the size of training data is small and doesn't represent the possibility space, this preforms that the network results are poor [13].

For preparing the data before the classification stage, each trajectory is divided to many segments. A combination of polynomial function [14] and ANFIS are efficiently extract the features from each segment, and provide these features to ANFIS for the classification purpose for getting high classification accuracy and least average error.

This paper is organized as follows: Section 2 stated the trajectories characterization of the fishing boat and tanker ship. Section 3 stated the features extraction using the polynomial function, subtractive clustering and ANFIS. Section 4 stated training parameters using the ANFIS. The efficiency of the proposed method for classification of the ships trajectories in the proposed TCA through ANFIS is demonstrated in Section 4. In Section 5, the related work is discussed. In Section 6, conclusions are presented. Finally, references are presented in Section 7.

2 Trajectories

The Trajectories of both objects (tanker ship and fishing boat) are represented as a sequence of 2-dimensional points (vectors or trajectory elements) [3, 4] $<x_1; y_1; t_1>; \ldots; < x_n; y_n; t_n>$ where x_n and y_n represent the position of the object at time t_n. In this paper, the temporal dimension is discarded, and each trajectory of both objects is presented as $T = < x_1; y_1 >; \ldots; < x_n; y_n >$. The sample rate of all trajectories is not constant where the temporal difference between the sequential samples is not always the same. Generally, the trajectories of both objects are different in temporal length, the number of data points and distance traveled. Each trajectory of both objects is partitioned to many line segments [1, 2, 4, 15], and the used real trajectories database [4] is represented two different types of ships in this study. This database is divided to two classes: Tanker ship class and fishing ship class. The total number of trajectories is 12 trajectories of both ships. The trajectories number of the tanker ship (points) is 6 (530), and the trajectories number of the fishing boat (points) is 6 (896).

3 Proposed Trajectories Classification Algorithm (TCA)

TCA consists of three parts that are: features extraction, subtractive clustering and ANFIS. The proposed algorithm is used to carry the goal of this research.

The proposed algorithm is illustrated in Fig.1.

The following subsections are clarified in the above system's block diagram.

3.1 Features Extraction

Polynomial function is used as a common base for extracting the coefficients (discriminating features) [16] from each trajectory of both objects (tanker ship and fishing boat). The features extraction is the stage before the trajectories' classification stage. For extracting the coefficients from each trajectory, the polyfit function is used "polyfit (*x*, *y*, *n*)" in MATLAB where '*x* ' and '*y* ' indicate the number of points of each trajectory of both objects which are represented as a Matrix and '*n*' indicates the degree of the polynomial function [14]. This function is used to return the coefficients for a polynomial *p(x)* of degree n of each trajectory of both objects that is a best fit (in a least-squares sense) for the trajectory data in *y*. A polynomial with a degree *n* is represented as in Eq. (1):

$$p(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1} \quad (1)$$

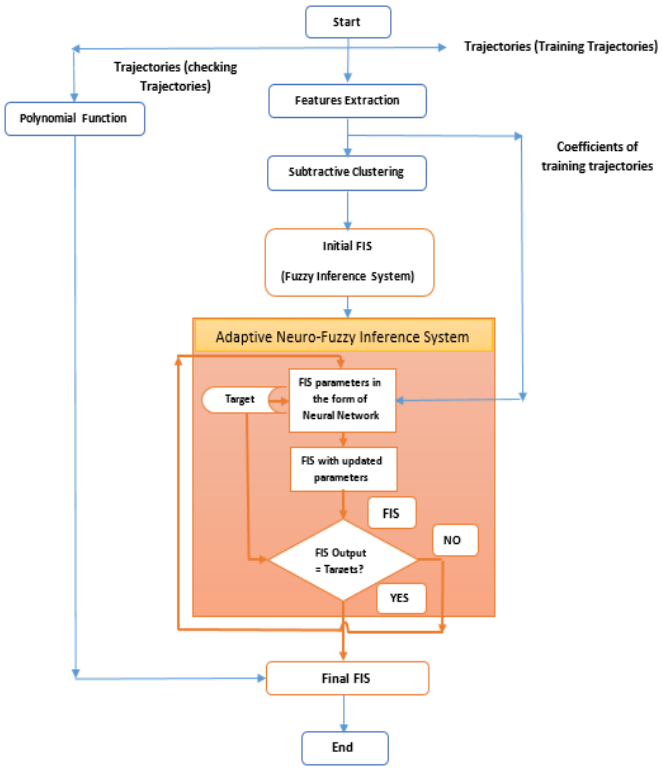


Fig.1. Proposed Algorithms of the Trajectories Classification (TCA) using Features Extraction, Subtractive Clustering and ANFIS.

Where p_1, p_2,…, p_{n+1} indicate the coefficients for the polynomial function. Because the coefficients don't represent the whole trajectories of both objects as shown in Fig.2 and Fig.3, each trajectory is partitioned to equal number of segments for getting a good representation of each segment by using the polynomial function. The polynomial function is used for extracting the same number of coefficients for each segment of all trajectories where each segment consists of the same number of points (three consecutive points) as much as possible as shown in Fig.4 and Fig.5. When the segments number in each trajectory of both objects is increased, the coefficients can be represented all segments of all trajectories of both objects well. This helps in getting real classification accuracy by using an ANFIS. Note that, the number of points of each trajectory depends on the type of object as we mentioned before.

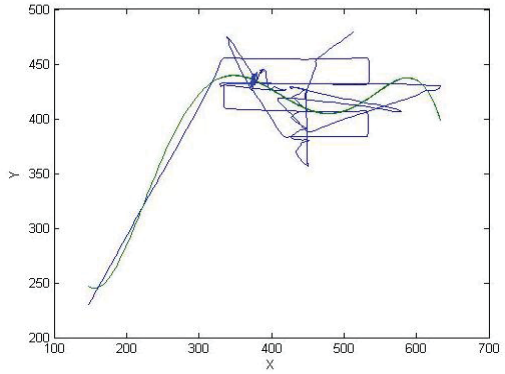


Fig.2. Blue line indicates the fishing boat trajectory (bending trajectory) at sea. The Green line depicts to plot Y through the constructed polynomial model evaluated at each X- component values. The generic symbols X and Y indicate the longitude and latitude coordinates (projections of these coordinates to an x; y-plane).

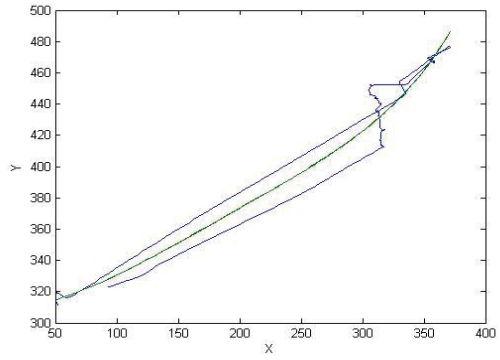


Fig.3. Blue line indicates the tanker ship trajectory (simple trajectory) at sea. The Green line depicts to plot Y through the constructed polynomial model evaluated at each X- component values. The generic symbols X and Y indicate the longitude and latitude coordinates (projections of these coordinates to an x; y-plane).

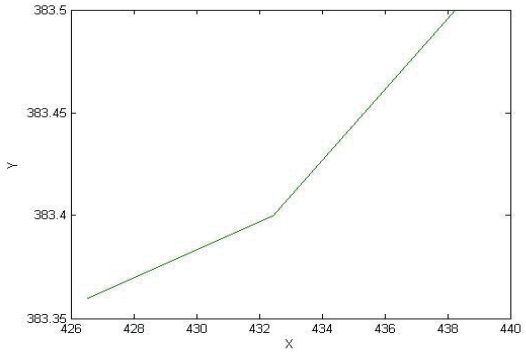


Fig.4. Good representation of a segment is shown by using a polynomial function. The segment of the fishing boat trajectory consists of three consecutive points. The generic symbols X and Y indicate the longitude and latitude coordinates (projections of these coordinates to an x; y-plane).

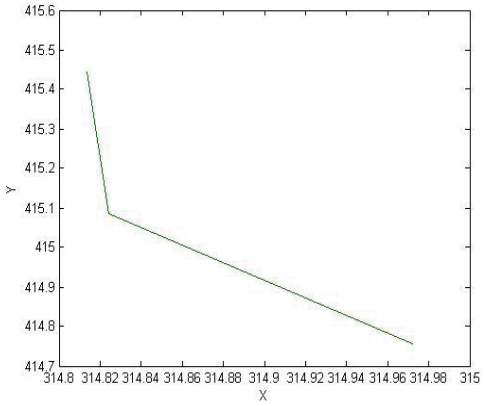


Fig.5. Good representation of a segment is shown by using a polynomial function. The segment of the tanker ship trajectory consists of three consecutive points. The generic symbols X and Y indicate the longitude and latitude coordinates (projections of these coordinates to an x; y-plane).

In simple and bending trajectories, when the coefficients number is increased by using the "polynomial function", this will perform to "over-fitting problem" where the coefficients don't represent the ships' trajectories. For that, the increase of segments number helps in getting a very good representation of the sub-trajectories (all segments) of both objects, and overcomes on the "over-fitting problem". There are a lot of experiments that are performed to specify the degree of the polynomial function (the degree of a polynomial is '4').

The dataset contains all coefficients of all segments of all trajectories (both objects). The preparation of data is very important to get high classification accuracy and least average error so the coefficients extracted from all segments of all trajectories of both objects are partitioned to equal number of groups where the total number of groups is 22. The number of coefficients in each group depends on the length of the trajectories which include different number of points based on the type of object. The 'standard deviation', 'variance', max function' and 'bsxfun function' in MATLAB [17, 18] are used for all coefficients of each group to reduce the size of dataset or to prepare the dataset well before the next stage (Subtractive Clustering) as shown in Fig. 6. This helped on reducing the number of rules which preformed to reduce the computational cost and training time [5]. In this study, 50 % of the dataset of each object is used for training by an artificial neural network, and 50 % of the dataset of each object is used for testing.

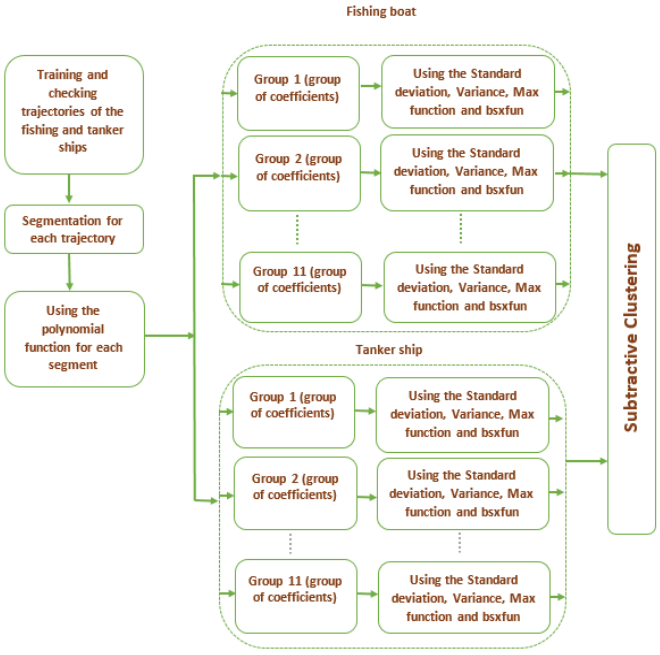


Fig.6. Analysis Framework

3.2 Subtractive Clustering

The training and checking data is uploaded to ANFIS EDITOR GUI of MATLAB. The training data is used in training the Sugeno based ANFIS in next phases. The purpose of the subtractive clustering is estimating the number of clusters and the cluster centers through the training data. Through the subtractive clustering, the training data is partitioned into clusters [19]. The procedure for grouping 22 data point clusters $\{z_1, z_2, z_3, \ldots, z_n = 22\}$ in the training set is described below [20].

1. Calculate the initial potential value for each data point (z_i) as in Eq. (2) [20].

$$P_i = \sum_{j=1}^{n} e^{-\alpha(z_i - z_j)} \qquad (2)$$

$\alpha = 4/r_a^2$ where r_a^2 is a positive constant representing a normalized neighborhood data radius for each cluster. Any point will have a little effect if it drops out outside this encircling region. The point is determined as a first cluster center if it has the highest potential value. This temporarily defines the first cluster center [20].

2. When potential value ($p^{(1)}$) is equal to the maximum of initial potential value ($p^{(1)*}$), the point is qualified as the first center as in Eq. (3) [15].

$$p^{(1)*} = max_i \left(p^{(1)}(z_i) \right) \qquad (3)$$

3. The threshold δ is defined as the decision to continue or stop in search about the cluster center. The search about the cluster center will go on if the current maximum potential remains greater than δ where δ is defined as δ = (reject ratio) × (potential value of the first cluster center) where η is the rejection ratio and $p^{(1)*}$ is the potential value of the first cluster center [20].

4. Remove the previous cluster center from further consideration [20].

5. Revise the potential value of the remaining points according to the Eq. (4) [20].

$$p_i = p_i - p_k^* e^{-\beta(z_i - z_k^*)^2} \qquad (4)$$

Where p_k^* is its potential value, z_k^* is the point of the k_{th} cluster center, and $\beta = 4/r_b^2$ and $r_b^2 > 0$ indicates the radius of the neighborhood for which significant potential reduction will occur [21].

The work on generating the cluster centers is repeated until the maximum potential value in the current situation is equal to less than the threshold δ. We get different cluster center numbers from 22 training samples (patterns) after using the subtractive clustering (depending on the rejection ratio) [20]. The IFIS is generated with number of rules and membership functions at the end of clustering by using the subtractive clustering [19]. The training data is used to get the primary parameters of the membership functions [5]. The center of each cluster is projected to get the centers of the membership functions, and the widths of the membership functions are obtained on the basis of the radius [21].

There are multiple parameters that are used for clustering: Range Of Influence, Squash Factor, Accept Ratio and Reject Ratio. Range of Influence indicates the radius of a cluster when the data space is considered as a unit hypercube [22]. The goal of the Squash Factor is that determines the neighborhood of a cluster center by multiplying the radii values in that factor [22]. The Accept Ratio contributes in determining the cluster center. When the value of Accept Ratio is high, this means that the data points have a very strong potential for being cluster centers [22]. The Reject Ratio contributes too in determining the condition to reject a data point to be a cluster center [20]. The criteria in determining the cluster center depends strongly on the Accept and Reject ratios [20]. Based on these parameters are mentioned before, the IFIS is constructed as shown in Fig.1.

The clustering parameters are used in this study are shown in Table. I

TABLE.I. the clustering parameters are used in this study

Clustering Parameters	Value
Range Of Influence	0.8
Squash Factor	1
Accept Ratio	0.5
Reject Ratio	0.15

3.3 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro- Fuzzy Inference System (ANFIS) is used in many applications [5, 23-25]. In this study, ANFIS is used for a classification purpose where it is used for predicting the type of object: Tanker ship or fishing boat through their trajectories. There are two kinds of fuzzy logic inference system: Mamdani fuzzy logic inference system and Takagi-Sugeno fuzzy logic inference system. The ANFIS has a good advantage where it merges artificial neural network and fuzzy logic system [24]. The Artificial Neural Network works on training the IFIS to access to least possible error between the desired output and FIS output through the dataset to obtain Final FIS as in Fig.1. The schematic of the architecture of ANFIS based on Sugeno fuzzy model is shown in Fig.7.

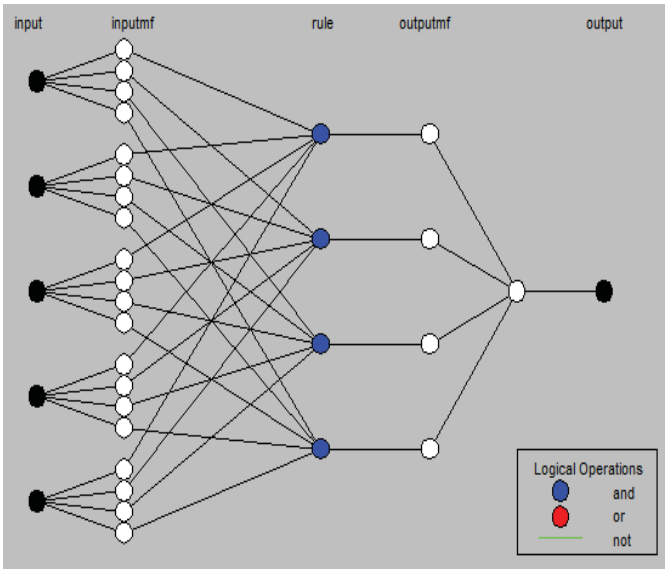
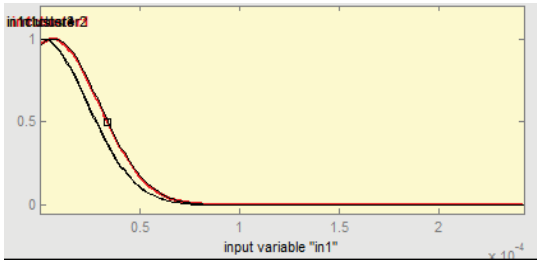


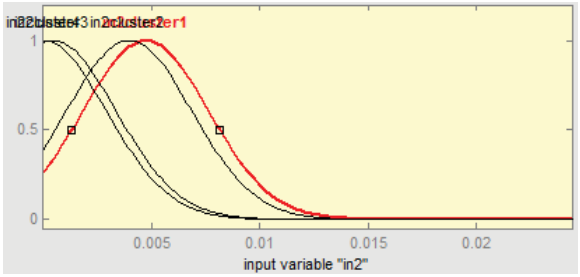
Fig.7 the schematic of ANFIS architecture based on Sugeno fuzzy model.

The ANFIS works on classifying the data based on the features extracted values [26] which represent the trajectories characteristics which apply as inputs of ANFIS. The structure of ANFIS [6] included input parameters, input membership functions, Fuzzy rules, output parameters which represent trajectories, output membership functions, and resultant prediction of trajectories. The hybrid-learning algorithm [27, 28] is employed here where it is used to fit the input and output membership parameters. The Hybrid learning rule is faster than the classical back-propagation method [29]. The Hybrid FIS is trained in 3 Epochs, and the Error tolerance is kept zero for the process. The Hybrid FIS is a combination of back-propagation algorithm and Least Squares [5]. The average error rate depends strongly on the number of membership functions. In addition, the average error rate and classification accuracy depend on the difference between the FIS output curve and checking data curve [5].

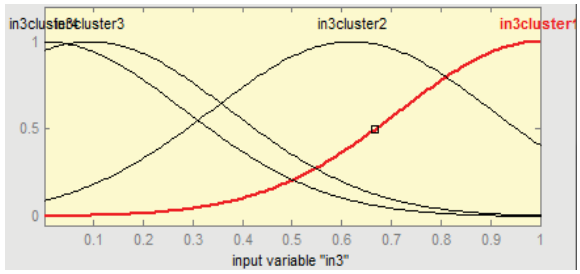
The ANFIS is evaluated with new input data (checking data) where an ANFIS simulates the checking data with the stored data [24]. After the training phase is successfully completed by using an artificial neural network, the Final FIS is tested with the checking data introduced where the training data is converted to Fuzzy data. The fuzzy data has value between '0 to 1' [23]. The ANFIS works on the basis of IF-Then rules [5] where the IF-Then rules change if any change (modify) happened in the membership functions. Based on the IF-Then rules, the checking data (input data) is compared with the trained data (output data). The ANFIS works on training the fuzzy logic system by changing the membership functions [24]. The membership functions represent the training and checking data which are uploaded from MATLAB workspace, and the input membership functions are shown in Fig. 8. The total number of IF-then rules is 4 rules where each IF-Then rule contains six coefficients where five coefficients are multiplied with five inputs in addition to the constant. IF-Then rules are the core of fuzzy logic, and are related varies stages with input parameters [5]. Each IF-Then rule explains some relationships between the input and output variables [25].



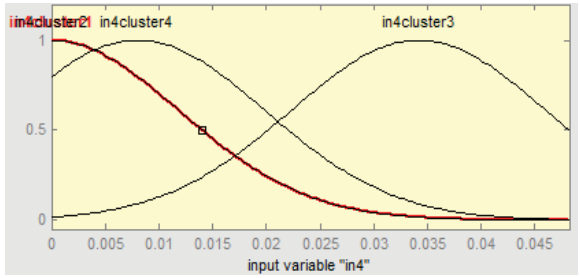
(A)



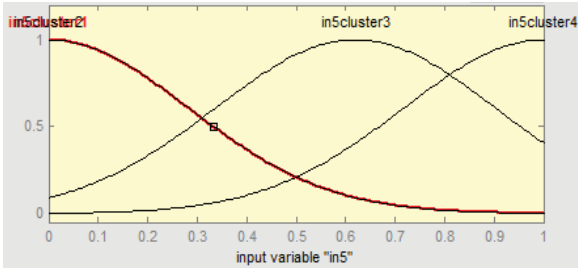
(B)



(C)



(D)



(E)

Fig.8 membership function plots of 5 inputs. (A) A typical initial MF setting, where first input (in 1) range is between 6.147e-012 and 0.0002427, (B) a typical initial MF setting, where second input (in 2) range is between 1.561e-008 and 0.02442, (C) a typical initial MF setting, where third input (in3) range is between 1.487e-005 and 1, (D) a typical initial MF setting, where fourth input (in 4) range is between 0 and 0.0482, (E) a typical initial MF setting, where fifth input (in 5) range is between 0 and 1.

4 Experiment Results

In this study, the training and checking of the proposed algorithm is done by using the ANFIS. There is 12 trajectories of both ships is applied to the proposed algorithm where six trajectories (three trajectories of both ships) is used as a training data and the other six trajectories (three trajectories of both ships) is used as a checking data for the proposed algorithm. The tanker and fishing trajectories are correctly classified by using the ANFIS. The '5' input parameters are used for training and checking aims and work on getting high classification accuracy and least average error where the training and checking data is uploaded to the ANFIS Editor. The input and output parameters on FIS is shown in Fig.9. The Designer of Neuro-Fuzzy shows the checking data appears as plus signs (+), and the training data appears as circles (O) [19]. The Hybrid is specified as an ANFIS model parameter optimization. The number of training epochs is kept to the default value '3', and the training error tolerance is kept to the default value '0'.

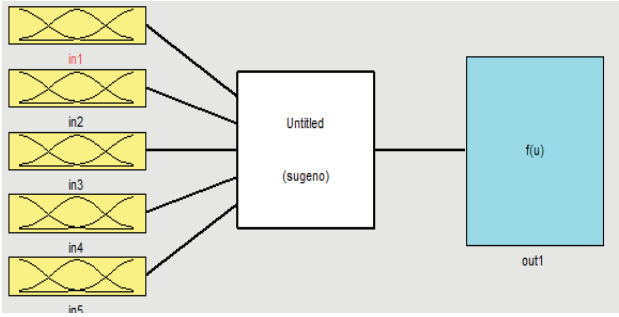


Fig. 9. Input and Output parameters on Fuzzy Inference System

The Sugeno ANFIS is implemented in MATLAB R2011a. 50% of the dataset is used as a training data, and 50% of the dataset is used as a checking data. In other words, the Sugeno ANFIS is tested on 50 % of dataset. After IFIS is generated, it is trained with 4 membership functions by using the Artificial Neural Network. The input and output parameters are presented with group of membership functions. The IFIS is trained many times to get high classification accuracy and least average error. The average testing error is very small which is near to the tolerance limit. 99.5% correct classification is obtained at the ANFIS training with checking data, which indicates (reflects) a good discrimination of the trajectories of both ships. After the IFIS is trained, the trained FIS is tested against the training and checking data to notice the difference in average testing error between both of them.

The output of the FIS appears on the plot as asterisks * * * * * * * *. The plot indicates that there is very less contradiction (conflict) between the training and checking data output and the output of the FIS as shown in Fig 10, 11.

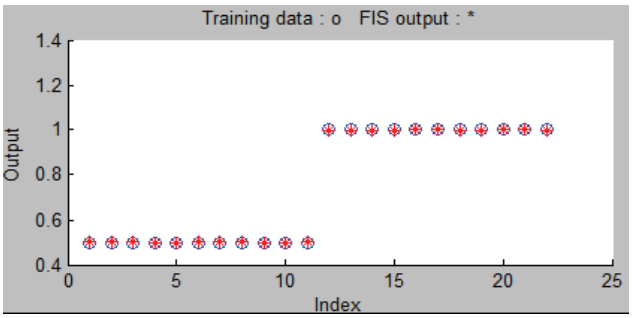


Fig.10 trained FIS against the training data" Trajectories"

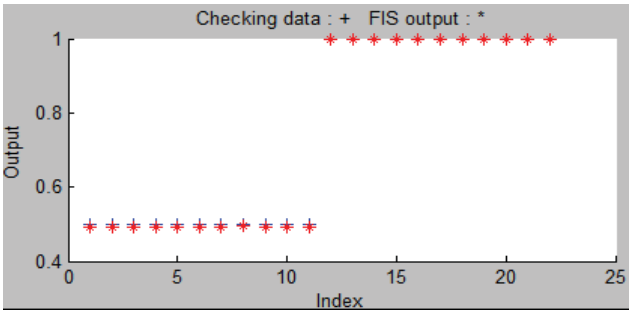


Fig.11. trained FIS against the checking data "Trajectories"

The average testing error is calculated by testing the training and checking data against the trained FIS as shown in Table. II.

TABLE.II. The average testing error of training and checking data

The average testing error	Value
Testing the training data against the trained FIS	1.4426e-5
Testing the checking data against the trained FIS	0.0051065

5 Related Works

In the field of pattern recognition, the trajectories classification is used widely in many applications such as automatic recognition of handwritten postal codes on postal envelopes, signature, handwriting image and automatic speech recognition. In many of proposed methods, the Hidden Markov Model (HMM) classifier is used for the trajectories classification. G. Vries, W. Hage and M. Someren [3] worked on predicting with the type of the vessels through their trajectories by clustering the trajectories into groups of similar movement patterns and used the SVM for the classification purpose. The trajectories haven't been partitioned to many segments before using the SVM classifier. They got classification accuracy 75.4%.

J. Lee, J. Han, X. Li and H. Gonzalez [4] worked on trajectories' classification of two difference types of ships: Tanker ship and fishing boat by partition each trajectory of both ships to number of segments which helped on discriminating the parts of trajectories identifiable and explored two types of clustering: region-based and trajectory-based which worked on finding the features of the sub-trajectories and then used the SVM for the classification purpose. The cooperation between two types of clusters helped on discriminating the features of sub-trajectories of both ships and got high classification accuracy 98.2%.

R. Pelot and Y. Wu [1] worked on boats' trajectories classification of 4 different types of the recreational boats to figure out the difference in the movements characteristics across boat types. The samples of the boats' trajectories were accumulated in two environments. To discriminate between different types of boats, there are 7 variables helped in that: Mean Speed (MS), Max1/20 Speed, Mean Turning Angle (MTA), Total Distance traveled (TD), Aspect Ratio (AR), Coverage Index (CI) and furthest Distance from Shore (DFS), but MS, MTA and DFS are most active variables in discriminating the boat type. The statistical multivariate approach helped in specifying the group to which an object belongs by forming discriminate function for each boat group. These functions, which include three active variables, played important role in specifying the boat types. Based on the forming discriminate functions, the highest classification accuracy for discriminating between boat types is 99.7%.

We worked on extracting the coefficients from each segment of all trajectories of two different types of ships using a polynomial function. This helped in getting a good representation of all segments in all trajectories. The used database [4] contains fishing boat's trajectories (bending trajectories) and tanker ship's trajectories (simple trajectories). The training and checking datasets are prepared well before uploading them to the ANFIS. ANFIS is used for the classification purpose. In this study, we could obtain a high classification accuracy (99.5%) compared with [3] and [4]. The innovated method is a simple method for getting high classification accuracy between two different types of ships compared with previous study as in [4]. The trajectories of the 4 types of recreational boats [1] are simple trajectories compared with the trajectories of fishing boat (bending trajectories) which were used in this study and previous study as in [4].

6 Conclusions

In this paper, a novel method for classifying the trajectories of two different types of objects (fishing boat and tanker ship) by using the Adaptive Neuro-Fuzzy Inference System has been presented. This method enables us to classify different types of objects that have different trajectories at sea. There are lot of experiments, which are made using a real database, for getting high classification accuracy (99.5%) and lowest average error that is near the predefined tolerance limit.

The performance of the trajectories classification system is evaluated through figuring out the average error where the trained FIS is tested against the training and checking data.

The ANFIS is used for the classification purpose. The ANFIS combines artificial neural network and fuzzy logic system which help in getting a very high level of the classification accuracy. Sugeno FIS is implemented to work on classifying the trajectories. The tasks of features extraction and classification were performed using polynomial function, subtractive clustering and ANFIS. The trajectories of the two ships are partitioned to many segments. The polynomial function is used to extract the features from each segment of all trajectories. These features make the system suitable for trajectories classification. The subtractive clustering is an effective tool to take the details of training data and place them in a group of clusters. The subtractive clustering method is used to estimate the number of clusters and cluster centers in a set of data where each point supposed as a potential cluster center. In other words, each data point is defined as a cluster center based on the density of surrounding data and is computed by the subtractive clustering.

The IFIS is generated by the subtractive clustering with minimum number of rules. The Artificial Neural Network works on training the IFIS for getting the Final FIS which perform to increase the ability of ANFIS in trajectories classification efficiently.

This study proves the ability of the trajectories classification system in classifying different trajectories of two different kinds of ships effectively by using the ANFIS.

7 References

[1] R. Pelot, Y. Wu. "Classification of recreational boat types based on trajectory patterns"; ScienceDirect, Vol. 28, Issue 15, pp. 1987–1994, 2007.

[2] J. Lee, J. Han, K. Whang. "Trajectory Clustering: A Partition-and-Group Framework"; Proceedings of the ACM SIGMOD international conference on Management of data, pp. 593-604, 2007.

[3] G. Vries, W. Hage, M. Someren. "Comparing Vessel Trajectories using Geographical Domain Knowledge and Alignments"; Data Mining Workshops (ICDMW), IEEE International Conference, Sydney, NSW, pp. 209-216, 2010.

[4] J. Lee, J. Han, X. Li, H. Gonzalez. "TraClass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering"; Proceedings of the VLDB Endowment, Vol. 1, Issue 1, pp. 1081-1094, 2008.

[5] A. Abbas, A. Mazhar, S. Hassan. "Measuring Weather Prediction Accuracy Using Sugeno Based Adaptive Neuro Fuzzy Inference System, Grid Partitioning and Guassmf"; Computing Technology and Information Management (ICCM), 8th International Conference, pp. 214 – 219, 2012.

[6] A. T. Azar. "Adaptive Neuro-Fuzzy Systems"; Vienna, Austria, 2010.

[7]Pakistan Meteorological Department's web-site: www.met.gov.pk.

[8] Wunderground web-site: www. wunderground.com.

[9] M.S.K. A wan, M.M. Awais. "Predicting weather events using fuzzy rule based system"; Vol. 11, Issue 1, pp. 56–63, 2011.

[10]P. Sallisl, M. Jarur, M. Koppen. "Frost Prediction Characteristics and Classification Using Computational Neural Networks"; (Eds.): ICONIP 2008, Part I, LNCS 5506, pp. 1211–1220, 2009.

[11]W. Myers, S. Linden, G. Wiener. "A data mining approach to soil temperature and moisture prediction"; National Center for Atmospheric Research, Boulder, CO, 2009.

[12]E. Visser, M. Otsuka, T.Lee "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments"; Ingentaconnect, Speech Communication, vol.41, Issue 2, pp. 393-407, 2003.

[13]P. Nava, J. Taylor. "The Optimization of Neural Network Performance through Incorporation of Fuzzy Theory"; Proceedings of the Eleventh International Conference on Systems Engineering, pp. 897-901.

[14]"Polynomial Function," Mathimatics Toolbox, MathWorks, Matlab R2011a.

[15]H. Li, X. Wang. "Automatic Recognition of Ship Types from Infrared Images Using Support Vector Machines"; Computer Science and Software Engineering, International Conference, Wuhan, Hubei, Vol. 6, pp. 483 – 486, 2008.

[16]B. Malakooti, Y. Zhou. "Approximating Polynomial Functions by Feedforward Artificial Neural Networks: Capacity Analysis and Design"; Applied Mathematics and Computation, ScienceDirect, Vol. 90, Issue 1, pp. 27–51, 1998.

[17]"mathematics", Statistics and Random Numbers, Descriptive Statistics, MathWorks, Matlab R2011a.

[18]"Mathematics", Linear Algebra, Matrix Functions, MathWorks, Matlab R2011a.

[19] Fuzzy logic Matlab toolbox, Mathworks, Matlab R2011a.

[20]S. Nakkrasae, P. sophatsathit, W.R. Edwards. "Fuzzy subtractive clustering based indexing approach for software components classification"; Vol. 14D, Issue 1, pp. 89-96, 2007.

[21]M. Eftekhari, S. D. Katebi. "Extracting compact fuzzy rules for nonlinear system modeling using subtractive clustering, GA and unscented filter"; Applied Mathematical Modelling, ScienceDirect, Vol. 32, Issue 12, pp. 2634–2651, 2008.

[22]S. Akbuluta, A.S. Hasiloglub, S. Pamukcu. "Data generation for shear modulus and damping ratio in reinforced sends using adaptive neuro-fuzzy inference system"; ScienceDirect, vol. 24, Issue 11, pp. 805-814, 2004.

[23]R. Singh, W. Bailey. "Fuzzy logic applications to multisensor-multitarget correlation"; IEEE transactions on aerospace and electronic system, , Vol. 33, Issue 3, pp. 1-18, 1997.

[24]M. Shah, P. D. Mehta. “Classification of vehicles using adaptive neuro fuzzy inference system”; Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference, pp. 1-6, 2014.

[25]A. Pandey, N. Gupta. “Stage Determination of Oral Cancer Using Neurofuzzy Inference System”; Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference, pp. 1-5, 2014.

[26]A. Dufaux. “Detection and recognition of impulsive sound signals”; Institute of microtechnology, Switzerland, pp.1-209, 2001.

[27]J. Jang. “ANFIS: adaptive-network-based fuzzy inference systems”. IEEE Trans Syst Man Cybern, 23, pp. 665- 685, 1993.

[28]J. Jang. “Fuzzy modeling using generalized neural networks and Kalman filter algorithm”; Proc. Ninth Nat. Conf. Artificial Intell, PP.762-767. 1991.

[29]E. Avci, Z. Akpolat. “Speech recognition using a wavelet packet adaptive network based fuzzy inference system”; SinceDirect, vol. 31, Issue 3, pp. 495- 503, 2006.

Resolution Enhancement of Single Image by Using Multiple Training Images

Yogesh Choudhary[1], Manish Kashyap[2], Sandeep Kumar[3], Swaraj Singh Pal[4] and Mahua Bhattacharya[5]
Indian Institute Of Information Technology and Management, Gwalior (M.P.) India
yogeshchoudhary135@gmail.com[1], manishkashyap.iiit@gmail.com[2], sandeep2006iiitm@gmail.com[3]
swarajsinghpal@gmail.com[4], mahuabhatta@gmail.com[5]

Abstract--In recent days most digital imaging devices i.e. the high resolution images or videos are playing a critical role in the areas of image processing and application. They too are becoming helpful in the areas of medical diagnosis. This type of images is basically helpful in the pictorial information of human understanding. The algorithm of single image super-resolution is presented which are basically based on spatial and wavelet domain and take the advantage of both. Now in virtual world when we take an object then its resolution of a target image is quite perfect and do not distort when we actually zoom it but this is not the case with the real world image data in which we need to apply some of the techniques by which the image do not get distorted.

Keywords-- Super Resolution, Nearest Neighbour, Low Resolution (LR), High Resolution (HR).

I. Introduction

Super resolution is a technique which is used now a day so often in the imaging technologies so as to get a clear view of the objects around us. Over nearly the past years we have been seeing how the latest technologies have impacted the size of the gadgets we are using, from about just a decimal value of number in pixels now we have reached a number close to 100 megapixels today which is seen in the professional models of the digital cameras.

The ever increasing demand of the number of pixels leads us to get a sharp and a high definition images or high resolution images and this have created a great interest in the areas of the imaging where we deal with the super resolution technique.

Here, our basic goal is to get various multiple 'low' resolution images of the scene and then combined all these several images to get a 'high' resolution image. And as a matter of fact it's a interesting work where we take four images of let say two megapixels and combines to get a single image of eight megapixels. So this is a technique in which we are taking various views of the images which are lacking in the high frequency details of the objects in it, this various images are super convolved to get a image which is a detailed image i.e. a HR image [1]. There are many techniques used in the literature to create a super resolved image:

1) Interpolation technique.
2) Training several images.
3) Reconstructing using smoothing process.

The interpolation are the methods which are basically the fine details of the image reconstruction based methods are applicable in various smoothness priors and imposing the constraint that if taken properly down sampled the HR (High Resolution) image then it should be able to reproduce the original LR (Low Resolution) image. Also in training several image based methods details texture is in delusion by searching across training set of LR/HR patch pairs [1]. This networking point is that here we have to select the training data properly otherwise the unexpected results can also be found.

From the basis of Example based approach, and nearest neighbour embedding (NNE) based methods [4] gives good results by showing using very less number of trained samples. In this context Glasner et. al. proposed a method base on image patches for natural images, which are combining the reconstruction and example based methods.

Now in this paper our approach is that we are trying to provide the real world depthness to a particular image by using the training data bases and hence giving the richness to the image. Which basically learns the fine details from the sample images and fine details will be based to predict the missing richness in the target image. For this purpose we use small patches of learned images for further image processing such as to make probable image information in the other image.

II. Methodology

The methods for increase the images resolution are as follows:

1) Amplifying the already present high frequencies in the image data provided noise should not be amplified.
2) Aggregating the feature from various frames and hence taking out a single fine detail resolution frame in a series of low resolution image.
3) Estimating missing high frequencies which is not present in the original image.

A. Creation of Training Set:

To create training set for the images we need to take the sample from the images and hence to get the patches from the sample images we first degrade them so that they can be done undo as we planned in the later process so that they would lack in the fine details of the image, and then after the process the resolution will be changed by higher factors by using single octave algorithm repetitively [3].

For this purpose we are basically applying the initial cubic interpolation technique to subsample the original which lacks in the fine details of the original image. We are taking the high resolution image patch to the corresponding low resolution image patches, now these images patches are typically are of 5x5 and 7x7 pixels respectively [4]. Now we know that the highest spatial frequency component will be the most important feature for predicting the extra fine details of the image.

B. Approaches:

If we are given the single image and said to find out the missing out details with the original image then the original image will look oatmeal image which does not show any kind of advancement [1]. So this approach does not work. Hence we can say that the local patch themselves are not sufficient to estimate the missing out frequencies.

C. Example Based Approach:

This work is being proposed by Freeman et al. In such approaches we work by maintaining two sets of training patches in which one is high sampled and other contains a low sampled patches, $\{x_i\}^n_{i=1}$ shows the sampled from high image patches. And $\{y_i\}^n_{i=1}$ shows the sampled from the low images[9].

This pair is connected by a behaviour model as $y_i = D*H*x_i+v$,[9] Now this High and Low resolution reoccurrence is hence applied to the target image to get a prediction of the high resolution image which is as modelled as a markov random field as shown in Fig 1.

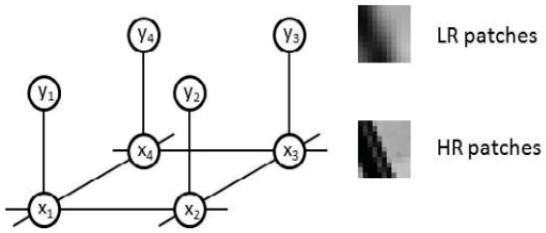


Figure 1: Markov Random Field model for a single image resolution.

The patch size should be chosen with utmost care i.e. if we are taking the co-occurrence is very poor and if we are taking a large patch size then we may need a very large training set to find the proximity patches.

Now this above methods are not showing very good results so here we apply methods on the images directly, requiring large training sets to include in any training sets, Also Chang et al. found a another method of performing the operation i.e. for each low resolution patch y_k^t from the image we find the k-nearest neighbour N_t from the $\{y_i\}^n_{i=1}$ and which computes the reconstruction weights by neighbour embedding [9].

$$\hat{w}_s = \arg\min_{w_s} \|\boldsymbol{y}_k^t - \sum_{\boldsymbol{y}_s \in \mathcal{N}_t} w_s \boldsymbol{y}_s\|^2,$$
$$s.t. \sum_{\boldsymbol{y}_s \in \mathcal{N}_t} w_s = 1. \quad (1)$$

These weights are used to apply to generate the corresponding high resolution patch given by[11][15].

$$\hat{x}_k^t = \sum_{y_s \epsilon N_t} \widehat{w_s} X_s \quad (2)$$

Hence for this purpose we explored two approaches which are [1]:

1) Markov Network.
2) Single pass technique.

Now in the markov network the relation between the low resolution image patch and high resolution image patch is shown which basically gives us the estimate of finding the distance between the two images patches, and the distance

nodes, it actually shows the statistical dependencies between the two [2].

Now in the second algorithm we have used a algorithm which actually do the work of the markov network more effectively and hence gives the same output as the former but with great accuracy. Here only we calculated high resolution image data to corresponding high resolution image patch in a single pass operation.

This operation takes two steps in which the first step includes the enlargement of a particular image so that the extra image details can be added by doubling the number of pixels by using cubic splines or bilinear interpolation and the second step includes the interpolation of a particular image. So here in this image we are actually predicting the missing details so as to create a super resolved output.

D. Overview of the algorithm:

1) Constructing the databases of matching LR-HR patches.
2) Now using Single pass technique discussed earlier we find the most coherent patch assignment to generate a good image.

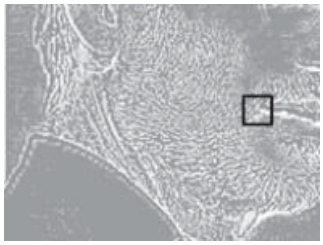
E. Constructing the databases:

We are given the databases of images:

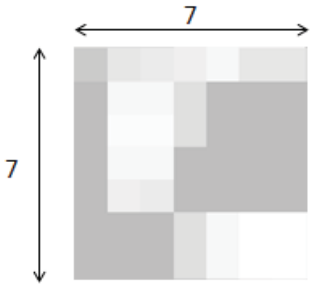
1) Make a sequence of LR patches to corresponding HR patches.
2) Now each image in the database is treated as:

a) Take each 7x7 patch from the image and de-resolute into a 5x5 patch.
b) Now normalize the 5x5 patches to have the same mean and relative constant.
c) Arrange the databases by the low frequency of the low resolution patches.

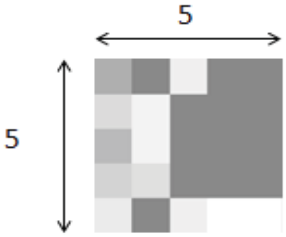
(a)



(b)



(c)



(d)

Figure 2: Fig 2(a) Shows the interpolated image of the original high resolution image. Fig 2(b) Shows the corresponding 7x7 patches. Fig 2(c) Shows the contrast normalized image of Fig 2(a). Fig 2(d) Shows the corresponding patches of 5x5 of the image Fig 2(b).

F.Construction of the Entire High Resolution (HR) image:

This step which is helpful in the reconstruction of the HR image; here we are applying two steps in which we have two steps which are as follows [7].

(1) Local Patch Matching.
(2) Global patch Matching.

Now in the former we are actually matching the LR image patch to the HR image patch from the database which we have previously formed from the low frequencies, now using this we will get the estimation of the unknown high frequencies which is again based on the match. After this we match between neighbouring overlapping patches.

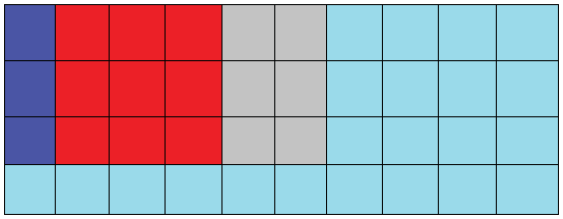


Figure: 3 Neighbouring patches are matched by using the local patch which is overlapped. The later patching scan in the order from left to right and then from top to bottom, and these we will match using the nearest neighbour in the databases which are a additional constraints.

III The Algorithm

The algorithm takes the input patch which is the LR patch and then we find out the compatible patches from the neighbouring high resolution patches which are already found out from the image for this purpose we scan in left to right and from top to bottom of the particular patch. This reconstruction of a patch is helped by using the training data which basically find's the best match and then this concatenated data is again sent to the high frequency image. Now this process is continued iteratively to give the HR image by using the factor β which is a trade-off in matching the low-resolution (LR) patch image data and then finding a high resolution patch which is helpful in adjusting with the neighbours [5].

$$\beta = 0.1 \frac{P^2}{2Q\text{-}1} \qquad (3)$$

Now in the algorithm we initially perform the searching by using the K-d tree algorithm in which the nearest neighbour is found out by plotting the data in the form of the tree, and thus this would help in finding the desired data very fast. The complexity of the K-d tree algorithm is very fast as compared to nearest neighbour which is O(log n). This tree is build by recursively distributing the training set in the direction of greater variation [10]. For e.g. if we training data as (1,9) (2,3) (4,1) (3,7) (5,4) (6,8) (7,2) (8,8) (7,9) (9,6)

Now K-d tree takes the following steps as:

1) Pick Random dimension.
2) Find the median
3) Split data
4) Repeat above steps..

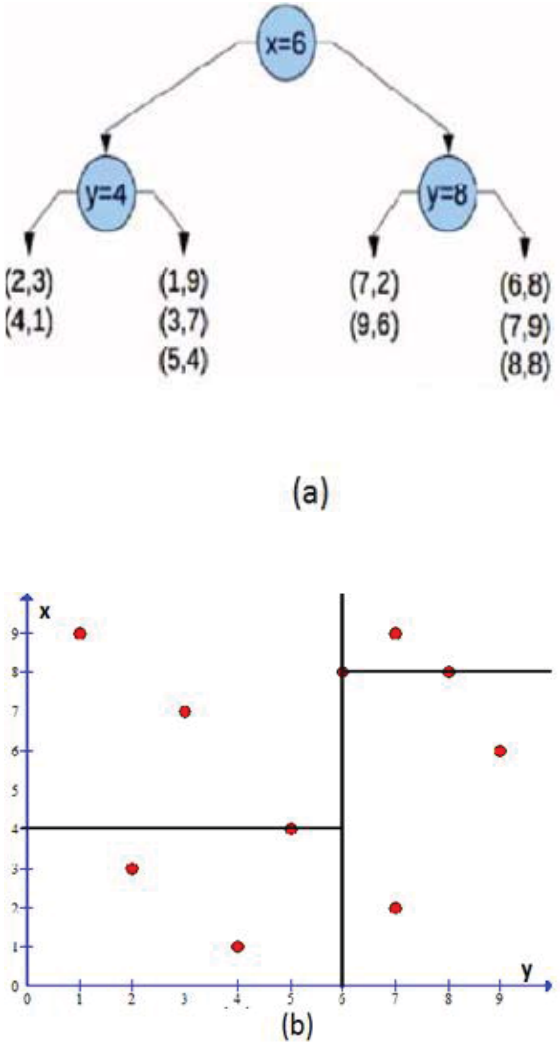


Figure 4: Fig 4(a) shows the tree representation of a training data set and Fig 4(b) shows its corresponding K-d representation of the above given data [6].

A. Steps followed in the Algorithm:

1) We first subsample the input original image so as to give the required degradation i.e. one half the original no. of pixels.
2) Now we apply an initial cubic spline interpolation for generating the image of the desired number of pixels which lacks in high resolution details [8].
3) Again after which we deposit the differences between the original input image and the interpolated image.
4) With this we stockpile the high resolution patch which is corresponding to every low resolution patch and they are typically of 5x5 and 7x7 pixels respectively.
5) Now this process is repetitively executed to get the low resolution patch to be totally converted into the high resolution image.

IV Block Diagram

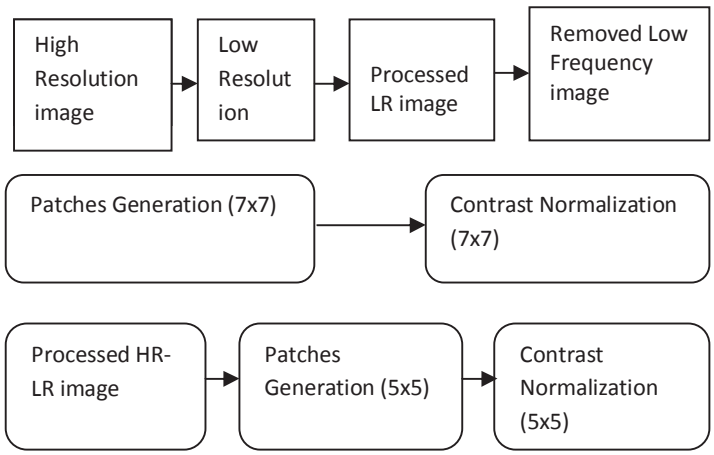


Figure 5: Training Set (Database) Generation.

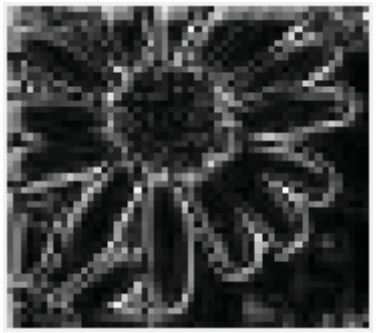
V RESULTS

The results thus generated are shown, these are generated from our algorithm [13] as we can see our algorithm shows a satisfactorily amount of output in the form of flower image shown in Fig 6(f). This image is thoroughly processed from the Fig 6(a) and thus various amount of pre-processing takes place. It differs from different image analogies in that we have used patches to create the new patch of image, and instead of operating in per-pixel. The difference clearly shows a performance benefit to super-resolution.

(a)

(b)

(c)



(d)

(e)

(f)

Figure 6: Fig 6(a) shows the original input image, Fig 6(b) is a sub-sampled image of Fig 6(a), Fig 6(c) is the interpolated image i.e. cubic spline interpolation, Fig 6(d) is a difference interpolated image of original image and interpolated image, Fig 6(e) is a difference super resolved image and Fig 6(f) is a final super resolved image.

VI Conclusion

The final super resolution image i.e. generated in this paper through learning approaches takes into consideration feature extracted from flowers of difference kind biologically different by observing the results we may conclude that different kind of flowers shares same features in context of image processing similar to the idea that human face can be generated from images other human face is a weighted combination.

References

[1] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution,"Computer Graphics and Applications, IEEE, vol. 22, no. 2, pp. 56–65, 2002.

[2] S. Vishnukumar, M. S. Nair, and M. Wilscy, "Edge preserving single image super-resolution with improved visual quality," Signal Processing, vol. 105,pp. 283–297, 2014.

[3] P. Cheng, Y. Qiu, K. Zhao, and X. Wang, "A transductive graphical model for single image super-resolution," Neurocomputing, vol. 148, pp. 376–387, 2015.

[4] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," Signal Processing Magazine, IEEE, vol. 20, no. 3, pp. 21–36, 2003.

[5] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in Computer Vision, 2009 IEEE 12th International Conference on, pp. 349–356, IEEE, 2009.

[6] P.Vandewalle, "Super-resolution from unregistered aliased images." Thesis Directors: Martin Vetterli and Sabine S¨usstrunk; Thesis No 3591, July 2006.

[7] Tian, Jing and Ma,kai-Kuang, "A Survey on super-resolution imaging","Signal Processing Magazine IEEE ", vol.no. 5, pp.329-342, 2011.

[8] Yang, Jianchao and Wright,John and Huang, Thomas S and Ma, Yi,"Image super-resolution via sparse representation" Image Processing, IEEE Transaction on, vol.no.19, pp. 2861-2873, 2010.

[9] Peyman Milanfar, "Super-Resolution imaging" 'CRC Press', "Taylor and Francis group"2010

[10] K. I. Kim and Y. Kwon, "Example-based learning for single-image super-resolution," in Pattern Recognition, pp. 456–465, Springer, 2008.

[11] Estner, H.L. ; Herzka, D.A. ; Mcveigh, E.R. ; Halperin, H.R., "Quantitative Assessment of Single-Image Super-Resolution in Myocardial Scar Imaging, "Translational Engineering in Health and Medicine, IEEE Journal of ", vol no:2,pp1-12, 2014

[12] Lingfeng Wang and Shiming Xiang and Gaofeng Meng and Huaiyu Wu and Chunhong Pan," Edge-Directed Single-Image Super-Resolution Via Adaptive Gradient Magnitude Self-Interpolation""Circuits and Systems for Video Technology, IEEE Transactions on",vol no.23,pp.1289-1299,2013.

[13] Baker, S.; Kanade, T," Limits on super-resolution and how to break them", Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.9, pp.1167,1183,2002

[14] Shechtman, E.; Caspi, Y.; Irani, M., "Space-time super-resolution," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.27, no.4, pp.531,545, April 2005.

[15] Lingfeng Wang; Shiming Xiang; Gaofeng Meng; Huaiyu Wu; Chunhong Pan, "Edge-Directed Single-Image Super-Resolution Via Adaptive Gradient Magnitude Self-Interpolation," Circuits and Systems for Video Technology, IEEE Transactions on , vol.23, no.8, pp.1289,1299, Aug. 2013.

SESSION

SIXTH WORKSHOP ON SOFT COMPUTING IN IMAGE PROCESSING AND COMPUTER VISION, SCIPCV + FUZZY LOGIC

Chair(s)

Dr. Gerald Schaefer
Dr. Pawan Lingras
Dr. Kouki Nagamune

Automatic Coronary Artery Segmentation Based on Matched Filters and Estimation of Distribution Algorithms

Ivan Cruz-Aceves[1], **Arturo Hernandez-Aguirre**[2], **and Ivvan Valdez-Peña**[2]
[1]Cátedras-CONACYT at Centro de Investigación en Matemáticas, A.C., Guanajuato, Guanajuato, México
[2]Centro de Investigación en Matemáticas, A.C., Guanajuato, Guanajuato, México

Abstract— *This paper presents an estimation of distribution algorithm (EDA) for improving the vessel detection performance of Gaussian matched filters (GMF) in X-ray angiographic images. The GMF method is governed by three discrete parameters and one continuous parameter. The optimal selection of the GMF parameters is highly desirable to maximize the detection rate of blood vessels in different types of medical images. The proposed optimization of these four parameters is carried out by applying a population-based method. From all the potential solutions found by EDA, the area (A_z) under the receiver operating characteristic (ROC) curve is used as fitness function to obtain the best GMF parameters and the corresponding Gaussian response. The detection performance of the proposed method is compared with those obtained using five different GMF methods of the state-of-the-art and the ground-truth vessels hand-labeled by a specialist. The experimental results applying the proposed method demonstrated high detection rate with $A_z = 0.9113$ using a training set of 40 angiograms and $A_z = 0.9343$ with a test set of 40 angiograms.*

Keywords: Automatic segmentation, Estimation of distribution Algorithm, Gaussian matched filters, Genetic algorithms, Vessel enhancement

1. Introduction

Coronary angiography is a specialized X-ray procedure for diagnosing and treating coronary artery disease. In recent years, the development of efficient and accurate computational methods has become essential for computer-aided diagnosis in cardiology. In general, there are two main drawbacks for automatic segmentation of vessels from coronary angiograms; nonuniform illumination along vessel structures and low contrast between vessels and image background. Due to these drawbacks, the vessel enhancement also called vessel detection problem plays an important and challenging role in most of the state-of-the-art coronary artery segmentation methods.

In literature, several methods have been proposed for automatic detection of coronary arteries from X-ray angiographic images. Most of the proposed automatic vessel detection methods are performed in the spatial image domain of the input angiogram such as single-scale top-hat operator [1], multiscale top-hat operator [2], [3], hit-or-miss transform [4], Hessian matrix [5], [6], [7], and Gaussian matched filters (GMF) [8], which have been used in different types of clinical studies including retinal image segmentation and registration [9], [10].

The GMF method is an spatial template matching technique used in the detection of different blood vessels. It works on the assumption that the shape of blood vessels can be approximated by a Gaussian curve as matching template. This Gaussian template is rotated at different angles and then convolved with the input image to form a filter bank of oriented responses. The maximum response at each pixel is recorded to obtain the final enhanced image. The GMF method has four main parameters, which have to be tuned to obtain the highest detection performance. The parameter L that determines the length of the vessel segment, σ that defines the spread of the intensity profile, T which is the position where the Gaussian curve trails will cut, and θ that represents the angular resolution of the filter bank.

Since the GMF was introduced, many researchers have suggested different values for each parameter of the filter. Kang et al. [11], [12], [13] applied the GMF taking into account different values for σ and angular resolution. Al-Rawi et al. [14] proposed extend the range of the variables L, T, and σ obtaining the best values by using an exhaustively deterministic search method and the area (A_z) under the receiver operating characteristic (ROC) curve. Cinsdikici and Aydin [15] used the original parameters of the method, just modifying the angular resolution. Al-Rawi and Karajeh [16] applied the population-based method of genetic algorithms in order to select the best L, T , and σ values for vessel detection keeping constant the angular resolution. The performance of the population-based method working together with the Gaussian matched filter is more accurate according to the tests than the empirically determined methods.

The population-based methods represent an effective way to solve discrete or continuous optimization problems. Recently, the Estimation of Distribution Algorithms (EDAs) have begun to attract more attention for solving global optimization problems. EDAs are stochastic methods consisting on a set of potential solutions called population that incorporate statistical knowledge to solve optimization problems [17], [18]. In EDAs a probabilistic model of the potential solutions is constructed at each generation, and the new solutions are generated from this model until a stopping

criterion is satisfied. EDAs have proven to be efficient in solving optimization problems such as cancer chemotherapy optimization [19], image segmentation [20], and proving to be more effective than genetic algorithms in discrete optimization systems [21].

Because of the importance of the Gaussian matched filters in many of the aforementioned segmentation methods, an optimization process for selecting the most suitable parameters is required. In this paper we propose the use of the Univariate Marginal Distribution Algorithm (UMDA) from the family of EDAs for improving the detection performance of the Gaussian matched filters. This proposed method addresses the problem of detecting coronary arteries in X-ray angriographic images. The optimization method is performed over the four GMF parameters, where L, T, and θ are discrete values, and σ represents a continuous value. The vessel detection performance of the proposed method is compared and analyzed with those obtained using five, previously described, state-of-the-art vessel detection methods by using the area A_z under the ROC curve.

The remainder of this paper is organized as follows. In Section 2, the basics of the Gaussian matched filter and Univariate Marginal Distribution Algorithm are introduced. In Section 3, the parameter optimization process is presented and analyzed. The experimental results are discussed in Section 4, and conclusions are given in Section 5.

2. Background

In this section, the fundamentals of the Gaussian matched filters and the univariate marginal distribution algorithm are described in detail.

2.1 Gaussian matched filters (GMF)

The Gaussian matched filters method [8] has been used for detecting blood vessels in different types of medical images. The main idea of GMF is to approximate the shape of blood vessels in the spatial image domain utilizing a Gaussian curve as matching template, which can be defined as follows:

$$G(x,y) = -\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \quad |y| \leq L/2, \qquad (1)$$

where L is the length of the vessel segment to be detected in pixels, and σ represents the average width of blood vessels, which uses an implicit parameter T (in pixels) to define the position in the template where the Gaussian curve trails will cut. This Gaussian kernel $G(x,y)$ is rotated at different angular resolutions (θ), obtaining $\kappa = 180/\theta$ oriented filters. These filters are convolved with the input image, and for each pixel the maximum response over all orientations is preserved to obtain the filtered resulting image.

In order to obtain the best detection performance of the GMF, the discrete parameters of L, T, and θ, and the continuous parameter of σ have to be optimized. In Figure 1, an X-ray coronary angiogram is illustrated along with its hand-labeled image (ground-truth) as drawn by a specialist. The Figure 1(c) presents the Gaussian template as it was defined in the work of Chaudhuri et al. [8] with values $L = 9$, $T = 13$, $\theta = 15$, $\kappa = 12$, and $\sigma = 2.0$, which have been used successfully in different segmentation applications [9], [10], and the enhancement result of the Gaussian filter method is given in Figure 1(d).

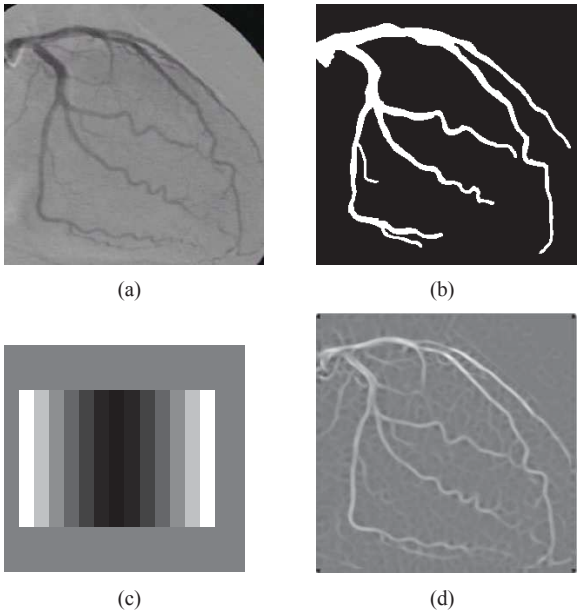


Fig. 1: (a) Original X-ray coronary angiogram. (b) Ground-truth of angiogram in (a). (c) Gaussian template with the predefined parameters of Chaudhuri et al. [8]. (d) Resulting enhanced image using the angiogram in (a) and the template in (c).

2.2 Estimation of Distribution Algorithms (EDAs)

Estimation of distribution algorithms represent population-based methods that incorporate statistical information of potential solutions to solve optimization problems in discrete and continuous domain [17], [18]. EDAs are similar to evolutionary computation (EC) techniques since they use a population of potential solutions called individuals, selection operators, and binary encoding. The main difference between EC methods and EDAs is the fact that the crossover and mutation operators are not required by EDAs, instead, they build a probabilistic model based on global statistical information of promising individuals at each generation. This probabilistic model is used to generate new potential solutions by inferring statistical dependencies between the variables. In the present work, the Univariate Marginal Distribution

Algorithm (UMDA) has been adopted because of ease of implementation and because it works ideally for linear problems with not many significant dependencies [21], [22].

UMDA uses a binary codification of the problem, and it generates a probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_n)^T$ to build the probabilistic model at each generation in order to create new individuals for each variable independently, where p_i is a probability rate. The key idea of UMDA is to approximate the probability distribution of the potential solutions in $\mathbb{P}_t$ as the product of the univariate frequencies computed from a subset of individuals assuming that all the variables are independent [23]. The steps of selection of promising solutions, estimation of probability distribution and creation of new individuals represent the evolutionary process of UMDA. To perform the selection step, the individuals in the search space Ω have to be arranged according to fitness value, then selection probability s is computed through proportional selection as follows:

$$\mathbb{P}^s(x) = \frac{\mathbb{P}(x)f(x)}{\sum_{\tilde{x}\in\Omega}\mathbb{P}(\tilde{x})f(\tilde{x})}. \quad (2)$$

The second step consisting in the estimation of a joint probability $\mathbb{P}$ is calculated as follows:

$$\mathbb{P}(x) = \prod_{i=1}^{n}\mathbb{P}(X_i = x_i), \quad (3)$$

where $x = (x_1, x_2, \ldots, x_n)^T$ represents the binary value of ith bit in the potential solution, and X_i is the ith value of the random vector X. Finally, the third step generates new individuals applying the estimated probability distribution, which is evaluated by the fitness function through generations, and these three steps are performed until a convergence criterion is satisfied.

According to the above description, UMDA can be implemented by the following procedure:

1) Establish number of generations t.
2) Generate n individuals randomly initialized.
3) Select a subset of individuals S of $m \leq n$ according to a selection method.
4) Calculate the univariate marginal probabilities $p_i^s(x_i, t)$ of S.
5) Generate n new individuals by using $p(x, t+1) = \prod_{i=1}^{n} p_i^s(x_i, t)$.
6) Stop if convergence criterion is satisfied (e.g., stability or number of generations), otherwise, repeat steps (3)-(5).

3. Optimization of the Gaussian matched filter

Since the GMF method was introduced by Chaudhuri et al. [8], many researchers have suggested different values for each parameter of the filter. The original work of Chaudhuri et al. fixed these parameters as $L = 9$, $T = 13$, $\sigma = 2.0$, and $\theta = 15$, obtaining $\kappa = 12$ oriented filters. Kang et al. [11], [12], [13] applied the GMF with an angular resolution of $\theta = 30$ degrees, obtaining 6 different kernels, with an average vessel width of $\sigma = 1.5$. Cinsdikici and Aydin [15], applied the original parameters of the GMF method, just modifying the angular resolution to $\theta = 10$, obtaining $\kappa = 18$ oriented filter responses. Al-Rawi et al. [14] proposed extend the range of variables to $L = \{7, 7.1, \ldots, 11\}$, $T = \{2, 2.25, \ldots, 10\}$, $\sigma = \{1.5, 1.6, \ldots, 3\}$, and keeping constant $\theta = 15$, with $\kappa = 12$ oriented filters. In this method, the best detection performance is obtained through an exhaustively search over all possible combinations, and using the area (A_z) under the receiver operating characteristic (ROC) curve. On the other hand, in order to avoid the exhaustively search to obtain the best GMF performance, Al-Rawi and Karajeh [16] applied the population-based method of Genetic Algorithms (GAs) to select the best L, T , and σ values keeping constant the angular resolution. The detection results acquired by the population-based method are promising to reduce the number of evaluations and obtaining superior performance than the empirically defined methods.

Due to the suitable performance of the population-based method, in the present work the univariate marginal distribution algorithm has been adopted to perform the optimization task in X-ray coronary angiograms. The search space of the GMF variables was defined according to the aforementioned methods and taking into account the analyzed coronary angiograms as $L = \{8, 9, \ldots, 15\}$, $T = \{8, 9, \ldots, 15\}$, and $\sigma = [1, 6]$ with an step size $\Delta = 0.001$. The number of oriented filters was set as $\kappa = 12$ with angular resolution $\theta = 15$. Similar A_z results were obtained with $\kappa = 15, 20, 30, 45$, and 60 filters. For further analysis, only $\kappa = 12$ was applied. The binary encoding of the UMDA population is set to 18 bits, where 12 bits are used for σ parameter, 3 bits for L parameter and the remaining 3 bits for T parameter. The objective function to be maximized is the area A_z under ROC curve which represents the true-positive fraction (TPF) against false-positive fraction (FPF). TPF is the rate of vessel pixels in the ground-truth image that are correctly detected by the method, also known as sensitivity metric. FPF is defined as the rate of nonvessel pixels that are incorrectly classified as vessel pixels by the computational method. The area under ROC curve is one for perfect detection, and zero otherwise.

4. Experimental results

The proposed GMF-UMDA method was tested on a computer with an Intel Core i3, 8GB of RAM, 2.13 GHz processor through Matlab software version 2014b.

The database used in the present work consists of 80 X-ray coronary angiograms of size 300 $\times$ 300 pixels of

27 different patients. Each coronary angiogram was hand-labeled by a specialist. Ethics approval was provided by the Cardiology Department of the Mexican Social Security Institute. To assess the detection performance of the computational methods, the database was divided into two subsets of images. The first subset consists of 40 angiograms, which is used as training set for tuning purpose, and the second subset of the remaining 40 angiograms is used as the test set for evaluation of vessel detection methods.

The performance of the proposed method is compared with five GMF of the state-of-the-art for automatic vessel detection. In Figure 2, the detection performance of the methods over a subset of the training angiograms is presented. These ROC curves are obtained by concatenating the filtered angiograms of the training set as only one large image, and applying the set of parameters discussed above. The genetic algorithm used in the work of Al-Rawi and Karajeh [16], is applied with the same set of parameters as: population size $= 30$, crossover fraction $= 0.7$, mutation fraction $= 0.3$, elite $= 1$, with an heuristic multi-point as crossover method. UMDA is applied with a population size $= 30$ and selection rate $= 0.6$. Both population-based methods employ 40 generations as stopping criterion. The comparative analysis suggests that the GMF-UMDA method provide superior performance in vessel pixel detection than the comparative five methods in the 40 training angiograms. To visualize the detection results of the comparative analysis, in Figure 3 the Gaussian filter response of the six methods is introduced. By visual analysis it can be observed that the proposed method presents in general, a higher discrimination of false-positive pixels and higher intensity in vessel pixels than the comparative five methods.

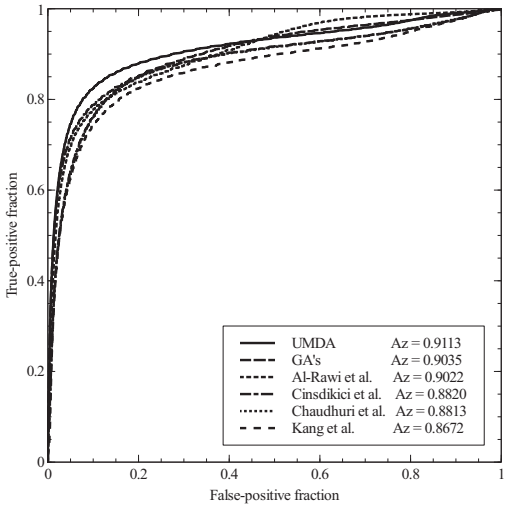


Fig. 2: Comparison of ROC curves for vessel detection with the training set of 40 angiograms, using the proposed method and the comparative methods.

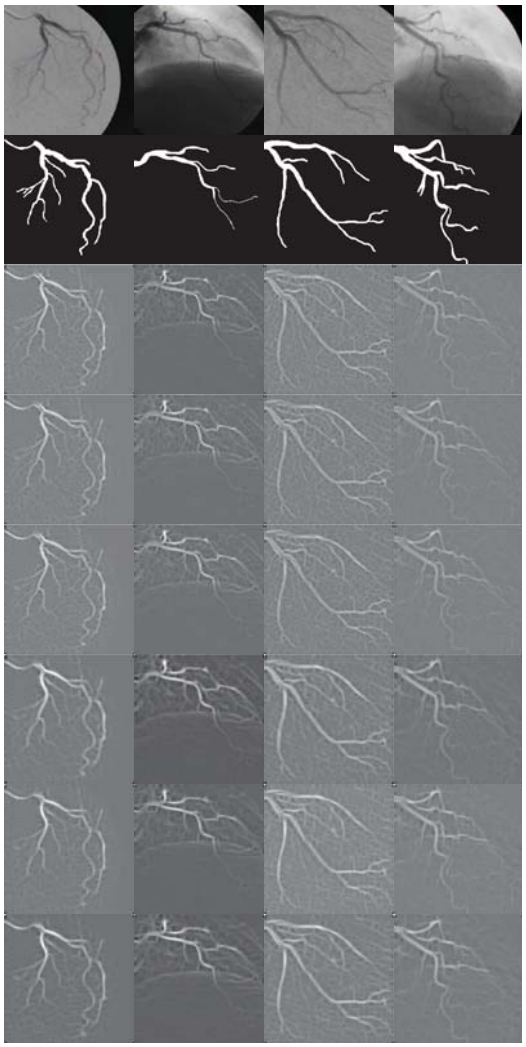


Fig. 3: First row: subset of angiographic images from the training set. Second row: ground-truth of the images in first row. The remaining six rows present the Gaussian filter response of the methods of Kang et al. [11], [12], [13], Chaudhuri et al. [8], Cinsdikici and Aydin [15], Al-Rawi et al. [14], Al-Rawi and Karajeh [16], and proposed GMF-UMDA method, respectively, applied on the angiograms in first row.

From the comparative analysis of A_z values over the training set of angiograms, the best GMF detection parameters of the two population-based methods were acquired. The genetic algorithm found as best parameters $L = 14$ pixels, $T = 13$ pixels, and $\sigma = 5.370$. UMDA found as best parameters $L = 15$ pixels, $T = 15$ pixels, and $\sigma = 2.414$. In Table 1, the detection performance of the analyzed methods is presented. The A_z rate of the empirical methods of Kang et al., Chaudhuri et al., and Cinsdikici and Aydin shows in general low performance. These three methods use similar values of L, T, and σ, and the main

difference among them is the number of oriented filters, which were defined as $\kappa = 6, 12, 18$, respectively. The exhaustively method of Al-Rawi et al. [14], presents superior performance than the empirical methods. This is mainly due to the search space in which the method is optimized. The best GMF parameters obtained by the method were found as $L = 11$, $T = 8$, and $\sigma = 1.9$. To perform the GA and UMDA strategies, the search space for three variables of the GMF method was extended in comparison with the method of Al-Rawi et al. discussed above. Due to this fact, both evolutionary methods present a superior performance than the exhaustively strategy. The detection results obtained by the proposed method using UMDA presents the highest A_z rate over the test set of 40 angiograms.

Table 1: Comparison of detection performance with the test set of 40 angiograms, using the proposed and comparative methods.

Method	Area under ROC curve (A_z)
Proposed GMF-UMDA	**0.9343**
GMF based on GA [16]	0.9239
Al-Rawi et al. [14]	0.9232
Cinsdikici and Aydin [15]	0.8934
Chaudhuri et al. [8], [9], [10]	0.8918
Kang et al. [11], [12], [13]	0.8843

Moreover, to illustrate the robustness of the UMDA against genetic algorithm for improving the Gaussian matched filter performance, in Table 2 an statistical analysis is shown. This analysis was performed with 30 runs over the test set of angiograms, where the mean and standard deviation values show that UMDA is more stable and with solutions closer to the mean than the best solutions found by the GA strategy.

Table 2: Statistical analysis with 30 runs of the GA and UMDA methods over the test set of 40 angiograms.

Method	Max.	Min.	Mean	Std. Dev	Median
GMF-GA [16]	0.9239	0.8596	0.8960	0.0148	0.9007
GMF-UMDA	**0.9343**	**0.8598**	**0.9101**	**0.0121**	**0.9106**

Generally, after applying a detection method, thresholding strategies are then used to classify vessel structures as white pixels and background information as black pixels. Although ROC analysis is used to quantify the vessel detection performance of the methods, also it is used to define a threshold value obtained with the best trade-off between true-positive and false-positive fractions. Figure 4 presents a subset of angiograms from the test set, which is filtered by the GMF method based on GA and UMDA strategies, and then it is thresholded by using the best classification value from ROC analysis. By this thresholding strategy, the segmentation results obtained from the proposed GMF-UMDA method show an appropriate rate of true-positive pixels with low rate of false-positive pixels compared with the GMF-GA method.

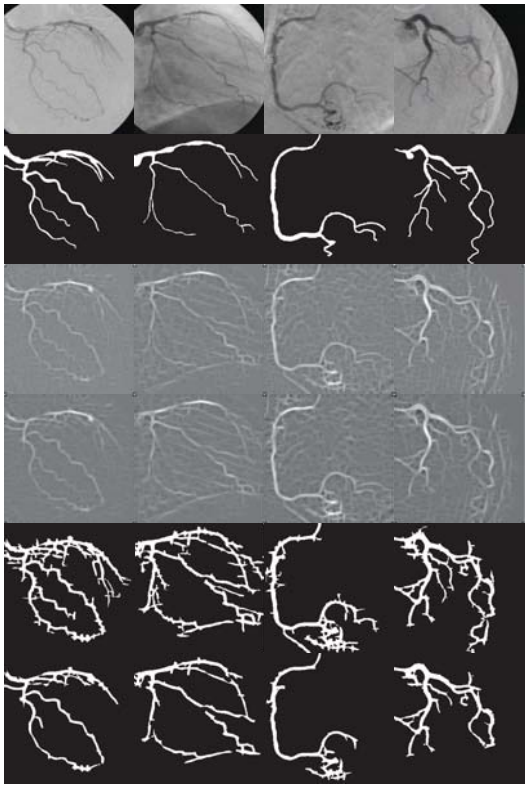


Fig. 4: First row: subset of angiographic images from the test set. Second row: ground-truth of the images in first row. Third row: Gaussian filter response based on GA strategy. Fourth row: Gaussian filter response based on proposed UMDA strategy. The remaining two rows present the thresholded responses of GA and UMDA by ROC analysis respectively.

The detection methods based on Gaussian matched filters of the state-of-the-art discussed above, provide suitable performance based on the area A_z under ROC curve. However, different comparative analysis reveal that the improved GFM method based on the univariate marginal distribution algorithm is robust and suitable for segmenting vessels in coronary angiograms. The results have also shown that the detection performance obtained from the proposed method is suitable for computer-aided diagnosis considering the coronary arteries hand-labeled by a specialist.

5. Conclusion

In this paper, an estimation of distribution algorithm has been applied for improving the Gaussian filter performance in coronary artery detection from X-ray angiograms. The use of UMDA produces a number of potential different Gaussian filters. All the candidate solutions were evaluated

by computing the area under ROC curve, and the best GMF parameters found were set as $L = 15$ pixels, $T = 15$ pixels, and $\sigma = 2.414$. The performance of the proposed strategy has demonstrated to be more efficient compared with five GMF methods of the state-of-the-art achieving $A_z = 0.9113$ with a training set of 40 angiograms. According to the experimental results, the proposed GMF method using UMDA as optimization strategy can lead to higher accuracy and efficiency than the comparative methods obtaining $A_z = 0.9343$ with the test set of 40 angiograms. In addition, considering the images hand-labeled by specialist and the segmentation results obtained from the improved GMF based on UMDA, it can be highly suitable for computer-aided diagnosis in cardiology. As future work, we plan to study continuous optimization methods, in order to apply our method to detect vessel abnormalities and prevent coronary artery stenosis.

Acknowledgment

This research has been supported by the National Council of Science and Technology of México (Project Cátedras-CONACYT No. 3150-3097). The authors would like to thank the Cardiology Department of the Mexican Social Security Institute, UMAE T1 León, for making this research possible providing us the sources of coronary angiograms and for valuable clinical advice.

References

[1] S. Eiho and Y. Qian, "Detection of coronary artery tree using morphological operator," *Computers in Cardiology*, vol. 24, pp. 525–528, 1997.

[2] Y. Qian, S. Eiho, N. Sugimoto, and M. Fujita, "Automatic extraction of coronary artery tree on coronary angiograms by morphological operators," *Computers in Cardiology*, vol. 25, pp. 765–768, 1998.

[3] K. Sun and N.Sang, "Morphological enhancement of vascular angiogram with multiscale detected by gabor filters," *Electronic letters*, vol. 44, no. 2, January 2008.

[4] B. Bouraoui, C. Ronse, J. Baruthio, N. Passat, and P. Germain, "Fully automatic 3D segmentation of coronary arteries based on mathematical morphology," *5th IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, pp. 1059–1062, 2008.

[5] C. Lorenz, I. Carlsen, T. Buzug, C. Fassnacht, and J. Weese, "A multi-scale line filter with automatic scale selection based on the Hessian matrix for medical image segmentation," *Proc. Scale-Space Theories in Computer Vision, Springer LNCS*, vol. 1252, pp. 152–163, 1997.

[6] A. Frangi, W. Niessen, K. Vincken, and M. Viergever, "Multiscale vessel enhancement filtering," *Medical Image Computing and Computer-Assisted Intervention (MICCAI'98), Springer LNCS*, vol. 1496, pp. 130–137, 1998.

[7] H. Shikata, E. Hoffman, and M. Sonka, "Automated segmentation of pulmonary vascular tree from 3D CT images," *Proc. SPIE International Symposium Medical Imaging*, vol. 5369, pp. 107–116, 2004.

[8] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 263–269, Sept. 1989.

[9] T. Chanwimaluang and G. Fan, "An efficient blood vessel detection algorithm for retinal images using local entropy thresholding," *Proc. IEEE International Symposium on Circuits and Systems*, vol. 5, pp. 21–24, May 2003.

[10] T. Chanwimaluang, G. Fan, and S. Fransen, "Hybrid retinal image registration," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 129–142, Jan 2006.

[11] W. Kang, K. Wang, W. Chen, and W. Kang, "Segmentation method based on fusion algorithm for coronary angiograms," *2nd International Congress on Image and Signal Processing (CISP)*, pp. 1–4, 2009.

[12] W. Kang, W. Kang, W. Chen, B. Liu, and W. Wu, "Segmentation method of degree-based transition region extraction for coronary angiograms," *2nd International Conference on Advanced Computer Control*, pp. 466–470, 2010.

[13] W. Kang, W. Kang, Y. Li, and Q. Wang, "The segmentation method of degree-based fusion algorithm for coronary angiograms," *2nd International Conference on Measurement, Information and Control*, pp. 696–699, 2013.

[14] M. Al-Rawi, M. Qutaishat, and M. Arrar, "An improved matched filter for blood vessel detection of digital retinal images," *Computers in Biology and Medicine*, vol. 37, pp. 262–267, 2007.

[15] M. Cinsdikici and D. Aydin, "Detection of blood vessels in ophthalmoscope images using MF/ant (matched filter/ant colony) algorithm," *Computer methods and programs in biomedicine*, vol. 96, pp. 85 –95, 2009.

[16] M. Al-Rawi and H. Karajeh, "Genetic algorithm matched filter optimization for automated detection of blood vessels from digital retinal images," *Computer methods and programs in biomedicine*, vol. 87, pp. 248 –253, 2007.

[17] P. Larrañaga and J. Lozano, *Estimation of distribution algorithms: A new tool for evolutionary computation*. Boston, Mass, USA: Kluwer Academic, 2002.

[18] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 111–128, 2011.

[19] A. Petrovski, S. Shakya, and J. McCall, "Optimising cancer chemotherapy using an estimation of distribution algorithm and genetic algorithms," *in Proceedings of the 8th Annual Genetic and Evolutionary Computation Conference*, pp. 413–418, 2006.

[20] I. Cruz-Aceves, J. Avina-Cervantes, J. Lopez-Hernandez, and et al., "Automatic image segmentation using active contours with univariate marginal distribution," *Mathematical Problems in Engineering*, vol. 419018, p. 9, 2013.

[21] S. Bashir, M. Naeem, and S. Shah, "A comparative study of heuristic algorithms: GA and UMDA in spatially multiplexed communication systems," *Engineering Applications of Artificial Intelligence*, vol. 23, pp. 95–101, 2010.

[22] H. Muehlenbein, "The equation for response to selection and its use of prediction." *Evolutionary Computation*, vol. 5, no. 3, pp. 303–346, 1997.

[23] L. Lozada-Chang and R. Santana, "Univariate marginal distribution algorithm dynamics for a class of parametric functions with unitation constraints," *Information Sciences*, vol. 181, pp. 2340–2355, 2011.

Summarizing Traffic Camera Images based on Weather, Traffic, and Lighting Attributes

Mathew Kallada
Dalhousie University
Halifax, Nova Scotia
Email: kallada@cs.dal.ca

Pawan Lingras
Saint Mary's University
Halifax, Nova Scotia
Email: pawan@cs.smu.ca

Jason Rhinelander
Saint Mary's University
Halifax, Nova Scotia
Email: jason.rhinelander@smu.ca

***Abstract*—Network attached cameras, or webcams, provide a relatively inexpensive source of visual information for Internet users. In recent years, live webcams have been used to monitor active regions of cities to allow users to view sites of interest. Live traffic images can be used for a more detailed analysis as well as for creating useful decision support systems using statistical learning techniques. The data set size produced by a streaming webcam grows rapidly, and methods for effectively summarizing the key traits of traffic images becomes increasingly important for reducing the overall data set size. This paper describes a feature extraction system for traffic webcams images, that represents an individual image using attributes based on three criteria that are important to drivers: weather, lighting, and traffic conditions. The processed data set is summarized by the use of unsupervised clustering to demonstrate the ability of the proposed feature extraction step.**

***Keywords*—feature extraction, activity detection, image analysis**

I. Introduction

The image stream provided by a traffic webcam a good source for data mining and predictive analytics. This paper uses the webcam traffic images from MacKay bridge of Nova Scotia, Canada. The MacKay bridge connects major metropolitan areas within the city of Halifax, and it is one of the most heaviest traffic areas in the region where thousands of vehicles cross the bridge each day. The bridge is an important determinant of the travel time for a large number of commuters. However, while a raw visual depiction of the traffic image is more meaningful to a user than textual or numeric summary, browsing through such a large sequence of images for discovering any meaningful patterns is a difficult task for an individual person or a computer application. In order to develop useful tools based on the images obtained from such a webcam, methods for summarizing the key traits from a traffic image is important to quantify the weather, traffic, and lighting conditions.

Therefore, it is essential to process the images and extract relevant information in order to summarize traffic data sets for easier browsing and further analysis. A number of researchers have proposed image processing techniques that make it possible to describe various features of these images [1]. This paper describes the details of an algorithm that processes images from a traffic webcam to create a semantically enhanced vector representation of the image. We further use unsupervised clustering to group the images and to understand the performance of our proposed extraction process through evaluation of the cluster quality. The medoids (the data point nearest to the centroid) of the resulting clusters show that the proposed feature representation of the images helps us distinguish images based on weather, traffic, and lighting conditions. The resultant cluster medoids has been used in previous work [2] as an input label for visual traffic predictions using supervised learning.

II. Dataset Preparation

The study data in this paper was provided by *Nova Scotia Webcams* consisting of 71 cameras in total, with the size of the MacKay Bridge road traffic data set being 4.3GB. The present study focuses on a single camera stream that monitors the traffic conditions on the MacKay bridge that links the communities of Halifax to Dartmouth. The traffic camera collects a snapshot every 10 minutes with a resolution of 960×600 pixels and was collected over a 21 month period from May 29th, 2012 to January 18th, 2014. We restricted our image analysis to the hours between 7 am and 7 pm which corresponds to peak activity. The resulting data set had a total of 45,899 observations.

III. Image Processing and Feature Extraction Method

Varying image processing techniques are used to capture weather, traffic, and lighting conditions. The collection of images came from a single geographical location. While the structure of the overall scene is essentially the same, the images differ from each other in terms of weather, lighting conditions, and traffic activity. The techniques used to process the images in our study include:

- **Image subtraction** to find the movement or traffic activity between subsequent images. This helps in identifying cars on the bridge. Each gray scaled image was subtracted from the previous ten images.
- **Blob detection** was achieved by filtering out small movements on the bridge by applying an intensity threshold to the subtracted image. The average number of blobs from the ten subtracted images was used to determine the blobs in the current image. We use attributes associated with the blobs detected to measure the activity level occurring on the bridge.
- **Weather feature extraction** was done using GLCM features and pixel-wise operations on a region of the

sky in the image. The use of GLCM features provides texture features in the sky, and average color intensity values can help to separate a clear blue sky from a uniform gray sky.

The remainder of the section describes the methods used to implement each stage of our image processing pipeline. The two primary image processing techniques employed in this work are texture analysis, and activity detection.

A. Region of Interests

Two Regions of Interest (ROI) polygons are defined and extracted from each image. The first ROI is the traffic region encompassing the bridge area, and the second ROI is in the sky region of the image for weather analysis. Pixel-level operations such as average red, green and blue pixel intensity operations are computed for the weather ROI, and gray scale conversion is performed on both ROIs.

Fig. 1. The traffic region of interest is highlighted in the image. The traffic region allows for extraction of activity features from the image.

Fig. 2. The sky condition region of interest is highlighted in the image. Texture analysis of the sky region allows for extraction of weather and lighting features from the image.

B. Gray-Scale Texture Analysis

Texture-based feature extraction is performed on both gray scale ROIs by the calculation of a gray-level co-occurrence matrix (GLCM) [3]. The features *contrast*, *dissimilarity*, *homogeneity*, *energy*, and *correlation* for the weather and traffic ROIs are extracted from the respective GLCM.

The contrast and dissimilarity attributes tend to be strongly correlated with high spatial frequencies in the texture of the ROI. Homogeneity is a feature that has smaller values for larger gray level differences across the ROI. As its name implies, the feature is sensitive to regular patterns. Higher values for the energy feature result from uniform or periodic gray levels across the ROI. Chaotic or random gray levels result in lower energy values. Finally, a high correlation feature value implies that there is a strong linear relationship between pixel values.

The GLCM has been found to be an effective method for the extraction of texture features from an image [4]. We use these features in our representation to distinguish between different textures in both the traffic and weather ROIs. The GLCM is a account of the number of times (or the probability) of co-occurrences of neighboring gray-level pixel values across a window of interest. It is a second order statistic where spatial variation is being measured. We compute the GLCM for pixels at a distance of 2, and at an angle of 90 degrees.

$$contrast = \sum_{i,j=0}^{l-1} P_{i,j} \left(i-j\right)^2 \tag{1}$$

$$dissimilarity = \sum_{i,j=0}^{l} P_{i,j} \left|i-j\right| \tag{2}$$

$$homogeneity = \sum_{i,j=0}^{l} \frac{P_{i,j}}{1+\left(i-j\right)^2} \tag{3}$$

$$energy = \sqrt{\sum_{i,j=0}^{l} P_{i,j}^2} \tag{4}$$

$$correlation = \sum_{i,j=0}^{l} P_{i,j} \left(\frac{\left(i-\mu_i\right)\left(j-\mu_j\right)}{\sqrt{\left(\sigma_i^2\right)\left(\sigma_j^2\right)}} \right) \tag{5}$$

Equations (1-5) are calculated over the GLCM where i is the row address and j is the column address in the GLCM, while $P_{i,j}$ represents the respective entry within the GLCM.

C. Traffic Activity Extraction

Since the webcam is at rest, detection of traffic activity within a sequence of images can be achieved through image subtraction. A Sobel filter is then applied to the image difference where "blobs" are formed. Differences found after Sobel filtering are thresholded to extract significant changes in the image. The pipeline for this process is summarized in Figure 3. After this process, the number of vehicles on the road is distinguishable. The number of blobs and the total area of all blobs is then recorded. The procedure can be repeated with the multiple previous images as a sliding window. The mean "blob" count and standard deviation in the "blob" count are both useful statistics and determine the activity level in the current image.

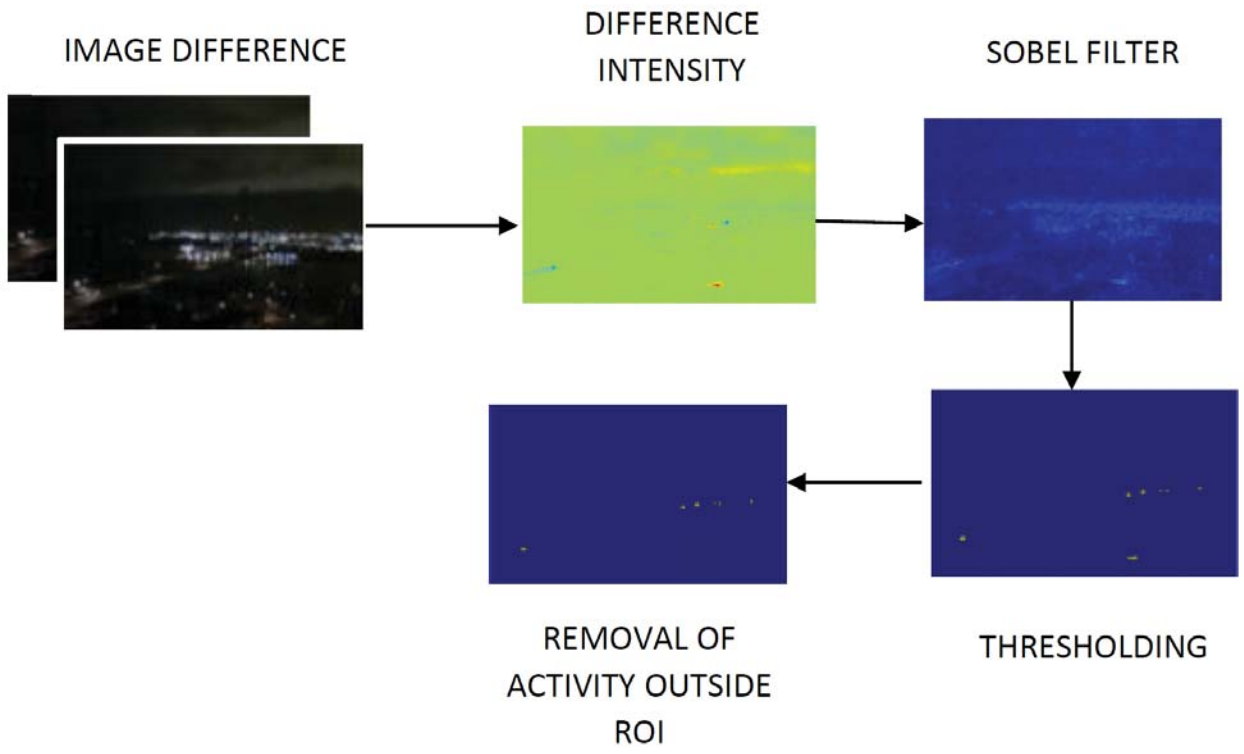


Fig. 3. Visual depiction of the pipeline for traffic activity extraction. The current image is subtracted from the previous 10 images resulting in a difference intensity image the highlights the temporal activity. A sobel filter is applied to enhance the edges around each activity blob. Thresholding is then applied to applied to the bridge ROI so that each independent blob is considered to be a vehicle.

D. Final Feature Vector Representation

Table I describes a list of attributes that were extracted from each image, their semantics, and their range of values (if applicable).

IV. Evaluation of Feature Representation with Cluster Analysis

The image representation obtained through the image processing described in Section III-D allows us to describe them in terms of the weather, lighting, time, and the amount of traffic present at the webcam location. In this section, we want to evaluate whether these attributes accurately represent the weather, traffic, and lighting conditions. We used unsupervised clustering to group webcam images based on similar attribute values.

We used a popular k-means clustering algorithm for grouping images. The objective of the k-means clustering algorithm [5], [6] is to find a locally optimal solution. The k-means algorithm creates k-clusters from a total of n objects. Since the k-means algorithm is a greedy approach, we run ten trials of k-means and pick the trial where the resulting clusters have the smallest cluster inertia. Since clustering is an unsupervised learning method, we used cluster validity measures to determine an appropriate value of number of clusters k. A number of cluster validity measures including Davies-Bouldin (DB) index [7] have been proposed to evaluate clustering schemes.

To assess the quality and semantic representation of each cluster, we visually analyzed the medoid image of each cluster. A medoid image is the image whose image attribute values are the closest to the centroid values of the cluster. Figure 4 shows the medoid images for each of the categories. These groups can be meaningfully described using the labels shown above each of the medoid images. Notably, each medoid image offers a unique visual reference for the weather and traffic conditions on the bridge.

A comparison of some of the key attribute values for the medoid images for each cluster suggest that the attributes derived from the images indeed help us distinguish between different weather, traffic and lighting conditions. One can see that most clusters have moderate traffic. There are more days with partly cloudy or cloudy weather conditions. The weather related attributes tend to distinguish images better than the traffic attributes. For example, weather_b and weather_g attributes for dark conditions in cluster 8 is significantly smaller than sunny bright days in clusters 7 and 9. A more detailed study to test the significance of different attributes in distinguishing the weather, traffic, and lighting conditions will help automate the image processing and prediction system.

1) **Cloudy moderate traffic:** The total number of images in this cluster is 8447.
2) **Cloudy light traffic:** The total number of images in this cluster is 3804.
3) **Mostly sunny moderate traffic:** The total number of images in this cluster is 8170.
4) **Cloudy dark moderate traffic:** The total number of images in this cluster is 1539.
5) **Cloudy rainy moderate traffic:** The total number of images in this cluster is 1510.

TABLE I. LIST OF FEATURE ATTRIBUTES EXTRACTED FROM TRAFFIC CAMERA IMAGES.

Attribute	Semantics	Value range
weather_b	The average blue channel intensity level for the sky region.	1.693-254.69
weather_homogeneity	The homogeneity feature from the weather GLCM.	13243-14021.5
blobs_found	The average number of blobs found in the image.	0-122.24
weather_g	The average green channel intensity level for the weather ROI.	1.69-255.0
weather_energy	The energy feature from the weather GLCM.	9845.13-14021.0
weather_dissimilarity	The dissimilarity feature from the weather GLCM.	1-1558
energy	The energy feature from the traffic GLCM.	1879.4-89879.5
computation_time	The time taken in processing the current image.	2.41-85.2
weather_contrast	The contrast feature from the weather GLCM.	1-1558
hour	The hour component (24 hour day cycle) of the time when the image was acquired.	7-19
homogeneity	The homogeneity feature from the traffic GLCM.	568848.5-574079.5
area	The average total area of all detected traffic blobs.	0-26550.4
weather_r	The average green channel intensity level for the weather ROI.	1.69-254.8
time	The time that the image was captured (yyyymmdd, y:year, m:month, d:day)	20120101-20121231
area_std	The standard deviation in the total area of all detected traffic blobs across the 10 previous images.	0-28500.5
dissimilarity	The dissimilarity feature from the traffic GLCM.	1-10463
correlation	The correlation feature from the traffic GLCM.	0.9999982581-1.0
day	The day of the month that the image was taken.	1-(31,30,28)
contrast	The contrast feature from the traffic GLCM.	1-10463
blob_std	The standard deviation in the total number of blobs detected across the 10 previous images.	0-122.243

Fig. 4. Medoid images for each cluster along with the corresponding feature vector representation. Semantic differences can be observed between each cluster medoid.

6) **Partly cloudy moderate traffic:** The total number of images in this cluster is 9256.
7) **Sunny bright heavy traffic:** The total number of images in this cluster is 1509.
8) **Dark light traffic:** The total number of images in this cluster is 3673.
9) **Sunny bright moderate traffic:** The total number of images in this cluster is 7991.

V. Conclusion and Future Work

This paper describes a strategy for summarizing traffic webcam images based on weather, traffic, and lighting conditions. The proposed image processing techniques are applied to a dataset of webcam images for a bridge that connects two major metropolitan areas. In order to assess the quality of our method, we apply k-means clustering of images. The experimental results from our cluster analysis shows that the images are indeed distinguished based on the three key aspects, namely weather, traffic, and lighting. Future publications will report the significance of the proposed attributes for predicting different aspects of the webcam images. Another area for future work is applying soft clustering techniques for evaluating the performance of our feature engineering method. In this analysis, we used a hard clustering approach to visualize the separation of the resulting clusters. A soft clustering approach, such as fuzzy c-means, could be used for identifying road traffic images with multiple weather condition cluster memberships. This would allow us to evaluate our representation on particular images that could share multiple weather conditions between clusters.

References

[1] T. Alexandropoulos, V. Loumos, and E. Kayafas, "Feature extraction on highways under time-varying illumination conditions," *Journal of Intelligent Transportation Systems*, vol. 13, no. 2, pp. 85–96, 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/15472450902858392

[2] J. Rhinelander, M. Kallada, and P. Lingras, "Visual predictions of traffic conditions," in *Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, D. Barbosa and E. Milios, Eds., vol. 9091. Springer International Publishing, 2015, pp. 122–129.

[3] A. Rampun, H. Strange, and R. Zwiggelaar, "Texture segmentation using different orientations of glcm features," in *Proceedings of the 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications*, ser. MIRAGE '13. New York, NY, USA: ACM, 2013, pp. 17:1–17:8.

[4] A. Baraldi and F. Parmiggiani, "An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 2, pp. 293–304, Mar 1995.

[5] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, 1967, pp. 281–297.

[7] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

Voice and Unvoiced Classification Using Fuzzy Logic

Mohammed Algabri[1], Mansour Alsulaiman[2], Ghulam Muhammad[2], Mohammed Zakariah[2], Mohamed Bencherif[2],Zulfiqar Ali[2]

[1] Computer Science Department , King Saud University, Riyadh, Saudi Arabia

[2] Computer Engineering Department , King Saud University, Riyadh, Saudi Arabia

{malgabri, msuliman, ghulam, mzakariah, mbencherif, zuali}@ksu.edu.sa

Abstract - *In this paper, we proposed a system for automatic classification of speech. A speech signal contains three different regions voiced, unvoiced and silence. In the proposed system, Zero-Crossing rate and short term energy are used in a fuzzy logic control this classification. Arabic digits of the KSU database is used to test our proposed method. The proposed method achieves 2.5 % error between human classification and automatic classification using this method.*

Keywords: *Voice detection, fuzzy logic controller, zero-crossing, and short term energy.*

1 Introduction

A speech signal contains three different regions –voiced, unvoiced and silence. It is to determine in speech recognition system it is very helpful to discriminate between speech and background noise. This well give better result in a shorter time. The discrimination between voiced and unvoiced speech is also very helpful in designing high performance speech recognition system. Voiced phonemes are produced due to the vibration of the vocal folds, while unvoiced phonemes are made without vocal cord vibration [1]. Researchers have done significant efforts during the recent years for classification of speech into voice and unvoiced detection [2-12]. Based on the statistical and non-statistical techniques and pattern recognition approach a decision is made on the given segment of speech signal to classify it as a voice or unvoiced speech [6,7,8, and 9].The speech segment is classified into voice and unvoiced based on the acoustic features and pattern recognition techniques[12]. Speech is said to be intelligible if two third of it is voiced. Non periodic speech is called unvoiced, sounds like random, when consonants and spoken air is passed through a narrow constraints of the vocal tract causing unvoiced speech. Identification and extraction of voice speech is done because of its periodic nature. The Energy and Zero-Crossing rate are two important parameter are used to classify voice and unvoiced speech. They are used as front-end processing in automatic speech recognition system. The energy present in the signal spectrum and its frequency is indicated by the zero crossing counts. The low zero crossing count shows that the air is flowing in periodic form due to the extraction of vocal cords producing voiced speech [13]. Voice activity detector (VAD) is an algorithm implemented to detect the presence and absence of the speech. Numerous techniques are applied to the art of VAD. The very common features used in the detection process in the early VAD algorithm stage were short-time energy, zero-crossing rate and linear prediction coefficients [14]. Cepstral coefficients [15], spectral entropy [16], a least-square periodicity measure [17], wavelet transform coefficients [18] are examples of recently proposed VAD features. Because of the variety and varying nature of human speech and also because of the background noise none of the above technique proves to be perfect for all the applications. The decision to classify a segment into voiced or unvoiced is based on the values of the energy and zero-crossing. Since these values are not precise it will be helpful to use fuzzy logic. So in this paper, we present a method that classify the speech based on fuzzy logic of short-time energy and zero-crossing. This paper will be structured as follows: Section 2 presents the literature review. Section 3 gives the details of our proposed method. Sections 4 give the experimental results. Section 5 concludes the paper and suggests some future works.

2 Literature Review

In [19] a speech recognition system based on zero-crossing rate and energy was presented. It used a vocabulary of ten Cantonese digits, and achieved a recognition accuracy of 97.2 percent has been achieved. Speech recognition using zero-crossing feature is presented in [20]. The zero-crossing features are extracted while speaking are done in the training phase, then stored in the database. Using the same technique the features for testing data are extracted and compared with the template in the database during the recognition phase. A VAD algorithm is presented in [21] for speech signal with very low signal to noise ration (SNR). The short-term energy of the speech signal is viewed as positive frequency of the magnitude spectrum of a minimum phase signal then the group delay function is computed for this signal. The speech regions of the signal are identified by well-defined peaks, while the non-speech regions are identified by well-defined valleys. Speech/ silence classification algorithm based on energy is proposed in [22]. The algorithm is able to track non-stationary signals and calculate the value of threshold using adaptive scaling parameter. Computed threshold can be obtained using maximum and minimum values of short-term energy. Voiced /Unvoiced classification based on clustering is developed in [23]. They used cepstral peak, zero crossing rate, and autocorrelation function peak of short time segments of speech by using some clustering methods. They achieved

good results for classification of voice and unvoiced segments of speech. Zero crossing and short-term energy function are used for VAD algorithm for speech recognition applications in [24]. The method is labeling the speech samples based on if they are silence, voiced or unvoiced speech. Zero crossing rate and short term energy of speech are extensively used to detect the endpoints of an utterance.

3 Proposed method

Energy for unvoiced speech is significantly smaller than for voiced speech hence the short time energy can be used to distinguish voice and unvoiced speech. The short time energy can also be used to distinguish speech from silence. But the use of short time energy is not sufficient alone hence it is used coupled with the zero crossing in the classification of speech. Hence in [25] the authors present an algorithm for a heuristically developed method that cooperate zero crossing and energy. The authors states that "voice speech should characterized by relatively high energy and relatively low zero crossing rate, while unvoiced speech will have relatively high zero crossing rate and relatively low energy". They also state "we have not said what we mean by high and low values of short time zero crossing rate, and it is really not possible to be precise". Hence we see that this is a problem that fit to be solved by fuzzy logic.

In this paper, we propose a method to classify the speech into silence voiced and unvoiced detection using short time energy and zero-crossing in a fuzzy logic system. Figure 1 presents the fuzzy logic system, where the zero-crossing (ZC) and short term energy (STE) are the inputs of fuzzy logic control and the (Detect) is the output. The signal was segmented into frames with duration 10 ms. Then, hamming window was applied to prevent discontinuity. The mean of zero-crossing and short term energy was computed for each frame and set as inputs to fuzzy logic control. Voice, unvoiced and silence detection is an output of fuzzy logic control.

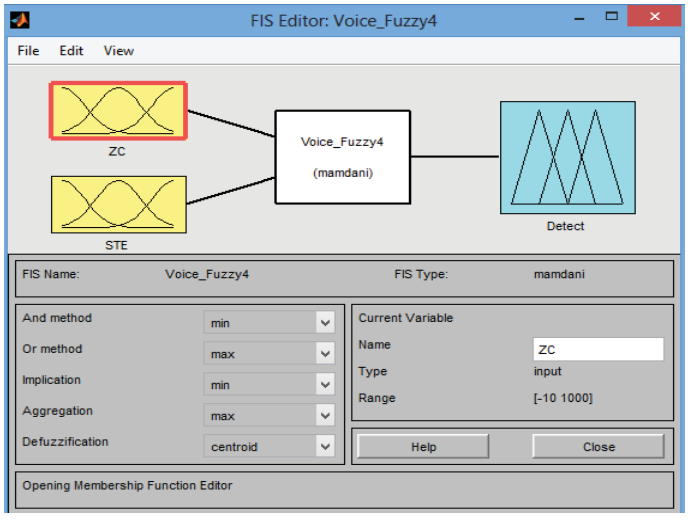


Figure 1: Fuzzy Logic System.

To define the membership function of each linguistic variables we used three membership functions for inputs (ZC and STE). The notation for Zero-crossing is Low, Mid and High as shown in Figure 2. The notation of short term energy (STE) id Low, Mid and High as shown in Figure 3. The notation of fuzzy output (Detect) is Sil, Unvoice and Voice as shown in Figure 4. The membership functions were tuned after many experiment manually to achieve good results.

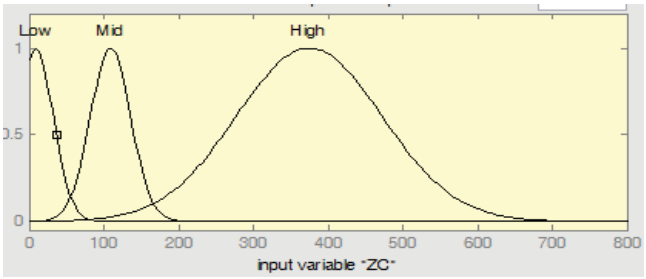


Figure 2: Membership Function of ZC.

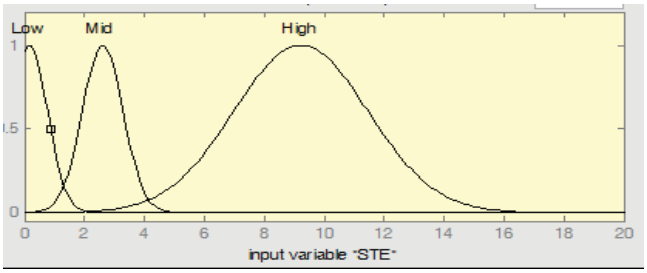


Figure 3: Membership Function of STE.

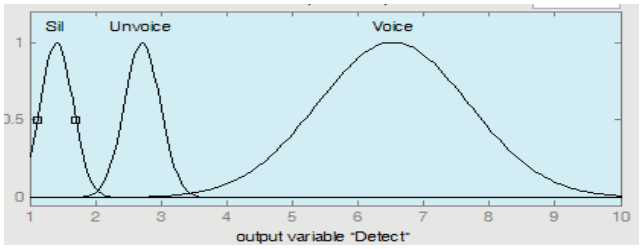


Figure 4: Membership Function of Detect.

The fuzzy rules of the control are defined in Figure 5. They were selected based on the study in [26].

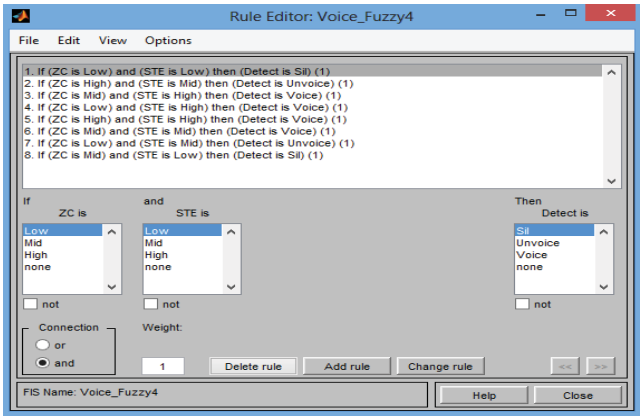


Figure 5: Rule based of fuzzy logic.

4 Experimental Results

The experimental results are obtained using the speech of Arabic digit of a King Saud Speech Database [27]. The database has 327 speakers and is rich in different aspects and different nationalities. We performed many experiments to segment the Arabic digits. The classification result was excellent. As an example Figure 6 shows the original wav file that contains the utterances of Arabic digits from zero to nine. We applied the proposed method on this file to classify it into voiced, unvoiced and silence. The output of our proposed method is shown in Figure 7.

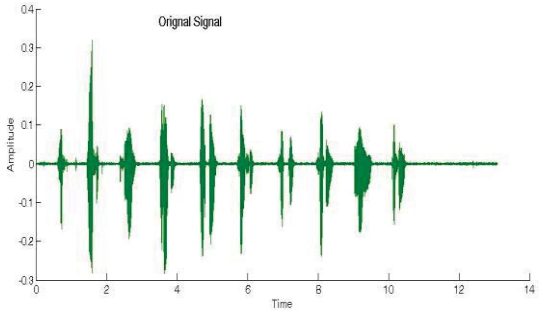


Figure 6: Original wav file.

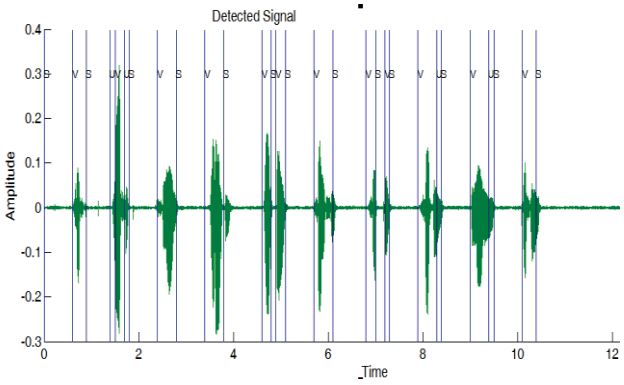


Figure 7: Segmented wav file [S-Silence-V-Voice, U-Unvoiced].

In order to evaluate the proposed method for voice and unvoiced detection using fuzzy logic, we used the error difference between voice detected by human and voice detected using this method. The error was calculated using equation (1).

$$Error = \frac{|TH - TC|}{TH} \times 100 \quad (1)$$

Where TH is the manually voice frame length detected by human as shown in table 1, and TC is the length of voice detected using the proposed approach.

Table 1: Automatic and manual voice segmentation length of digit file.

	Voice frame length detected (ms)												total
TH	0.262	0.26	0.416	0.43	0.19	0.21	0.397	0.17	0.097	0.416	0.445	0.33	3.623
TC	0.3	0.2	0.4	0.4	0.2	0.2	0.4	0.2	0.1	0.4	0.4	0.3	3.5
TH-TC	-0.038	0.06	0.016	0.03	-0.01	0.01	-0.003	-0.03	-0.003	0.016	0.045	0.03	0.093
Error	14.50	23.08	3.85	6.98	5.26	4.76	0.76	17.65	3.09	3.85	10.11	9.09	**2.57**

So from the results in table 1, we can calculated the overall detection error using equation in (2).

$$OverallError = \frac{\sum_{i=1}^{n} |TH - TC|}{\sum_{i=1}^{n} TH} \times 100 \quad (2)$$

Where n is a number of voice frames detected. So from the table 1 above the overall error segmentation is approximate **2.5%**.

As another example, Figure 8 shows the original wav file that contains the speech of ten phonetic distinctive words in Arabic language. We applied the proposed method on this file to classify it into voiced, unvoiced and silence. The output of our proposed method is shown in Figure 9.

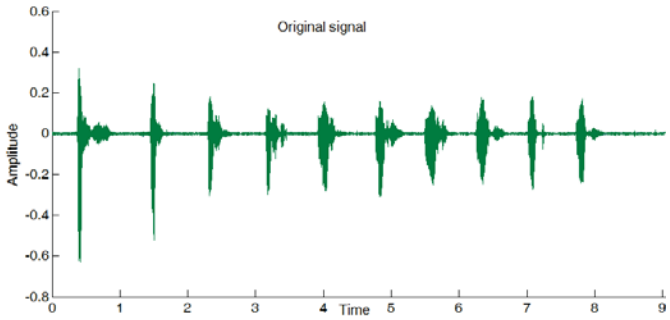


Figure 8: Original wav file.

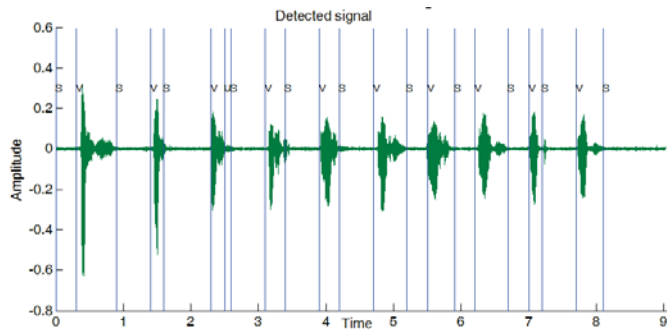


Figure 9: Segmented wav file [S-Silence-V-Voice, U-Unvoiced].

The calculated error based on equations (1, 2) between human segmented and our proposed method shown in Table 2. The overall error segmentation is approximate **1.84%.**

Table 2: Automatic and manual voice segmentation length of words file.

	Voice frame length detected (ms)										total
TH	0.52	0.24	0.25	0.314	0.294	0.447	0.387	0.453	0.25	0.38	3.535
TC	0.6	0.2	0.2	0.3	0.3	0.5	0.4	0.5	0.2	0.4	3.6
TH-TC	-0.08	0.04	0.05	0.014	-0.006	-0.053	-0.013	-0.047	0.05	-0.02	-0.065
Error	15.38	16.67	20.00	4.46	2.04	11.86	3.36	10.38	20.00	5.26	**1.84**

5 Conclusion

Silence/unvoiced/voiced classification using fuzzy logic of Arabic speech was proposed in this paper. Zero crossing and short term energy were used as features. Fuzzy logic controller used to distinguish three categorize of speech (Silence, Voice and unvoiced). The experiments were performed on Arabic KSU database in MATALB environment and their results showed that the proposed method successfully classified the speech.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED2474-02).

References

[1] D. Jurafsky, J. H. Martin "Speech and Language Processing", Publisher Pearson.

[2] E. Fisher, J. Tabrikian and S. Dubnov, "Generalized Li-kelihood Ratio Test for Voiced-Unvoiced Decision in Noisy Speech Using the Harmonic Model," IEEE Trans-actions on Audio, Speech, and Language Processing, Vol. 14, No. 2, 2006, pp. 502-510. doi:10.1109/TSA.2005.857806

[3] Y. Qi and B. R. Hunt, "Voiced-Unvoiced-Silence Classi-fications of Speech Using Hybrid Features and a Network Classifier," IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 2, 2002, pp. 250-255. doi:10.1109/89.222883

[4] B. Atal and L. Rabiner, "A Pattern Recognition Approach to Voicedunvoiced-Silence Classification with Applica-tions to Speech Recognition," IEEE Transactions on Ac- oustics, Speech and Signal Processing, Vol. 24, No. 3, 2003, pp. 201-212. doi:10.1109/TASSP.1976.1162800

[5] F. Y. Qi and C. C. Bao, "A Method for Voiced/Unvoiced/Silence Classification of Speech with Noise Using SVM," Acta Electronica Sinica, Vol. 34, No. 4, 2006, pp. 605-611.

[6] B. Atal, and L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," IEEE Trans. On ASSP, vol. ASSP-24, pp. 201-212, 1976.

[7] S. Ahmadi, and A.S. Spanias, "Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm," IEEE Trans. Speech Audio Processing, vol. 7 No. 3, pp. 333-338, 1999.

[8] Y. Qi, and B.R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," IEEE Trans. Speech Audio Processing, vol. 1 No. 2, pp. 250-255, 1993.

[9] L. Siegel, "A Procedure for using Pattern Classification Techniques to obtain a Voiced/Unvoiced Classifier", IEEE Trans. on ASSP, vol. ASSP-27, pp. 83- 88, 1979.

[10] T.L. Burrows, "Speech Processing with Linear and Neural Network Models", Ph.D. thesis, Cambridge University Engineering Department, U.K., 1996.

[11] D.G. Childers, M. Hahn, and J.N. Larar, "Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech," IEEE Trans. on ASSP, vol. 37 No. 11, pp. 1771-1774, 1989.

[12] Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno "Robust voiced/unvoiced classification using novel features and Gaussian Mixture model", Speech Processing Lab., ECE Dept., Temple University, 1947 N 12th St., Philadelphia, PA 19122-6077, USA.

[13] JaberMarvan, "Voice Activity detection Method and Apparatus for voiced/unvoiced decision and Pitch Estimation in a Noisy speech feature extraction", 08/23/2007, United States Patent 20070198251.

[14] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced- silence classi_cation with applications to speech recognition, " IEEE Trans. Acoustics, Speech, Signal Processing, vol. 24, pp. 201-212, June 1976.

[15] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstralfea-tures, " in Proc. of IEEE Region 10 Annual Conf. Speech and Image Technologies for Computing and Telecommunications, (Beijing), pp. 321-324, Oct. 1993.

[16] S. A. McClellan and J. D. Gibson, "Spectral entropy: An alternative indicator for rate allocation, " in IEEE Int. Conf. on Acoustics, Speech, Signal Processing, (Adelaide, Australia), pp. 201-204, Apr. 1994.

[17] R. Tucker, "Voice activity detection using a periodicity measure, "IEE Proc.-I, vol. 139, pp. 377-380, Aug. 1992.

[18] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform, " in Proc. IEEE Workshop on Speech Coding for Telecommunications, (Pocono Manor, PN), pp. 99-100, Sept. 1997.

[19] Lau, Yiu-Kei, and Chok-Ki Chan. "Speech recognition based on zero crossing rate and energy." IEEE transactions on acoustics, speech, and signal processing 33, no. 1 (1985): 320-323.

[20] Aye, Yin. "Speech Recognition Using Zero-Crossing Features." InElectronic Computer Technology, 2009 International Conference on, pp. 689-692. IEEE, 2009.

[21] Hari Krishnan P, S., R. Padmanabhan, and Hema A. Murthy. "Robust voice activity detection using group delay functions." In Industrial Technology, 2006. ICIT 2006. IEEE International Conference on, pp. 2603-2607. IEEE, 2006.

[22] Sakhnov, Kirill, Ekaterina Verteletskaya, and Boris Simak. "Approach for energy-based voice detector with adaptive scaling factor." IAENG International Journal of Computer Science 36, no. 4 (2009): 394.
[23] Radmard, Mojtaba, Mahdi Hadavi, and Mohammad Mahdi Nayebi. "A New Method of Voiced/Unvoiced Classification Based on Clustering." Journal of Signal and Information Processing 2, no. 04 (2011): 336.
[24] Lokhande, N. N., D. S. Nehe, and P. S. Vikhe. "Voice activity detection algorithm for speech recognition applications." In IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia, no. 6, pp. 1-4. 2012.
[25] Rabiner, Lawrence R., and Ronald W. Schafer. "Theory and application of digital speech processing.", Person , 2011.
[26] Greenwood, M., & Kinghorn, A. (1999). Suving: Automatic silence/unvoiced/voiced classification of speech. Undergraduate Coursework, Department of Computer Science, the University of Sheffield, UK.
[27] Alsulaiman, M., Ali, Z., Muhammed, G., Bencherif, M., & Mahmood, A. (2013). KSU Speech Database: Text Selection, Recording and Verification. In Modelling Symposium (EMS), 2013 European (pp. 237-242). IEEE.

Peripheral Detail-based Edge Preserving Image Interpolation Scheme

Abhinash Kumar Jha[1], **Vishwajeet Narwal**[1], **Rohit Patwa**[1] **and Gerald Schaefer**[2]
[1]The LNM Institute of Information Technology, Jaipur, Indiat
[2]Department of Computer Science, Loughborough University, U.K.

Abstract— *Interpolation is a common technique for predicting unknown data points within the range of discrete known data points. In image processing and analysis applications, image interpolation is employed for changing image resolution but also for tasks like image rotation and other transformations. Classical image interpolation algorithms such as nearest neighbour, bilinear and bicubic interpolation are simple and fast. However, due to the high distortion along image details such as edges, the resulting images are often low in quality. In order to reduce these distortions and preserve fine image details, we propose an edge preserving interpolation algorithm in this paper. Our proposed algorithm extracts and recognises the direction of edges in an image. Based on the extracted information about localisation of edges, interpolated pixels are either replicated or predicted from known neighbourhood pixels. Experimental results confirm our approach to give good image quality, outperforming various other interpolation algorithms.*

Keywords: Image interpolation, edge preservation, image detail, upsampling.

1. Introduction

Image interpolation is a technique often employed for obtaining a high resolution (HR) image from a low resolution (LR) image. For such an upsampling process, image interpolation can also be interpreted as increasing the pixel density per unit area of an image to obtain a HR version from a LR counterpart. As illustrated in Fig. 1, if an image is assumed to be a sequence $f(x_K)$ of length N and this sequence is downsampled by a factor of 2 to obtain another sequence $g(x_n)$ of length $N/2$, then interpolation will yield a sequence $l(x_k)$ that should approximate $f(x_k)$.

Among the most popular image interpolation algorithms are nearest neighbour interpolation [1], bilinear interpolation [2], bicubic interpolation [3], and spline interpolation [4]. In the nearest neighbour method, the value of a new pixel simply takes on the value of the nearest original pixel. In bilinear interpolation, the interpolated value is calculated as the weighted average of its neighbours. While this takes only four pixels into account, bicubic interpolation considers 16 pixels and cubic fitting functions, whereas spline interpolation is based on a special type of piecewise

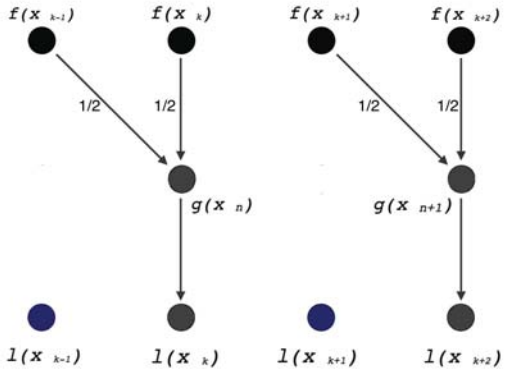


Fig. 1: Signal downsampling and interpolation.

polynomial interpolant. However, these algorithms, despite being fast and simple, suffer from resulting in relatively low image quality due to aliasing effects, pixelated curves and blurring in particular of edges and other fine image detail.

Interpolation techniques focussing on preserving edge structures in an image are computationally costly compared to the above methods. The new edge directed interpolation algorithm (NEDI) by Li *et al.* [5] predicts unknown pixels by estimating the covariance of the HR image from the covariance of the LR image. Tang *et al.* [6] used an autoregressive method based on Gauss-Seidel optimisation, while Jaiswal *et al.* [7] presented an algorithm based on least squares estimation. Kumar *et al.*'s approach [8] is less complex and based on a set of predictors, however these predictors do not adapt well to all types of images. Chan *et al.* [9] suggested a content adaptive interpolation scheme that is computationally simple but typically fails to give sufficiently high image quality. Jha *et al.* [10] proposed an edge preserving interpolation technique based on inverse gradient weights but proper estimation of these weights is difficult.

In this paper, we propose an image interpolation algorithm that preserves the edges in the HR image by extracting and estimating them from its LR counterpart. Our proposed method is simple and is shown to give good image quality, outperforming various existing interpolation algorithms.

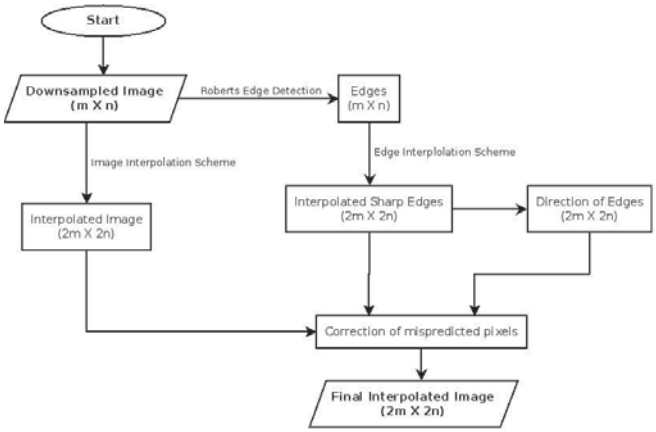


Fig. 2: Overview of our proposed interpolation algorithm.

2. Proposed Algorithm

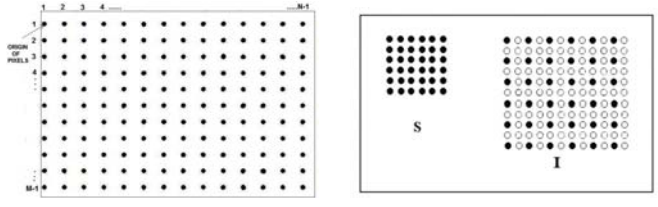
In the following, $S(i,j)$ denotes a pixel at the i-th row and j-th column of an image S, while the sign convention for spatial relative pixel locations is given in Fig. 3(a).

Our proposed algorithm consists of an initial phase of extraction and localisation of image details, which is followed by an interpolation and correction phase based on the extracted details. An overview of our approach is depicted in Fig. 2; in the following we explain the individual steps in detail.

2.1 Extraction and Analysis of Image Details

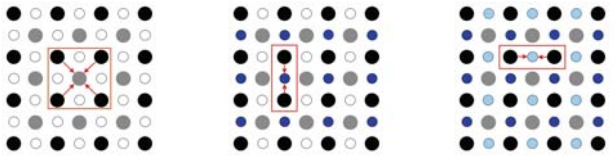
Natural images comprise mostly smooth variations, with fine details typically being represented as sharp edges. In our proposed algorithm, image edges are extracted from the low resolution image using the Roberts filter [11], with the value of the threshold set so that we obtain only sharp edges.

Analysis of the edges involves interpolating the edges so as to make a prediction about intensity transitions. For this, the direction of edges provides important information so that proper localisation of edges in the high resolution (i.e., the interpolated image) can be deduced.



(a) (b)

Fig. 3: (a) Sign convention used for determining the current spatial location of pixels (b) Transition of low resolution image to odd-odd locations of high resolution image.



(a) (b) (c)

Fig. 4: The boxed region shows the pixels used for the prediction of edges at (a) even-even, (b) even-odd and (c) odd-even locations.

2.1.1 Interpolation of Edges

The extracted edges are then interpolated using the following procedure:

1) **Expansion:** The first step of the interpolation focusses on the expansion of a source LR grid e into an HR grid E (i.e., $M \times N$ grid into a grid of size $2M \times 2N$) . The transition from source LR image to interpolated HR image follows the mapping $M : e \mapsto E$ governed by
$$M(e(i,j)) = E(2i-1, 2j-1). \quad (1)$$
The effect of this initial transition can be seen in Fig. 3(b). The remaining $\frac{3}{4}$ pixels of E are filled depending on the type of edge present there.
2) **Edge Enlargement:** This prediction for unknown pixels is performed as detailed in Algorithm 1 using the illustrations in Fig. 4.

for *edge pixels* $e_{i,j} \in E$ **do**
  if i *is* ***even*** *and* j *is* ***even*** *as in Fig. 4(a)* **then**
    if $e_{i-1,j-1} = 1$ *and* $e_{i+1,j+1} = 1$ **then**
      $e_{i,j} = 1$
    else if $e_{i+1,j-1} = 1$ *and* $e_{i-1,j+1} = 1$ **then**
      $e_{i,j} = 1$
    else
      $e_{i,j} = 0$
    end
  else if i *is* ***odd*** *and* j *is* ***even*** *as in Fig. 4(b)* **then**
    if $e_{i,j-1} = 1$ *and* $e_{i,j+1} = 1$ **then**
      $e_{i,j} = 1$
    else
      $e_{i,j} = 0$
    end
  else if i *is* ***even*** *and* j *is* ***odd*** *as in Fig. 4(c)* **then**
    if $e_{i-1,j} = 1$ *and* $e_{i+1,j} = 1$ **then**
      $e_{i,j} = 1$
    else
      $e_{i,j} = 0$
    end
end

Algorithm 1: Enlargement of Edges.

2.1.2 Direction of edges

For prediction of unknown pixels according to image edges, we must know the direction of the edges so that pixels are predicted in a way that preserves the edges. Directions of edges are found and stored following Algorithm 2.

2.2 Interpolation of LR image to HR image.

The technique used for inerpolation in this phase is inspired by that of [12]. Pixels are classified into four categories: odd-odd, even-even, odd-even and even-odd pixels, based on their current spatial locations. Odd-odd pixels were already predicted through the initial expansion of the LR image as explained above. The remaining pixels are predicted using three prediction schemes based on their spatial locations, as follows:

2.2.1 Even-even position pixels:

The prediction of pixels at even-even locations is performed by

$$I(2i,2j) = \frac{1}{2(n+1)} \sum_{y=1}^{n<4} \sum_{z=1}^{n<4} I(2(i-y)\pm 1, 2(j-z)\pm 1). \tag{2}$$

The larger n, the more computationally demanding the algorithm would be; therefore only close neighbours ($n < 4$) are considered.

for *pixels* $e_{i,j} \in E$ **do**
  if $e_{i,j-1}$ *=1 and* $e_{i,j+1}$ *= 1* **then**
    Edge is **Horizontal**
  else if $e_{i-1,j}$ *=1 and* $e_{i+1,j}$ *=1* **then**
    Edge is **Vertical**
  else if $e_{i-1,j+1}$ *=1 and* $e_{i+1,j-1}$ *= 1* **then**
    Edge is at an angle of $\mathbf{45}^\circ$ from horizontal
  else if $e_{i-1,j-1}$ *= 1 and* $e_{i+1,j+1}$ *= 1* **then**
    Edge is at an angle of $\mathbf{135}^\circ$ from horizontal
  else
    if $e_{i,j-1}$ *=1 or* $e_{i,j+1}$ *= 1* **then**
      Edge is **Horizontal**
    else if $e_{i-1,j}$ *=1 or* $e_{i+1,j}$ *=1* **then**
      Edge is **Vertical**
    else if $e_{i-1,j+1}$ *=1 or* $e_{i+1,j-1}$ *= 1* **then**
      Edge is at an angle of $\mathbf{45}^\circ$ from horizontal
    else if $e_{i-1,j-1}$ *= 1 or* $e_{i+1,j+1}$ *= 1* **then**
      Edge is at an angle of $\mathbf{135}^\circ$ from horizontal
    else
      $e_{i,j} = 0$
    end
  end
end

Algorithm 2: Extraction of Direction of Edges.

2.2.2 Even-odd position pixels:

Pixels at even-odd locations are predicted as

$$I(2i,2j-1) = \frac{1}{n+1} \sum_{z=1}^{n<4} I(2i\pm z, 2j-1). \tag{3}$$

2.2.3 Odd-even position pixels:

Finally, pixels at odd-even image locations are determined as

$$I(2i,2j-1) = \frac{1}{n+1} \sum_{z=1}^{n<4} I(2i-1, 2j\pm z). \tag{4}$$

2.3 Correction of Mispredicted Pixels.

Both edge information and interpolated pixels are then used in a correction step. For this purpose, pixels are classified into two groups:

- Pixels in smooth area (correctly predicted pixels)
- Pixels on edges (incorrectly predicted pixels)

Correction of the mispredicted pixels is performed as described in Algorithm 3.

for *unknown* ***pixel*** $p_{i,j} \in I$ *and* ***edge*** $e_{i,j} \in E$ **do**
  if $e_{i,j}$ *= 1* **then**
    if *Edge is horizontal* **then**
      $p_{i,j} = (p_{i,j+1} + p_{i,j+1})/2$
    else if *Edge is vertical* **then**
      $p_{i,j} = (p_{i-1,j} + p_{i+1,j})/2$
    else if *Edge is at angle of* 45° *from horizontal* **then**
      $p_{i,j} = (p_{i-1,j+1} + p_{i+1,j-1})/2$
    else
      (Edge is at angle of 135° *from horizontal)*
      $p_{i,j} = (p_{i-1,j-1} + p_{i+1,j+1})/2$
    end
  else
    Leave pixel unchanged.
  end
end

Algorithm 3: Correction of mispredicted pixels.

3. Experimental results

We performed an extensive set of experiments to appropriately test our proposed algorithm. As is common, we evaluate the performance of an interpolation technique using the peak signal to noise ratio (PSNR), defined as

$$\mathrm{PSNR}(O,I) = 10\log_{10}\frac{255^2}{\mathrm{MSE}(O,I)} \tag{5}$$

between the original image O and the interpolated image I generated from the downsampled version S of O, where $\mathrm{MSE}(O,I)$ is the mean squared error between O and I.

Fig. 5: Dataset of 30 images used in the experiments.

To put our obtained results into context, we have implemented six other image interpolation algorithms, namely bicubic interpolation, content adaptive interpolation (CAI) [9], context-based image dependent (CBID) interpolation [13], context-based image independent (CBII) interpolation [13], efficient edge preserving interpolation (EEPI) [10], and interlinear interpolation (ILI) [12].

Our test database comprises a total of 30 images [13], all with a resolution of 512×512 pixels. Fig. 5 shows the image dataset, while Table 1 gives the PSNR results for all evaluated algorithms.

As can be seen from Table 1, our proposed algorithm provides the highest average PSNR and hence the best image quality. This clearly demonstrates that our approach is able to deliver superior interpolation performance compared to other methods.

A visual comparision of the quality of the different methods is presented in Fig. 6 which shows an original image together with interpolated versions obtained by all tested methods, and confirms that our presented approach is capable of correctly preserving edges and of providing superior image quality.

The edge preservation ability is further illustrated in Fig. 7 which shows image edges extracted from interpolated images obtained by running the different interpolation algorithms. Again, it is clear to observe the edges are indeed better preserved in our presented approach.

Fig. 6: (a)Visual comparison of sample image interpolated by the different methods

Table 1: PSNR results for all interpolation techniques on the images of Fig. 5. Bolded results indicate the best performing algorithm.

image	bicubic	CAI	CBID	CBII	EEPI	ILI	Proposed
1	27.38	29.13	29.49	29.42	29.61	29.77	**29.81**
2	29.34	31.08	31.04	30.97	31.18	31.38	**31.39**
3	26.02	27.26	27.83	27.67	27.88	28.13	**28.18**
4	27.74	29.99	32.41	29.96	29.65	32.39	**32.99**
5	26.25	27.56	28.17	28.11	28.31	28.33	**28.35**
6	26.15	27.94	30.18	28.08	27.62	29.89	**30.69**
7	23.58	24.36	25.07	25.04	25.16	25.35	**25.36**
8	22.54	24.12	24.51	24.38	24.68	24.91	**24.97**
9	19.52	20.41	21.11	21.06	21.17	21.27	**21.30**
10	25.81	27.2	27.96	27.85	28.16	**28.24**	28.21
11	31.97	34.84	34.45	34.55	34.74	34.86	**34.90**
12	30.81	32.69	32.84	32.83	32.90	32.96	**32.97**
13	30.82	32.70	32.89	32.9	33.02	33.04	**33.07**
14	26.03	27.52	27.84	27.64	27.81	28.07	**28.11**
15	24.47	25.95	26.61	26.51	26.67	26.73	**26.74**
16	26.42	28.14	28.13	28.04	28.21	28.35	**28.37**
17	24.45	25.8	26.11	26.05	26.22	**26.39**	26.31
18	27.24	28.65	28.82	28.78	28.90	**29.21**	29.05
19	29.40	31.14	31.07	30.98	31.26	31.47	**31.49**
20	30.82	32.48	32.75	32.69	32.70	32.73	**32.76**
21	33.40	36.04	35.72	35.65	35.86	**36.13**	**36.19**
22	26.76	28.67	28.8	28.78	28.95	**29.07**	**29.07**
23	25.60	26.63	27.03	26.85	26.96	**27.16**	**27.16**
24	28.50	31.63	31.54	31.33	31.91	**32.17**	32.13
25	17.09	18.12	18.51	18.43	18.39	18.45	**18.46**
26	26.77	28.72	29.68	29.27	28.91	29.51	**29.98**
27	23.99	25.41	25.82	25.65	25.66	25.83	**25.89**
28	21.35	22.25	23.15	22.62	22.62	23.14	**23.15**
29	27.73	30.29	30.24	30.18	30.42	30.57	**30.59**
30	27.59	29.13	29.25	29.14	29.21	29.28	**29.34**
average	26.68	28.43	28.58	28.66	28.77	28.81	**28.89**

4. Conclusions

In this paper, we have presented an effective edge preserving image interpolation algorithm. Our proposed method extracts image details and predicts unknown pixels so that edges are accurately preserved by correcting mispredicted pixels based on the extracted image (edge) information. Experimental results confirm that our proposed approach provides an effective image interpolation algorithm and outperforms various existing interpolation schemes from the literature.

References

[1] J. A. Parker, R. V. Kenyon, and D. Troxel, "Comparison of interpolating methods for image resampling," *Medical Imaging, IEEE Transactions on*, vol. 2, no. 1, pp. 31–39, 1983.

[2] G. Vendroux and W. Knauss, "Submicron deformation field measurements: Part 2. improved digital image correlation," *Experimental Mechanics*, vol. 38, no. 2, pp. 86–92, 1998.

[3] F. N. Fritsch and R. Carlson, "Monotonicity preserving bicubic interpolation: A progress report," *Computer Aided Geometric Design*, vol. 2, no. 1, pp. 117–121, 1985.

[4] H. S. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 6, pp. 508–517, 1978.

[5] X. Li and M. T. Orchard, "New edge-directed interpolation," *Image Processing, IEEE Transactions on*, vol. 10, no. 10, pp. 1521–1527, 2001.

[6] K. Tang, O. C. Au, L. Fang, Z. Yu, and Y. Guo, "Image interpolation using autoregressive model and gauss-seidel optimization," in *Image and Graphics (ICIG), 2011 Sixth International Conference on*. IEEE, 2011, pp. 66–69.

[7] S. P. Jaiswal, V. Jakhetiya, and A. K. Tiwari, "An efficient image interpolation algorithm based upon the switching and self learned characteristics for natural images," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 861–864.

[8] A. Kumar, N. Agarwal, J. Bhadviya, and A. Tiwari, "An efficient 2-d jacobian iteration modeling for image interpolation," in *Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on*. IEEE, Dec 2012, pp. 977–980.

[9] T.-W. Chan, O. C. Au, T.-S. Chong, and W.-S. Chau, "A novel content-adaptive interpolation," in *Circuits and Systems, 2005. ISCAS 2005.*

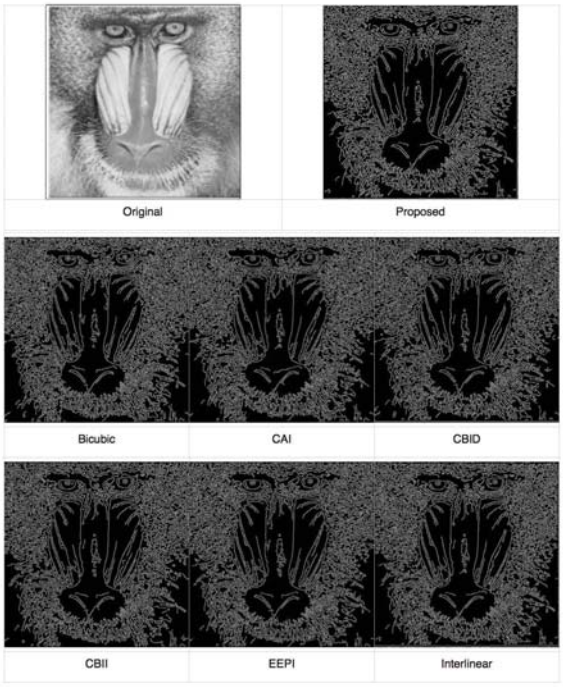


Fig. 7: Comparision of edge preservation ability by the different methods

IEEE International Symposium on. IEEE, 2005, pp. 6260–6263.

[10] A. K. Jha, A. Kumar, G. Schaefer, and M. R. A. Ahad, "An efficient edge preserving image interpolation algorithm," in *International Conference on Informatics, Electronics & Vision*. IEEE, 2014, pp. 1–4.

[11] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *Solid-State Circuits, IEEE Journal of*, vol. 23, no. 2, pp. 358–367, 1988.

[12] A. K. Jha, V. Narwal, Y. Gupta, and A. Kumar, "Interlinear image interpolation scheme for real time application," in *and innovation in Technology,2014, Indicon Emerging Trends, 2014 IEEE India Conference on*. IEEE, 2014.

[13] S. Prasad Jaiswal, V. Jakhetiya, A. Kumar, and A. K. Tiwari, "A low complex context adaptive image interpolation algorithm for real-time applications," in *Instrumentation and Measurement Technology Conference (I2MTC), 2012 IEEE International*. IEEE, 2012, pp. 969–972.

A Simple Approach for Abandoned Object Detection

Fahian Shahriar Mahin[1], Md. Nazmul Islam[1], Gerald Schaefer[2], and Md. Atiqur Rahman Ahad[1]
[1]Department of Electrical and Electronic Engineering, University of Dhaka, Bangladesh
[2]Department of Computer Science, Loughborough University, U.K.

Abstract - *In this paper, we implement a low-cost solution to detect abandoned and removed objects. To detect abandoned and stolen objects, the focus is to determine static regions that have recently changed in the scene by performing background subtraction. The time and presence of static objects, which may be either abandoned or stolen, are marked on the video feed and may be used to alert security personnel. Our system can detect abandoned objects and is capable of performing this in real-time. No special sensors are required and the results are shown to be satisfactory.*

Keywords: video processing, object detection, abandoned object detection, background subtraction.

1 Introduction

Low-cost and low-power video surveillance systems based on networks of wireless video sensors come with many advantages including flexibility, quick deployment and the ability of providing accurate and real-time visual data. Energy autonomy and efficiency of the implemented algorithms are undoubtedly the primary design challenges to be addressed on systems subject to low computational capabilities and memory constraints.

The demand for reliable surveillance systems is increasing, especially for mass transit and public areas such as airports, railway and subway stations, sports and event venues. For this reason, video surveillance systems that, through the analysis of video sequences, perform automatic detection of security-related events or aid human personnel in monitoring a place are gaining increasing interest. An important aspect for current video surveillance systems is the capability of reliably detecting common events such as abandoned or removed object within a scene. Typical scenarios are for example the detection of unattended packages in a railway station or an airport [1,2] and the detection of stolen objects in a museum [3].

In this paper, we implement a low cost-solution to detect abandoned or unattended objects both in real-time and from recorded video feeds. We thus present a video surveillance system that is able to detect objects abandoned or removed in indoor environments.

The remainder of the paper is structured as follows: Section 2 presents the background and related works. Section 3 presents our system in detail. In Section 4, we demonstrate experimental results. Finally, we conclude the paper in Section 5.

2 Related Works

The increasing need for automated surveillance of indoor environments, such as airports, warehouses, production plants, etc., has stimulated the development of intelligent systems based on mobile systems.

There exists a significant body of research addressing the task of robustly identifying abandoned baggage in public spaces [4]. Most authors treat detection of abandoned (or left) objects, especially luggage, as a task of static object detection, with the application of tracking [5,6], or without any tracking [7,8]. Tian *et al.* [6] presented a framework to detect abandoned and removed scene objects based on background subtraction and foreground analysis, combined with tracking output to reduce false positives. A comparative evaluation of stationary foreground detection algorithms based on background subtraction is given in [9].

There has been some attempt at human activity recognition and association to scene objects. The most widely used datasets with which to evaluate approaches to abandoned bag detection have been from PETS [10] and from the UK Home Office i-LIDS [11]. The dataset provided for the PETS2006 challenge consists of 7 multi-camera scenarios involving an increasing number of people and passers-by. Most of the submissions to PETS2006 were based on background subtraction combined with a blob tracker [12-16].

Lv *et al.* [17] rely on a more realistic human model by incorporating a human detector. Two approaches beyond classical methods to blob tracking and split-track analysis were developed by Arsic *et al.* [18] and Dalley *et al.* [19]. They used a temporal-joint-boost algorithm for each blob being tracked to classify it into a number of actions, namely a person-walking, not moving, a person picking-up/leaving a bag, or an abandoned bag. They incorporated temporal features as optical flow vectors and motion energy into the classification process over some temporal window. Ardo and Astrom [20] used an HMM to improve the temporal consistency of the tracking, and demonstrated the appropriateness of the development of HMM.

The PETS 2007 challenge is very simple to reduce to comparing the distance of a bag to its owner (abandoned bag, theft) or measuring the time for which a person stays in the scene (i.e., loitering). i-LIDS [11] is developed by the UK Home Office in order to evaluate video-based detection systems to meet government requirements. This library includes an abandoned luggage dataset including several challenges of single instances of left luggage on a metro platform in the presence of passing passengers and trains.

Though the dataset is useful for evaluating detection algorithms, it has some limitations. For example, it is monocular and it lacks any behavioural interactions.

3 System Development

In this paper, we present a simple and low-cost solution for detecting abandoned or stolen objects in videos. Figure 1 gives an overview of our proposed approach.

Our system proceeds in the following stages:

1. Store background image;
2. Segmentation using background subtraction;
3. Blob analysis;
4. Detect stationary objects.

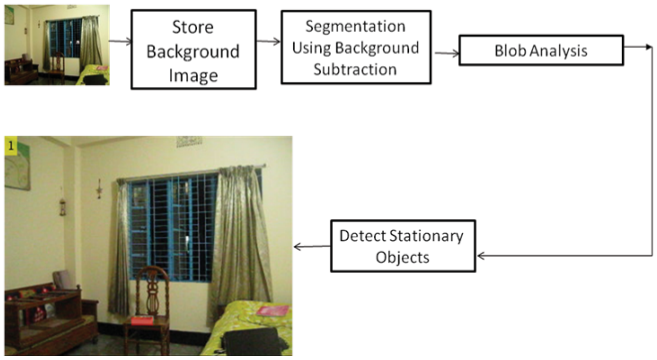


Figure 1: Flow diagram of the detection system

Our method involves the following steps:

- Get input video and select a region of interest (ROI);
- Perform video segmentation using background subtraction;
- Calculate object statistics using blob analysis;
- Detect stationary objects based on their area and centroid statistics;
- Show output video with boundary box around the detected objects.

In the following we describe each step in more detail.

Store Background Image

This program uses the first frame of the video as the background. The RGB image frames are converted to YCbCr colour space. Later for the background subtraction operation both intensity and color information are used [22]. For Blob Analysis it is converted to binary images using a ThresholdScaleFactor. To fill the gaps of the object, morphological closing is performed using neighbouring pixels as the structural element [23].

Segmentation Using Background Subtraction

Inside this subsystem, the Luminance Segmentation and Color Segmentation subsystems perform background subtraction using the intensity and color data respectively [24,25]. The program combines these two segmentation results using a binary OR operator**.**

Blob Analysis

The Blob Analysis block computes statistics of the objects present in the scene. It computes statistics for labelled regions, including area, centroid, count, maximum number of tracks, and feeds them to the core object detection function subsystem.

Abandoned Object Tracker

The Abandoned Object Tracker subsystem [21] uses the object statistics to determine which objects are stationary. This function gets the count, area, centroid, etc. from the Blob Analysis, checks whether the area and centroid of the blob has changed less than a ratio, and then determines which objects are stationary.

4 Experimental Results

In this paper, we present our dataset that is used for various experimentations. Table 1 shows some descriptions of the dataset based on the actions with object(s) and the corresponding object(s). Book, ball, bag, and box are taken as objects in this dataset.

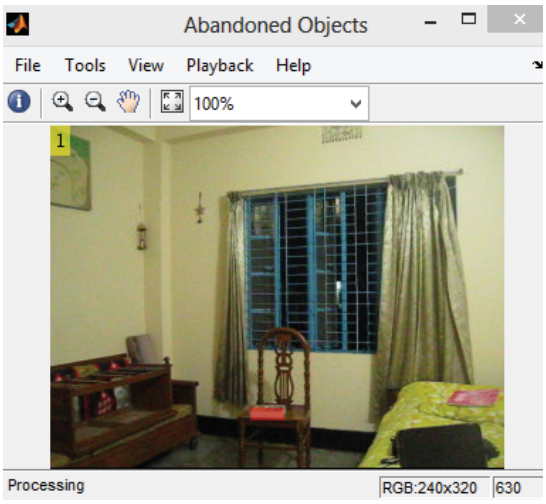
Table 1**:** Properties of our abandoned object detection dataset.

Action with object	**Object(s)**
Placement at chair	book
Placement at table	ball
Placed book, later a box	book, box
Placed bag at table	bag
First book then bag placed at a table	book, bag
Book removed from original place and placed box nearby	book, bag
Swapped book with box	book, box
Placed book and box at two ends of a table, one is outside ROI	book, box

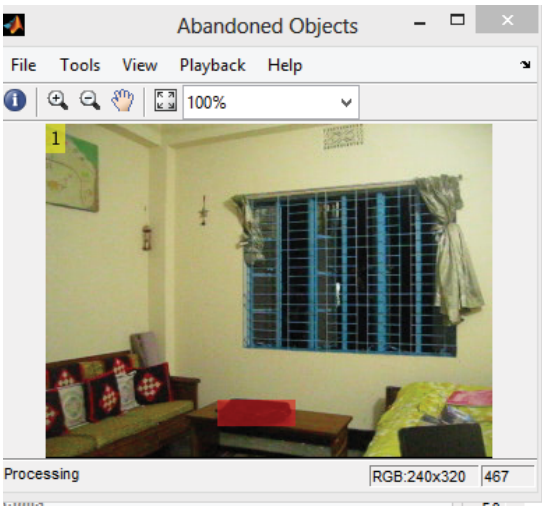
We present some experimental results. Figure 2 shows an example of detection of an abandoned object. After processing, in the Abandoned Object window, the book is marked as red which means that this object was not in the first frame (as seen in Figure 2(a)). As soon as the book was placed and the subject left the scene, the threshold window starts showing the book as a white blob which means it has subtracted the current frame from the background frame, which is the first frame of the video and found that blob. The All Objects window shows the region of interest area marked by a yellow line.

Figure 3 shows another example. Here, the object is placed on a table that is marked in a red box. In the final image, it is marked properly and detected.

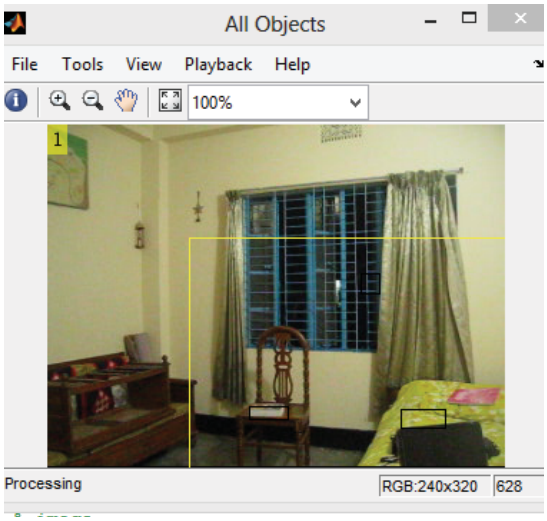
We thus demonstrate the performance of our system on real-time video feeds (using a simple webcam for testing the system).



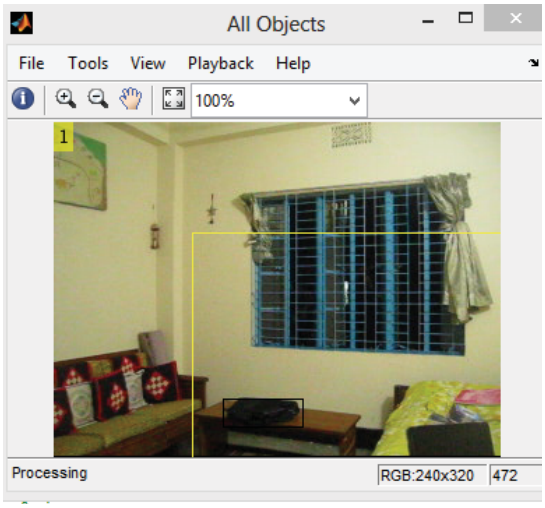
(a) Abandoned object in a chair.



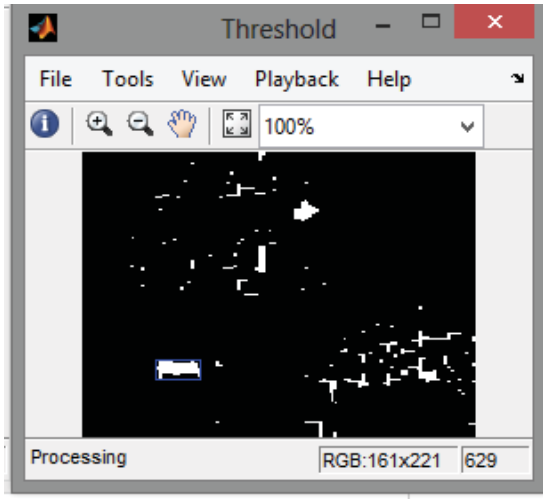
(a) Abandoned object on a table.



(b) Location of objects

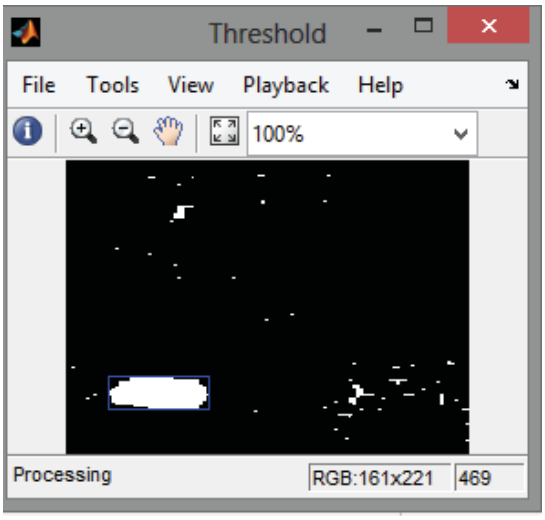


(b) Location of the objects.



(c) Location of objects in a bounding box in the scene.

Figure 2: An example of the system during operation (the object as shown in (a) is in a chair, marked by a rectangular red box).



(c) Detected results.

Figure 3: Another example, for a different object placed on the table.

5 Conclusions

In this paper, we have implemented an abandoned object detection system based on relatively simple operations that is thus able to run in real-time. We also prepared our own data set, which is open for anyone to test their respective algorithms. In our experiments, we used a normal laptop, and no additional sensors were employed. We used various objects to ensure it can detect all types of objects.

An indoor environment was used to test the detection capabilities with a variety of lighting conditions. The system successfully detected objects also under low light condition, and was able to distinguish between replaced objects.

Acknowledgements

The authors are thankful to the Center for Natural Science and Engineering Research (CNSER) for its support for this research.

References

[1] S. Lu, J. Zhang, and D. Feng, "A knowledge-based approach for detecting unattended packages in surveillance video," IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp. 2-3, 2006.

[2] S. Lim and L. Davis, "A one-threshold algorithm for detecting abandoned packages under severe occlusions using a single camera," University of Maryland, Tech. Rep. CS-TR-4784, 2006.

[3] S. Ferrando, G. Gera, and C. Regazzoni, "Classification of unattended and stolen objects in video-surveillance system," IEEE Int. Conf. on Advanced Video & Signal Based Surveillance, 2006.

[4] J. Ferryman, D. Hogg, J. Sochman, A. Behera, J. Rodriguez-Serrano, S. Worgan, L. Li, "Robust abandoned object detection integrating wide area visual surveillance and social context", Pattern Recognition Letters, 34, 7, pp. 789-798, 2013.

[5] W. Hassan, P. Birch, B. Mitra, N. Bangalore, R. Young, and C. Chatwin, "Illumination invariant stationary object detection", IET Computer Vision, 7, 1, pp. 1-8, 2013.

[6] Y. Tian, R. Schmidt Feris, H. Liu, A. Hampapur, and M-T. Sun. "Robust detection of abandoned and removed objects in complex surveillance videos.", IEEE Transactions on Systems, Man, and Cybernetics Part C, 41, 5, pp. 565-576, 2011.

[7] R. Evangelio, M. Patzold, and T. Sikora, "A system for automatic and interactive detection of static objects", in IEEE Workshop on Person-Oriented Vision, pp. 27-32, 2011.

[8] F. Porikli, "Detection of temporarily static regions by processing video at different frame rates", in Advanced Video and Signal Based Surveillance, pp. 236-241, 2007.

[9] A. Bayona, J.C. SanMiguel, and J.M. Martínez, "Stationary foreground detection using background subtraction and temporal difference in video surveillance", in Intl. Conf. Image Processing, pp. 4657-4660, 2010.

[10] PETS2007, http://www.cvg.rdg.ac.uk/PETS2007/

[11] http://tna.europarchive.org/20100413151426/sciencea ndresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/index.html

[12] E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier, "Fall detection with multiple cameras: An occlusion-resistant method based on 3-d silhouette vertical distribution", IEEE Trans. Information Technology in Biomedicine, 15, 2, pp. 290-300, 2011.

[13] S. Guler, J.A. Silverstein, and I.H. Pushee, "Stationary objects in multiple object tracking", in IEEE Conf. Advanced Video and Signal Based Surveillance, pp. 248-253, 2007.

[14] X. Liu, P.H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications", in Advanced Video and Signal Based Surveillance, pp. 306-311, 2005.

[15] A. Bayona, J.C. SanMiguel, and J.M. Martinez, "Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques", in Advanced Video and Signal Based Surveillance, pp. 25-30, 2009.

[16] M. Smith and K. Takeo, "Video skimming and characterization through the combination of image and language understanding", in Content-Based Access of Image & Video Databases, pp. 61-70, 1998.

[17] L.C. Hardin, and N. Larry, "Electro-optical range finding and speed detection system", U.S. Patent 5,642,299, issued June 24, 1997.

[18] D. Arsic, H. Martin, S. Björn, and R. Gerhard, "Multi-camera person tracking and left luggage detection applying homographic transformation", in IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2007.

[19] G. Dalley, W. Xiaogang, and W. Eric, "Event detection using an attention-based tracker", in 10th International Workshop on Performance Evaluation for Tracking and Surveillance, pp. 71-79. 2007.

[20] H. Ardo and A. Kalle, "Multi sensor loitering detection using online viterbi", in IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp-34-36, 2007.

[21] http://www.mathworks.com/help/vision/examples/aba ndoned-object-detection-1.html

[22] R. Gonzalez, R. Woods, and S. Eddins, "Digital Image Processing Using MATLAB", Pearson Prentice Hall, 2nd edition, 2003.

[23] O. Marques, "Practical Image and Video Processing Using MATLAB", Wiley, 2011.

[24] M.A.R. Ahad, "Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding", Springer, 2011.

[25] M.A.R. Ahad, "Motion History Images for Action Recognition and Understanding", Springer, 2013.

SESSION

POSTER PAPERS

Chair(s)

TBA

Fusion method of visible and infrared images in foggy environment

Jun-Woo Son[1], **Hyuk-Ju Kwon**[1], **Tae-Eun Shim**[2], **Young-Choon Kim**[3], **Sang-Ho Ahn**[4], **Kyu-Ik Sohng**[1]
[1]School of Electronics Engineering, Kyungpook National University, Republic of Korea
[2]Dept. of IT, Gyeongbuk Provincial College, Republic of Korea
[3]Dept. of Information Communication & Security, Youngdong University, Republic of Korea
[4]Dept. of Electronics Engineering, Inje University, Republic of Korea

Abstract – *In this paper, we propose a fusion method of visible and infrared images in foggy condition. Using Dark Channel Prior, we remove the fog each of the visible and infrared image. After that, the images are fused by the fusion rule. The fusion method selects a maximum gray level value taken from visible or infrared images. As a result, a fused image has good image quality in spite of a fog condition. So, a proposed method is expected to be applied to the observation system.*

Keywords: Visible and infrared image fusion, Dark Channel Prior, fog removal, atmospheric transmittance

1 Introduction

"This paper is being submitted as a poster". An image fusion based on multi-sensors is a way to get more information combining the image signals of the various wavelength-bands. Especially, the fusion of visible and infrared (IR) images has been studied in the perspective of military observation. A visible image represents the reflected light from an object. Accordingly, an object in a visible image can be recognized under good lighting condition such as daylight. On the other hand, an IR image, which describes the thermal radiance of an object, can be used in the nighttime because lighting condition has little effect on object recognition in the IR image. However, in foggy weather, the atmospheric transmittance is largely decreased because reflected light in the visible wavelength-band and radiance in the IR wavelength band are absorbed and scattered. Consequently, the identification of objects is difficult in the visible and IR images under cloudy and foggy weather conditions.

In this paper, we propose a fusion method of visible and IR images to remove the fog in two images. Especially, using the Dark Channel Prior from a visible image, a fog is removed not only from a visible image but also from an IR image. Then, the simple fusion method selecting a maximum gray level value is conducted with two fog-free images. As a result, a fused image has good image quality in spite of a fog condition.

2 Fusion of visible and IR images in fog condition

2.1 Fog removal in a visible image using Dark Channel Prior

The optical model for a visible image in a fog condition can be expressed as

$$I_v(x) = J_v(x)t + A(1-t) \tag{1}$$

where I_v is the visual image in a fog condition, J_v is the fog-free image, t is the global atmospheric transmittance, and A is the atmospheric light. In the case of homogenous atmosphere, an atmospheric transmittance is exponentially attenuated with the distance from the camera to the object.

In order to yield the fog-free image, J_v, He et al.[1] estimated local atmospheric transmittances, $t(x)$, using the estimation of Dark Channel Prior, DCP, as follows;

$$t(x) = 1 - w\mathrm{DCP}\left(\frac{I_v(x)}{A}\right) \tag{2}$$

$$\mathrm{DCP}(I(x)) = \min_{y\in\Omega(x)}\left(\min_{c\in\{r,g,b\}} I^c(y)\right) \tag{3}$$

where w is a weight factor equal to 0.95 and $\Omega(x)$ is a local patch centered at pixel location x. I^c is a color channel of an arbitrary image, I. Finally, a fog-free image obtained from Eq. (1) and estimated local atmospheric transmittances can be expressed as

$$J_v(x) = \frac{I_v(x) - A}{\max(t(x), t_o)} + A\,. \tag{4}$$

Because the fog-free image can be prone to noise when the $t(x)$ is close to zero, the $t(x)$ is restricted by a lower bound, $t_0 = 0.1$. In addition, the atmospheric light, A, is approximated by the top 0.1 percent of brightest pixels in the result of DCP using I_v.

2.2 Fog removal in IR image

An IR image, I_{IR}, in a fog condition can be expressed as

$$I_{IR}(x) = J_{IR}(x)t(x) \tag{5}$$

where J_{IR} is the fog-free IR image and $t(x)$ the atmospheric transmittance. In comparison to Eq. (1), an IR image in a fog condition is mostly influenced by radiance of an object, therefore, the atmospheric light, A, can be discarded. Consequently, a fog-free IR image, J_{IR}, which is motivated by Eq. (4), can be expressed as

$$J_{IR}(x) = \frac{I_{IR}(x)}{\max(t(x), t_o)}. \tag{6}$$

The atmospheric transmittance, $t(x)$, is calculated from Eq. (2).

2.3 Fusion method of visible and IR images

Figure 1 indicates the block diagram of proposed fusion method. Processing step of the proposed method is as follows.

(1) Estimating the atmospheric light A and the atmospheric transmittance $t(x)$ of the visible image using Dark Channel Prior in RGB color space.
(2) Refining the atmospheric transmittance map using bilateral filter to prevent the block noise in next the fog removal process.[2]
(3) Computing the fog removal images for visible and IR images using Eq. (4) and (6), respectively.
(4) Fusing visible and IR images with the fusion rule selecting a higher gray level value of the two images.

3 Simulation

We perform a simulation in order to investigate the validity of proposed method. We use 640×480 pixels visible camera and IR image is obtained by IR camera (FLIR, TAU640) having a resolution of 640×480 pixels of LWIR wavelength-band. The used original foggy images are shown in Figs. 2(a) and 2(c). Figures 2(b) and 2(d) are the fog removed images. The fusion results of the visible and IR images are shown in Fig. 3. As shown in Fig. 3 (white circles), the edge and contrast around buildings are enhanced.

4 Conclusions

In this paper, we propose a fusion method of visible and IR images. Especially, a fog removal method using estimated atmospheric transmittance from visible image is applied to IR image. The refinement method of estimated atmospheric transmittance is used by the bilateral filter, and the fusion method selects a maximum gray level value taken from visible or IR images. The proposed method, in foggy condition, can identify the object easily through the fog removal image processing and the fusion, and is expected to be applied to the observation system.

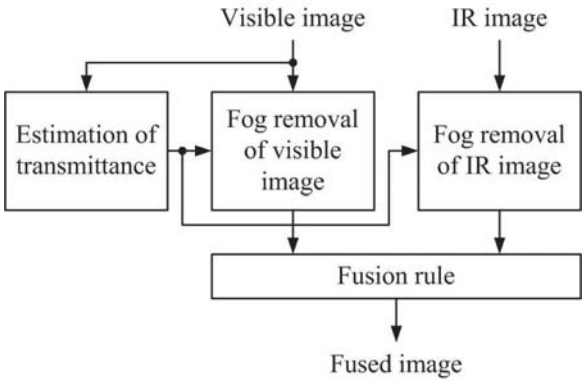


Fig. 1. Block diagram of proposed fusion method.

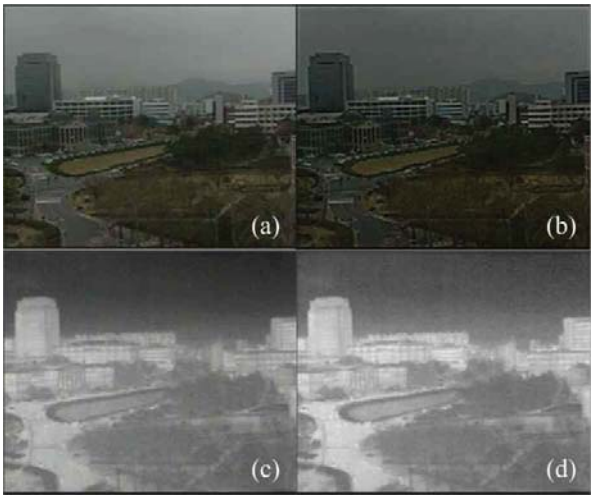


Fig. 2. The original images; (a) visible and (c) IR images, and result images of fog removal; (b) visible and (d) IR images.

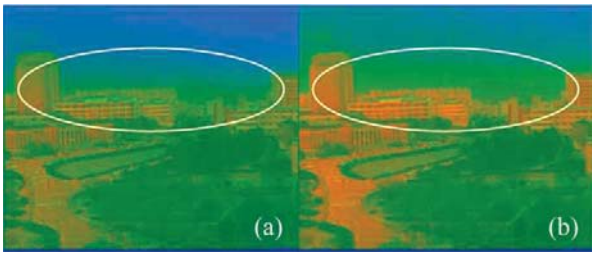


Fig. 3. Result images of the fusion with visible and IR images; (a) without fog removal in IR image and (b) with fog removal in IR image.

5 References

[1] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341-2353, 2011.
[2] J. H. Kim and C. S. Kim, "Hierarchical Haze Removal Using Dark Channel Prior," *The Transactions of Korean Institute of Electrical Engineers*, vol. 59, no. 2, pp. 457-464, 2010.

An Analysis of Trihalomethanes Formation in Chlorinated Water by using Ensembles of Regression Models

Sami M. Halawani
Faculty of Computing and Information Technology, King Abdul Aziz University, Jeddah, Saudi Arabia

Abstract - *Trihalomethanes (TMHs) formation in chlorinated water is an important problem for human health. The correct prediction of TMHs is necessary to tackle this problem. An ensemble consists of many accurate and diverse models. In this paper, we study the use of regression model ensembles for the prediction of formation of THMs in chlorinated water. The results suggest that regression model ensembles can be useful for the problem.*

Keywords: Trihalomethanes, Chlorinated Water, Ensembles, Regression.

1 Introduction

Chlorine is commonly used to disinfect the water. Chlorine reacts with organic substances present in waters and form THMs. High level of THMs is harmful for the human health. Different models [1] have been proposed to study the THMs formation.

Singh et al. [1] presented a dataset to study THMs formation. This dataset was generated from the Gomti river water, Lucknow city, India. This dataset has five input variables; chlorine dose/dissolvedorganic carbon ratio, pH, temperature, bromide concentration, and reaction time. Each data point also has the THMs concentration. The task is to predict the THMs concentration on the basis of 5 input variables. They divided the data set into two subsets; training dataset (45 data points) and validation dataset (18 data points).

Regression trees [2], support vector machines [3] and neural networks [3] are popular regression models. Ensembles are combinations of multiple base models [3]; the final result depends on the combined outputs of individual models. Ensembles [4, 5] have shown to produce better results than single models, provided the models are *accurate* and *diverse*. Bagging [6] is a common method to produce ensembles. In this paper, we will use ensembles of different models created by Bagging [6] to model the THMs formation for the dataset [1].

.

2 Methods

Regression is a supervised learning problem in which the task is to predict the output with given inputs. The output is a real number. There are many regression methods such that linear regression methods, regression trees, support vector machines and neural networks which are used extensively for regression problems. Regression trees, support vector machines and neural networks can handle both linear and nonlinear regression problems.

2.1 Regression trees

Regression trees [2, 10] have shown excellent performance for many regression problems. The key principle of regression trees is to recursively binary split the training data. The split tries to minimize the variance of the output in the left and right parts of training data points corresponding to that split. The splitting process stops when the data points in a node have same output or some predefined stopping criterion is reached. The mean of these points in a node is taken as the output of the leave.

2.2 Neural Networks

Neural networks [3] are very popular for regression problems. A neural network consists of one or many interconnected artificial neurons. They can approximate highly nonlinear regression functions. However, neural networks are slow and the selection of proper hidden layers is a problem.

2.3 Support vector machines

Support vector machines [3] are useful for regression problems. Similar to neural networks they can approximate highly nonlinear problems. However, the selection of a proper kernel is very important for the accuracy of the support vector machines.

2.4 Ensembles

Ensembles [4, 5] are a combination of multiple base models; the final classification depends on the combined outputs of individual models. Classifier ensembles have shown to produce better results than individual models provided the models are accurate and diverse.

Several methods have been proposed to build ensembles. In these methods, randomization is introduced to build diverse model. Bagging [6] and Boosting [7] introduce randomization by manipulating the training data supplied to each model. Ho [8] proposed Random Subspaces that selects random subsets of input features for training an ensemble. Breiman [9] combined Random Subspaces technique with Bagging to create Random Forests. To build a tree, it uses a bootstrap replica of the training sample, then during the tree growing phase, at each node the optimal split is selected from a random subset of size K of candidate features, then during the tree growing phase, at each node the optimal split is selected from a random subset of size K of candidate features.

2.5 Bagging

Bagging [6] generates different bootstrap training datasets from the original training dataset and uses each of them to train one of the models in the ensemble. For example, to create a training set of *N* data points, it selects one point from the training dataset, *N* times without replacement. Each point has equal probability of selection. In one training dataset, some of the points get selected more than once, whereas some of them are not selected at all. Different training datasets are created by this process. When different models of the ensemble are trained on different training datasets, diverse models are created. Bagging does more to reduce the variance part of the error of the base model than the bias part of the error.

3 Results and Discussion

We used WEKA software [10] for our experiments. The size of the ensemble was 10. The kernel for support vector machine was RBF kernel. All other parameters were default parameters. The model were made on the training dataset and tested on the validation dataset. The results are calculated in Root Mean Square Error (RMSE). Low values of RMSE suggest better regression models. The results for single model were presented in Table 1 and the results for ensembles are presented in Table 2.

Table 1. Root Mean square Error (RMSE) for single regression models

Name of the Model	REP tree	Neural Networks	Support Vector Machine
RMSE	15.20	84.87	27.73

Table 2. Root Mean square Error (RMSE) for ensembles (the size = 10) of regression models.

Name of the Model	REP tree	Neural Networks	Support Vector Machine
RMSE	13.91	68.91	24.64

We summarize our results in following points.

1- Ensembles of regression models performed better than single models

2- Trees regression models were the best for the problem. support vector machines performed better than the neural networks.

4 Conclusions

The high level of THMs is harmful for the human health. The proper prediction model is necessary to handle this problem. In this paper, we show the effectiveness of ensembles of regression models to predict the formation of THMs in chlorinated water.

5 References

[1] Kunwar P. Singh ; Premanjali Rai ; Priyanka Pandey ; Sarita Sinha, , *Environ Sci Pollut Res*, 2012, 19:113–127.
[2] Breiman, L. Friedman, J., Olshen R. and Stone C.. *Classification and Regression Trees*. CA: Wadsworth International Group, 1984.
[3] Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. 1st Edn., Springer-Verlag, New York, 2006.
[4] Dietterich. T. G., Ensemble Methods in Machine Learning. In *Proc. Of Conf. Multiple Classifier Systems*, volume 1857, pages 1–15, 2000.
[5] Kuncheva, L. I., *Combining Pattern Classifiers: Methods and Algorithms*: Wiley-Interscience, 2004.
[6] Breiman, L., *Mach. Learn.* 1996, 24(2):123–140.
[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, pp. 119-139, 1997.
[8] Ho. T. K., The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
[9] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
[10] Hall, M, ; Frank, E. ; Holmes, G. ; Pfahringer, B. Reutemann, P. ; Witten, I. H. ; The WEKA Data Mining Software: An Update; SIGKDD *Explorations*, 2009, Volume 11, Issue 1.

SESSION

LATE BREAKING PAPERS: IMAGE PROCESSING, COMPUTER VISION, AND PATTERN RECOGNITION

Chair(s)

TBA

Efficient Integration of Image Encryption with Compression Using Optimal Entropy Coder

Lei Huang
Department of Electrical Engineering and Computer Science
Loyola Marymount University, Los Angeles, CA, USA

Abstract – *A novel scheme of efficiently performing both image encryption and compression is proposed in this paper. The proposed scheme is a combination of two existing approaches, namely, selective encryption and in-compression encryption schemes. The image is compressed using an adaptively optimized entropy codec. The coding table of the entropy codec, i.e., the Huffman table, is selected as the only subset of data to be encrypted by state-of-art encryption algorithms. The proposed scheme can achieve both high computational efficiency and high level of security. Moreover, it can be adapted to different multimedia compression systems with similar entropy codecs.*

Keywords: Image Encryption; Image Compression; Selective Encryption; Multiple Huffman Tables; Optimal Entropy Coder

1 Introduction

Recently there has been a rapid increase in multimedia applications with image data shared among computer networks and mobile networks. It has been widely recognized that image data should be compressed for efficient utilization of storage and/or transmission resources. Advanced image compression schemes, such as JPEG and JPEG-2000 have been developed, standardized and employed in multimedia networking applications. Meanwhile, with greater concerns about data security in the new era of cloud computing, there have been increased awareness of and demand for privacy preservation in many emerging applications, such as social networking, mobile commerce, electronic medical record systems, etc. It can be foreseen that image encryption, as one of the major instrumentations for confidentiality protection of digital image information, is expected to become more critical to the success of those applications. Therefore, it is imperative to have digital image data both compressed and encrypted for future multimedia applications, which require both data efficiency and data security.

How to efficiently perform both compression and encryption on digital imagery data is the particular problem we are addressing in this paper. The most straightforward solution is to cascade the two processes, utilizing existing image compression and data encryption techniques. Since most image compression schemes reduce the redundancy of data by removing the inherent correlation among data, while most encryption schemes aims at randomization of the original data to hide the real information represented, it is intuitive that encrypted image data will be hard to be compressed due to little correlation among encrypted data. Therefore, the encryption process should follow the compression process. Nevertheless, it is not an efficient solution, because even after compression, image data usually is still much larger in size than other types such as text data. Most existing encryption methods involve extensive computational complexity, thus become bottleneck of the process, especially in real-time applications.

To address the above problem, researchers have proposed some selective encryption schemes (also referred to as partial encryption schemes) [1-6], which encrypted only a portion of data in the compressed image using conventional bit stream encryption algorithms, such as AES, IDEA, and RSA. The rest of the data was not encrypted at all or was encrypted using some simple scrambling schemes. Thus, encryption and overall computational efficiency can be improved.

A noticeable advantage of selective encryption was its fully compliance with current image compression standards, as well as existing state-of-art data encryption primitives. As a result, it has been incorporated into the newest image compression standard, JPEG-2000, as a solution to data confidentiality in Part 8 (known as JPSEC)[6] of the standard. However, selective encryption had limitations in some applications as examined in [8].

An alternative approach was to integrate the compression and encryption as one process, by adding a randomization mechanism into an existing compression system. This approach was also referred to as in-compression encryption. The most notable in-compression image encryption schemes were based on Multiple Huffman Tables (MHT)[8-10]. By alternating through multiple Huffman tables according to a randomly generated key stream, these schemes turned the built-in entropy coder in most image compression systems into a cryptographic cipher. To achieve a higher security level, more Huffman tables should be used to increase the model space. The major drawback of this scheme was that the multiple Huffman tables used in the coding process had to be included in the final bit stream, thus degraded the compression ratio.

In this paper, we propose a new efficient and secure scheme that combines the above-mentioned two approaches, namely, selective encryption and MHT scheme. In our scheme, we use one optimized Huffman table that is specifically tailored to the statistical model of each image to reduce the resulted data size overhead. Then, we select this optimized Huffman table as the portion of data to be encrypted by strong bit stream encryption primitives. Consequently, our proposed scheme achieved higher security level with smaller data size than the MHT schemes, while avoided the limitations of previous selective encryptions.

The rest of this paper is organized as followed. In Section II, we review related work of selective image encryption and MHT schemes in the literature. In Section III, we describe and analyze our proposed scheme in detail. In Section IV, experimental results are presented and discussed to show the efficiency of our proposed scheme. Finally, conclusion and future work will be provided in Section V.

2 Related work

In this section, we review two categories of recently developed image encryption and compression schemes, namely, selective encryption and in-compression encryption using MHT, which are closely related to our proposed scheme.

2.1 Selective Encryption

The critical issue in designing an effective selective encryption scheme is how to choose the portion of data that needs to be encrypted. Since most multimedia compression codecs take advantage of energy concentration property in the frequency domain of multimedia data by either DCT or wavelet transform, previous selective encryption schemes focused on selecting the most energy concentrated coefficients in the transform domain. Tang [1] proposed to encrypt DC coefficients with DES and scramble AC coefficients with block permutation. Shi and Bhargava [2] proposed to encrypt the single most significant bit, i.e., the sign bit of every DC coefficient in JPEG. Although the reconstructed image suffered severe visual degradation without correct knowledge of the most energy concentrated DC coefficients, it was still possible to reveal some information of the image [8] even without any knowledge of DC coefficients. As pointed out in [8], energy concentration was not necessarily related to the intelligibility of image data. Therefore, these selective encryption methods were not suitable for applications that require protection of image intelligibility. Another problem with selection of certain coefficients to encrypt was the encryption of those coefficients resulted in the disruption of statistical model of data, thus degraded the performance of the later entropy coding stage. Cheng and Li [4] proposed a selective encryption scheme for zero-tree wavelet based image compression scheme by encrypting the significance information of the highest two levels of coefficients. This scheme was successful in protecting the intelligibility of image data. However, it was not suitable for compression algorithms with scalar quantization and entropy coding, such as JPEG/MPEG.

2.2 In-Compression Encryption with Multiple Huffman Tables

The integration of image compression with image encryption using MHT was first proposed by Wu and Kuo [7][8]. The basic idea was to hide the statistical model information that was critical to entropy coding, by increasing the model space from one standard Huffman table to a variety of its permutations. However, the original schemes were vulnerable to some attacks because there were only a limited number of fixed Huffman tables used. To increase its security level, enhanced MHT schemes using further randomization in the bit stream such as key hopping or random rotation of partitioned bit streams have been proposed in [9] and [10]. More recently, El-said et al. [11] proposed the OMHT method, which used more number of dynamically generated Huffman tables from training of subsets of image dataset. Since the Huffman tables were dynamically generated from different subsets of images to be compressed and encrypted, the compression performance was improved than previous MHT schemes using fixed tables. At the same time, using more number of tables increased the statistical model space and hence the security level of the MHT scheme. However, the major drawback of this scheme was the large computational overhead resulted from training multiple subsets of images. Moreover, even though the trained Huffman tables were optimized for each subset of image set, the compression performance was not necessarily optimized for each individual image.

3 Proposed integrated image encryption and compression scheme

Based on the two approaches described in the previous section, we propose a new scheme integrating image encryption and compression. Similar to the MHT approaches, it utilizes the important entropy codec that are commonly found in image and other multimedia data compression standards. However, our proposed scheme differs from the previous MHT approaches in the number of Huffman table used for each individual image. Only one particular Huffman table is optimally generated from the statistical model of the specific image to be compressed and encrypted. Therefore, the compression performance of the entropy coder can be optimized for each specific image. Since each table is generated from one single image data, the computational cost of the training process can be greatly reduced compared to that in the OMHT scheme.

On the other hand, we propose to apply selective decryption concept to achieve high level of security. We select the customized Huffman table as the subset of data to be encrypted, based on the following observations. First, the size of data representing a Huffman table is significantly smaller

than that of resulting compressed image bit stream, therefore high computational efficiency can be achieved in the encryption process. Second, the Huffman table is critical to the intelligibility of the original image. It has been shown in [12] that decoding a Huffman coded bit stream without correct knowledge of the Huffman table is extremely difficult. Third, encryption of Huffman table data does not affect the entire compression process, or the resulting compressed bit streams. Therefore, it is highly compliant with existing image and video compression standards such as JPEG and MPEG.

A conceptual illustration of our proposed scheme, when applied to a typical image compression system consisting of transformation, quantization, run-length coding and entropy coding stages, can be found in Fig. 1. Note that our proposed scheme is flexible in two folds. First, it can be applied to any image compression scheme with entropy coding. In fact, it can even be extended to multimedia compression schemes with entropy codec for other types of media data, such as speech, audio, and video. With an additional block of adaptive Huffman table generation, the optimal Huffman table can be generated. Consequently, the subsequent entropy codec becomes adaptive to the current input data. Second, the optimal Huffman table can be encrypted by any state-of-art encryption algorithm according to the desired security level of the application.

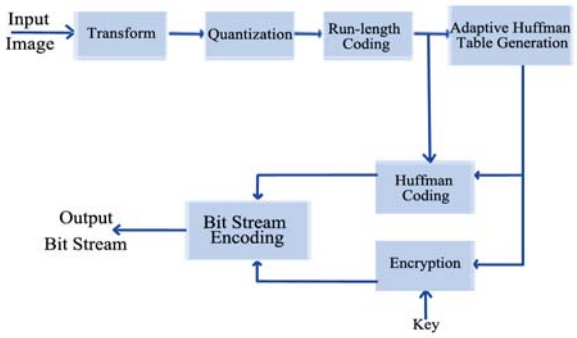


Figure 1. Conceptual Illustration of the Proposed Scheme

4 Experimental results

To evaluate the performance of our proposed scheme, preliminary experimental results as shown in Table 1 have been obtained using a variety of test images, including two natural images, one medical MRI image, and one biometric fingerprint image. These test images represent those found in different types of application demanding efficient and secure image compression. Our experiments were performed for the lossless image compression first, i.e., without the lossy quantization step. A standard public key encryption algorithm, RSA, was employed as the encryption method. Results were evaluated in terms of two efficiency metrics, namely, the Encryption Ratio (ER) and the Compression Ratio (CR). The Encryption Ratio is defined as the ratio of the number of bits being encrypted to the total number of bits in the output bit stream. A lower Encryption Ratio implies a lower computational cost in the encryption process, which is usually the bottleneck of the entire scheme. The Compression Ratio is the ratio of the number of bits in the output bit stream to the number of bits in the original image. According to Shannon's information theory, the compression ratio of a lossless compression scheme is bounded by the entropy of the image.

TABLE I. EFFICENCY RESULTS OF THE PROPOSED SCHEME

Test Image	Original Size	Overhead (bits)	ER	CR
Lena	512*512	11459	0.0152	1 : 2.747
Photo-grapher	256*256	10753	0.0628	1 : 2.882
Hand MRI	256*256	11079	0.0603	1 : 2.695
Finger-print	256*256	11779	0.0391	1 : 1.675

From Table 1, it can be seen that the Encryption Ratio of all test images are reasonably low. The overhead size lists the number of bits used to represent the optimal Huffman table. It should be noted that we did not perform any quantization for these lossless compression experiments. Therefore the size of the Huffman table is the largest possible with all entries of symbols. When applied to lossy compression, the number of entries in the table is expected to be much smaller, which will result in both lower Encryption Ratio and Compression Ratio. It should be also noted that with the increased image size, there is a slight increase in computational cost resulted from the generation of optimal Huffman table. However, the Encryption Ratio will actually decrease significantly, which implies greater savings in the computational cost resulted from the encryption process.

The compression ratios for different types of image reflect the redundancy inherent in different types of images. The fingerprint image has rich textures, thus higher entropy than the other types of images, resulting in a lower compression ratio. For lossless compression, it is apparent that our proposed scheme using optimized Huffman table for each image can achieve near-perfect compression ratio, with a slight overhead of the data for the Huffman table.

We have not performed any crypto analysis for our proposed scheme yet. Based upon our design analysis described in Section III, our proposed scheme, as a carefully designed selective encryption scheme, should achieve the same or nearly the same security level as that of the encryption algorithm employed.

5 Conclusion and future work

In this paper, we reviewed two different approaches of performing both image encryption and image compression

efficiently, namely, selective encryption and in-compression encryption using MHT, which were previously investigated for the emerging digital image applications requiring both data security and data efficiency. Based on our review and evaluation, we proposed a new approach that combines both approaches. As a selective encryption approach, our combined approach can take advantage of existing standard image compression schemes and data encryption algorithms Meanwhile, by using adaptive entropy codec which was optimized for each individual image, we expect the proposed scheme to achieve high compression efficiency and high security level at the same time.

Preliminary experimental results have partially confirmed our expectation using different types of test images with different size. We are in the process of running more experiments to demonstrate the efficiency and effectiveness of our proposed schemes when applying to different image compression schemes and/or employing different data encryption algorithms. In the future, we plan to extend our proposed scheme to lossy image compression scenario, and perform more in-depth security analysis.

6 References

[1] L. Tang, “Methods for encrypting and decrypting MPEG video data efficiently,” 4th ACM Int. Conf. on Multimedia, p. 219, Nov. 1996.

[2] C. Shi and B. Bhargava, “A fast MPEG video encryption algorithm,” in 6th ACM Int. conf. Multimedia, Sept. 1998.

[3] C. Shi, S. Wang, and B. Bhargava, “MPEG video encryption in real-time using secret key cryptography,” in Proc. PDPTA’99, 1999.

[4] H. Cheng and X. Li, “Partial encryption of compressed images and videos,” IEEE Trans. Signal Processing, vol. 48, p. 2439, Aug. 2000.

[5] Y. Sadourny and V. Conan, "A proposal for supporting selective encryption in JPSEC," IEEE Transactions on Consumer Electronics, vol.49, no.4, pp.846,849, Nov. 2003.

[6] J. Apostolopoulos, S. Wee, F. Dufaux, T. Ebrahimi, Q. Sun, and Z. Zhang, "The emerging JPEG-2000 security (JPSEC) standard," Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on , vol., no., pp.4 pp.,3885, 21-24 May 2006.

[7] C.-P. Wu and C.-C. J. Kuo, “Fast encryption methods for audiovisual data confidentiality,” SPIE Int. Symposium on Information Technologies 2000, vol. 4209, p. 284, Nov. 2000.

[8] C.-P. Wu and C.-C.J. Kuo, “Design of integrated multimedia compression and encryption systems”, IEEE Transactions on Multimedia Vol. 7, No. 5, pp. 828-839, 2005.

[9] D. Xie and C.-C.J. Kuo, "Enhanced multiple Huffman table (MHT) encryption scheme using key hopping," Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on , vol.5, no., pp.V-568,V-571 Vol.5, 23-26 May 2004.

[10] D. Xie, and C.-C.J. Kuo, “Multimedia Data Encryption via Random Rotation in Partitioned Bit Stream,” In Proceedings of IEEE International Symposium on Circuits and Systems, pp.568–571, May2004.

[11] S. A. El-said, K. F. A. Hussein, and M. M. Fouad, “Securing Multimedia Transmission Using Optimized Multiple Huffman Tables Technique”, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 1, March 2011.

[12] D. Gillman, M. Mohtashemi, and R. Rivest, “On breaking a huffman code,” IEEE Trans. Inform. Theory, vol. 42, no. 3, p. 972, May 1996.

Foreign Object Detection (FOD) using multi-class classifier with single camera vs. distance map with stereo configuration

Haoyuan Lin, Raj Aggarwal, and Arun Somani
Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa, USA

Abstract— *The reliable detection of foreign objects is a key requirement for the safe operation of autonomous machines. A foreign object is defined as an object that can be hurt by the machine or can damage the machine during its operation.*

In this paper we describe two fundamentally different approaches to detecting foreign objects. The first technique is to use video from a single camera and detect each object type individually using object shape information and integrate the results from each object type classifier. This also requires training the classifier for all possible shapes. The second technique is to use video from two cameras in a stereo configuration and utilize range and size information to detect foreign objects.

The test results for both of these techniques using real imagery are described in this paper. Both techniques perform satisfactorily. However, the techniques based on stereo imagery is computationally efficient and more robust.

Keywords: foreign object detection, cluster, vector boosting, stereo cameras, depth map

1. Introduction

Most object detection research focuses on how to design both accurate and fast detector for one particular class of objects [1], [2]. In order to detect multiple classes of objects, multiple detectors are deployed simultaneously, each detecting the respective objects for a given input. This is not efficient. On the other hand, many applications do not care about the categories of the objects.

Detecting objects but not necessarily classifying them may be a requirement for many safety applications. The classification step can be considered as redundant for the system. In other applications, our first interest, the most useful information is whether any of the objects is present or not. If it does, the location and size may be of secondary interest.

We consider objects of interest as foreign objects which can be classified in one ensemble positive class. The merit of this method is by considering the similarity among all foreign objects, a group of similar features can be used and shared in ensemble positive class. Foreign objects can be of any arbitrary shape with certain volume in a specified range. Based on the experiments, we observe that the deformation of objects may degrade the performance of such detectors. In our work, we consider using two cameras to generate depth map and then select blocks in the depth map whose depth is below the threshold. Size discrimination is applied to obtain blocks with certain area. The merit of this method is to consider these foreign objects as a block with certain area and not to consider which categories they belong to.

2. Background

Object detection is a fundamental problem in computer vision which has seen a great progress in both performance and detection speed in the last few years [3]. For example, pedestrian protection system [4](implemented in the advanced driver assistant system) and intelligent robot [5] that can follow the specified target are two good examples of such applications.

To achieve the goal of detecting foreign objects, an idea is to use several categories of objects such as pedestrian, vehicle or animal to cover the most common foreign objects. The foreign objects detection task is converted to a multi-classification task. Haar-like feature [6] considers the pixel intensity difference between adjacent rectangular regions at a specific location as a vector. Haar-like feature has been proved to be efficient in human face detection problem. However, Haar-like feature cannot perform well under the condition of clustered background.

Histogram of oriented gradients (HOG) [1] is using edge orientation histogram to describe the shape of the objects. Algorithm using HOG feature was developed [1] to boost the performance for object detection since HOG feature can tolerate the mis-position or deformation of samples.

Support vector machine (SVM) [7] is using optimization method to find a separating hyper-plane to solve the pattern recognition problems. Boosting algorithm [8] is another powerful learning algorithm which composes several weak classifiers into a strong classifier. Compared to SVM , boosting algorithm chooses a small amount of representative features to construct the detector rather than all features as is the case for SVM.

Real AdaBoost [9] also has been proposed to handle multi-class multi-label problems. Vector boosting [10] is proposed to extend the output of the Real AdaBoost from scalar to vector which allows one sample to be assigned into several classes.

3. Our Approach

For object detection with multiple categories of objects, we use a divide-and-conquer strategy. Our detector is based on a tree-structure which can be easily adopted for a coarse-to-fine strategy to handle multiple category objects. Each node uses a cluster algorithm to split the data into several clusters before training the classifier. Since different categories of objects can also be similar in certain parts, several different categories of objects can be handled at one node in the tree. Each node in the tree is a strong classifier that puts emphasis on the diversities among these clusters. The overview of our method is shown in Fig. 1.

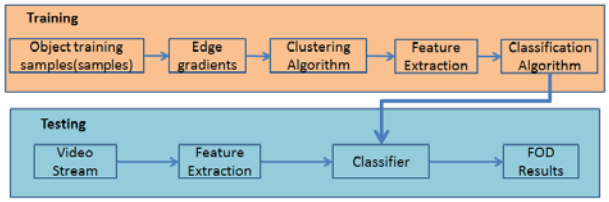


Fig.1: The overview of our single camera multi-class classifier method

3.1 Clustering algorithm

In this paper, we use vector-space model to represent clustering algorithm which is an iterative process. The goal of clustering algorithm is to maximize the internal similarity within each single cluster and minimize the external similarity between any two clusters. Under this criterion, we consider not only the single cluster, but also the relationship between any two clusters and the whole sample space.

Assume there are k clusters. We denote i_{th} cluster as $C_i \quad (i = 1, 2, ...k)$. Let n_i be the number of training samples in C_i. Each training sample x_i is a matrix of pixel values. The feature of each training sample x_i is considered as one vector as $\varphi(x_i)$. The function of $\varphi()$ is to map from a set of pixel values to a vector of feature values. $\varphi(x_i)$ is computed by applying feature extraction algorithm to each training sample. An example is shown in Fig. 2.

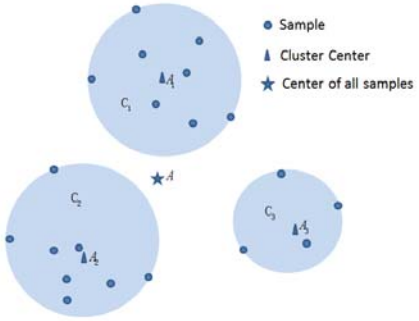


Fig.2: Schematic diagram depicting multiple clusters and their centers

The similarity between two samples x and y using Euclidean distance is defined as

$$\| \varphi(x) - \varphi(y) \|^2 \tag{1}$$

Based on this definition, the similarity of the C_i cluster s_{1i} is defined as the average sum of relative distances among all sample points in C_i to the center of C_i which is denoted as A_i.

$$s_{1i} = \sum_{x \in C_i} \| \varphi(x) - \varphi(A_i) \|^2 \tag{2}$$

The similarity between any two clusters s_{2i} is defined as the sum of relative distances from the center of each cluster A_i to the center of all clusters which is denoted as A. Since the size of each cluster is not the same, the contribution of each cluster is weighted by its size as follows.

$$s_{2i} = n_i \| \varphi(A_i) - \varphi(A) \|^2 \tag{3}$$

We use both internal evaluation and external evaluation to evaluate the quality of clustering, the score s is defined by

$$s = \frac{s_1}{s_2} = \frac{\sum_{i=1}^{k} s_{1i}}{\sum_{j=1}^{k} s_{2j}} \tag{4}$$

where s_1 denotes as internal evaluation, and s_2 denotes as external evaluation. The goal of clustering algorithm is to maximize the score s.

Not all features extracted from samples are suitable for clustering. The classification power of the feature can be measured by Z value which is defined as follows.

$$Z = 2 \sum_j \sqrt{W_+^j W_-^j} \tag{5}$$

We define the foreign objects as positive samples and image patches excluding foreign objects as negative samples. W_+ is the sum of weights for positive samples and W_- is the sum of weights for negative samples. The weight for all samples is normalized. The feature is more discriminating if the difference between W_+ and W_- is bigger. Therefore, the smaller value of Z implies that the classification power of the feature is more discriminating. With the help of Z value, a threshold is set to determine if the features selected by boosting training is close to saturation and no more features could be selected. Then boosting algorithm should stop. If the selected features are not close to saturation, clustering algorithm should be called to split the training samples into more clusters. The results of the boosting algorithm are a set of features which are selected to construct a strong classifier in the node. Meanwhile, these features are also used by clustering algorithm to prepare the samples for the next level node.

Since the background of positive training samples can be regarded as noise, some samples may not be assigned to a representative cluster due to the noise. This will produce clustering bias in the result of the clustering algorithm and will result in worse effects in the subsequent training stages. In the next step, the boosting algorithm will make biased decision based on the worse effects which are then used to select the most representative feature to train a classifier. To solve this problem, we set a safe barrier that can tolerate the bias effect from the cluster. In a addition, some examples could be assigned to multiple clusters when they are on the cluster edge or close to other clusters. This is a modification of the original clustering algorithm. The detail of our clustering algorithm is shown as follows.

Algorithm 1 Clustering Algorithm

Input:

The sample set S=$\{\vec{x}_1, ..., \vec{x}_N\}$(set of samples to be clustered)

The maximum cluster number K

The overlap parameter b

Output:

The cluster set W=$\{\vec{w}_1, ..., \vec{w}_k\}$(samples in each cluster)

The cluster centroids set C=$\{\vec{c}_1, ..., \vec{c}_k\}$ (set of cluster centroids)

for $k \leftarrow$2 to K **do**
  $\{\vec{\mu}_1, ..., \vec{\mu}_k\} \leftarrow RandomSeeds(\{\vec{x}_1, ..., \vec{x}_N\}, k)$
  for $i \leftarrow 1$ to N **do**
    $j \leftarrow \text{argmin} \|\vec{x}_i - \vec{\mu}_{jj}\|$
    $\vec{w}_j \leftarrow \vec{w}_j \cup \vec{x}_i^{jj}$
  end for
  $\{\vec{w}_1, ..., \vec{w}_k\} \leftarrow Reassignment(\{\vec{w}_1, ..., \vec{w}_k\})$
  for $i \leftarrow 1$ to k **do**
    $d_i \leftarrow b$ * $CalculateClusterDiameter\ (\vec{w}_i)$
    $\vec{c}_i \leftarrow CalculateClusterCenter\ (\vec{w}_i)$
  end for
  for $i \leftarrow$1 to k **do**
    $\vec{w}_i \leftarrow \{\}$
  end for
  for $i \leftarrow 1$ to N **do**
    for $j \leftarrow 1$ to k **do**
      if $\|\vec{x}_i - \vec{c}_j\| < d_j$ **then**
        $\vec{w}_j \leftarrow \vec{w}_j \cup \vec{x}_i$
      end if
    end for
  end for
end for

3.2 Classification algorithm

In this section, we compare Real AdaBoost to vector boosting with an example, and then describe the detail of our extend vector boosting algorithm.

Algorithm 2 Reassignment

Input:

The sample set S=$\{\vec{x}_1, ..., \vec{x}_N\}$

The initial cluster set W_1=$\{\vec{w}_1, ..., \vec{w}_k\}$

Output:

The final cluster set W_2=$\{\vec{w}_1, ..., \vec{w}_k\}$

repeat
  $flag \leftarrow false$
  $score \leftarrow CalculateScore\ (\{\vec{w}_1, ..., \vec{w}_k\})$
  for i$\leftarrow$1 to N **do**
    $p \leftarrow GetClusterId(\vec{x}_i)$
    $\vec{w}_p \leftarrow \vec{w}_p - \vec{x}_i$
    $[s, q] \leftarrow FindMaxScore\,(\{\vec{w}_1, ..., \vec{w}_k\}, \vec{x}_i)$
    if $s > score$ **then**
      $flag \leftarrow true$
      $\vec{w}_q \leftarrow \vec{w}_q \cup \vec{x}_i$
    else
      $\vec{w}_p \leftarrow \vec{w}_p \cup \vec{x}_i$
    end if
  end for
until $flag$

Real AdaBoost [9] has been proposed to handle multi-class multi-label problems. Multi-class problem is to classify instance into more than two classes instead of only positive class and negative class. Multi-label problem is that each instance could be assigned to more than one class. The idea of Real AdaBoost method is to make each class orthogonal and consider each label independently so that this problem can be converted into original binary classification problem. One prediction is considered as correct if and only if all class labels of the instance are predicted correctly. However, this condition is not tenable since some attributes are dependent.

Vector boosting [10] is proposed to extend the output of the Real AdaBoost from scalar to vector. The difference is that Real AdaBoost converts multi-label problem to several binary classification problem (the attributes are required to be independent) while vector boosting algorithm consider all labels for one instance together (the attributes are not required to be independent). Vector boosting algorithm modifies the sample weight redistribution formula using vector dot production which can avoid some independent status. For example, there are three classes horse, deer, human. The ground-true label of horse is $\{1, 0, 0\}$, the ground-true label of deer is $\{0, 1, 0\}$, and the ground-true label of human is $\{0, 0, 1\}$. One sample is clustered by the clustering algorithm as $\{1, 1, 0\}$ which means it is like horse and deer and may be any of them (the deer and horse are four-leg animals that look like similar from a certain distance). In Real AdaBoost, classifier output of $\{1, 1, 0\}$ is considered as the only right estimation. However, in vector boosting, besides $\{1, 1, 0\}$, classifier outputs $\{1, 1, 1\}$ is also considered as right estimation since the third attribute is independent of

the other two. Real AdaBoost is suitable for problems in which the label of sample is absolutely with no ambiguity while vector boosting algorithm allows ambiguity to be handled. Evaluating all attributes together is the merit of vector boosting algorithm.

In our problem, we use a label denoted as either 1 or 0 to indicate if the sample belongs to the cluster or not. A label of value 0 means the sample does not belong to the cluster. A label of value 1 means the sample may belong to the cluster. Label 0 means the instance is not in this class which shares the same concept in Real AdaBoost. When the number of label 1 in one instance equals to one, the meaning is the same as in Real AdaBoost. When the number of label 1 in one instance is more than one, different from Real AdaBoost, the instance could be in any class of these labels but cannot be in those classes where the label is 0. The label meaning is determined by our clustering algorithm. So Real AdaBoost and vector boosting algorithm cannot work for our problem.

The following example illustrates the meaning of the label in our problem. One sample is in the area safe barrier between class 1 and class 2. The sample is clustered as class 1 and class 2. The real label should be $\{1,1,0\}$. The correct prediction of the Real Adaboost is $\{1,1,0\}$ since all class label of the instance should predicted correctly. The correct predictions of the vector boosting are $\{1,1,0\}$, $\{1,1,1\}$ since the labels of the first two classes are correct and the label of last class is independent of the other two. The correct predictions of the extended vector boosting are $\{1,1,0\}$, $\{1,0,0\}$, $\{0,1,0\}$ since any label of the first two classes is correct and the label of the last class is 0. The difference is shown is Table 1. The algorithm pseudo code is shown as follows.

Table 1: Correct predictions with different boosting algorithm.

Real label	$\{1,1,0\}$
Real Adaboost	$\{1,1,0\}$
Vector Boosting	$\{1,1,0\}\{1,1,1\}$
Extended Vector Boosting	$\{1,1,0\}\{1,0,0\}\{0,1,0\}$

3.3 FOD system framework

For foreign object detection with any shape, it is necessary to find a certain stable pattern for the foreign object. The basic idea behind our method is that the distance between cameras and any part of one object are much more similar than other objects in the scene. Disparity map [11], [12] is widely used in the computer vision area to recover the 3D structure of a scene using two or more images of the 3D scene, each acquired from a different viewpoint in space. With a set of well configured cameras, disparity, which means the distance between the two corresponding points, can be calculated. The disparity is calculated by finding the corresponding points in the two frames which have a similar

Algorithm 3 Classification algorithm

Input:

The sample set S=$\{(\vec{x}_1,\vec{v}_1),...,(\vec{x}_N,\vec{v}_N)\}$ where $\vec{v}_i$ is got by our proposed cluster algorithm

1: Initialize the sample weight as $D_1(i)=1/N$
2: **for** $t=1,...,T$ **do**
3: Under the updated sample weight, train a weak classifier $h_t(x)$
4: Update the sample weight $D_{t+1}(i)=\frac{D_t(i)exp[v_i\otimes h_t(x_i)]}{Z_t}$ where Z_t is a normalization factor to make $D_{t+1}(i)$ a distribution
5: The final strong classifier $H(x)=\sum_{t=1}^{T}h_t(x)$
6: **end for**

feature. One method to calculate the disparity is using the feature matching [13].

We use stereo vision to address the problem in hand. The two camera model is shown in Fig. 3. The parameters used in this paper are shown in Table 2.

In Fig. 3, based on the triangulation principle, we have $\frac{x_l}{X_l}=\frac{f}{Z}$ and $\frac{x_r}{X_r}=\frac{f}{Z}$. Manipulation of these equations give us $X_l=\frac{Z}{f}x_1$ and $X_r=\frac{Z}{f}x_r$. Also $X_l+X_r=T$. Therefore, $\frac{Z}{f}(x_l+x_r)=T$ where x_l+x_r is the disparity d. Then the distance between target and camera is given by $Z=\frac{Tf}{d}$.

Table 2: Parameters of the camera model.

Distance between the two cameras	T
Focus length of the cameras	f
Distance between target and camera	Z
Position of middle point of camera film	c
Position of object in camera file	x
Disparity of the target	d
Displacement between target and camera	X

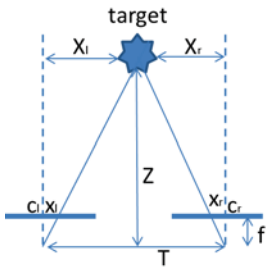


Fig.3: The structure of FOD

The depth map is an important resource in the algorithm. In order to make the distance calculation robust, a block of pixels are grouped together. The distance calculation is based on the blocks instead of pixel that can eliminate single error. The calculation of the disparity also uses the feature matching algorithm [13]. Edge feature within a block in

the left frame can be used as the pattern to search in the right frame. For the purpose of the FOD task instead of 3D construction, the depth calculation is calculated coarsely. If the area of the block is too small, holes will be seen in the depth map. If the area of the block is too big, the object may not be detected especially when the object is far away from the camera.

In the Fig. 3, the number denotes the distance between background and camera in meters. Once the depth map for the initial scene is generated, the system is ready to work. The distance of the new frame is calculated and compared with the initial depth map. If the value of depth is smaller than the initial value for the block, foreign object may exist in this block and the position of the block is memorized. The distance filter can be used to filter the background object. Any feature block may be ignored beyond the range of interest. Only feature block within the range is marked. After the processing of the filter, blood fill algorithm is used to connect the neighbor block into one integrated block. Each integrated block is represented as one object. The distance is also related to the integrated block. In our experimental setup, we use 8-neighbor rule to recognize the neighbor candidate around one block. We consider that the foreign object can be any kind of shape, therefore, we believe that all eight directions should be considered as the extension of the foreign objects.

Size discrimination is implemented using a make-up table with the distance information and object size information. When a small object is too close to the camera, the size of the pixel block may appear as a big block. By using the make-up table we store the lower-bound level of the size which has high confidence. Fig. 4 shows the result obtained from the filter using the depth map when a foreign object (a vehicle) stopped in front of the camera. The number is depth of the block in meter. The rectangle denotes the position of the foreign object.

Fig.4: The result after the blood fill algorithm when one foreign object (vehicle) is in front of the camera

3.4 Adaptive depth calculation algorithm

The value of depth may not be a constant for one particular block if the camera is moving on the rugged ground. The initial depth map generated for the initial scene cannot be the reference for the following frames. For example, the vehicle is stand-by on an even ground. The depth map is generated based on the initial scene at this time. Once the vehicle is running on the down ramp, the distance between the camera and ground will be smaller than the initial scene since the camera is pointing down than the initial state. In this case, there will be false alarms even if there is no foreign object in the scene. The reason is that there will be areas of quite small value of the depth when the camera is pointing down to the ground. The system with high rate of false alarms is not useful at all.

An easy way is to solve this problem is to generate a depth map with the minimum depth value for each block. Only when the value of the depth is smaller than the minimum value on the depth map, the foreign object can be determined with a higher level of confidence. In order to get enough data to train the initial depth map, the vehicle with mounted cameras is set to run on the experiment site to collect the data. Then the depth map for each frame is calculated. The minimum value of depth for each block will be the value in the depth map. The advantage of this method is to reduce the false alarm sharply. However, this method will lose a certain number of blocks which the foreign objects are located in. The reason is that the comparison is based on the minimum depth map which is a very adverse situation.

A better solution is to use information on the current frame appropriately. The method will not require the initial scene to generate the initial depth map. The area of the object shows in the frame can be part of scene if the object is in a certain distance from the vehicle or full of scene if the object is very close to the vehicle. Both cases can be handled using only the current frame without any help of the initial frame. If the area of the object takes the part of scene, the rest of the scene could be the clue to get the distance from the camera to the ground. If the area of the object takes full of scene, it is easy to handle using the depth map with the minimum depth as described above. The detail of the method is described as follows.

In order to make the distance calculation robust, the area of each frame is divided into sets of blocks instead of each pixel. The dimension of the block map is m*n. The denotations used in this algorithm are shown in Table 3.

For each block b_{ij}, the distance d_{ij} will be calculated using the disparity. For each row, the maximum depth value is calculated. In order to eliminate the outlier value, the median value of the first three big numbers is picked as the maximum depth value m_i. If the depth value of any block in row i is smaller than $m_i - t_i$, the block will be marked as a dangerous points. Like the last section, a breadth first search algorithm will be used to group. The optimization objective

Table 3: Correct Predictions with different boosting algorithm.

The number of blocks in a row	m
The number of blocks in a column	n
The denotation for each block	b_{ij}
The depth value of block b_{ij}	d_{ij}
The maximum depth value in row i	m_i
Threshold of depth different value in row i	t_i

is to find such a set of threshold t_i for each row i such that the error rate of detecting is high and the false alarm is low.

4. Experiment and discussion

We apply our method to the problem of FOD in a real scenario. To demonstrate the feasibility, we trained multiple categories of object detector. In this experiment, we consider objects that often appear on the road that can cause a traffic accident. We apply our algorithm on four big object categories (pedestrian, bicycle, vehicle, and four-legged animal). Reference detector is also trained for each category to compare the performance with multiple categories object detector. A big variance of shape that cannot be handled by one classifier could be seen in our setup. Our goal is to make all objects in these categories distinct without knowing the sub-category in advanced.

4.1 Clustering performance

At the first stage, all positive samples are fed into the general Real boosting algorithm to select the feature that can be used to distinguish the positive samples against the negative samples. The value of the threshold Z is obtained by cross-validation to make sure that the training is not overfitting (Z value is too big) or underfitting (Z value is too small). When the Z value achieves the threshold, this set of features is used for clustering algorithm. A small portion of features may come from the background when Z value is close to threshold, therefore we set a tolerant ratio. This tolerant ratio is defined as the executing diameter divided by diameter of each cluster. In this experiment, the ratio is set to 1.2. Any instance locating in the overlapping area may belong to multiple clusters. After the clustering algorithm is computed, each cluster is trained by using modified vector boosting algorithm instead of general boosting algorithm to select discriminative features to construct strong classifier. These steps are performed iteratively until the training error is under the predefined level. At the end of training, we get 58 leaf nodes in the FOD tree. In Fig. 5 we show parts of results in the node which contains the samples in each cluster.

Since the width-to-height ratio of objects is different, we propose a simple method to shrink the area that can pass all stages of the FOD classifier. The idea is based on that the discriminating features are mostly located around object outline and seldom located in background area. The operation is to squeeze the area from four directions in three steps. First, get four strips from boundary of the output box and get the number of discriminating features in the area. Second, squeeze the output box in the direction which has the least number of discriminating features. Third, calculate the total number of features that has been removed. If it is more than 10% of the total features, then stop. If not, go to first step.

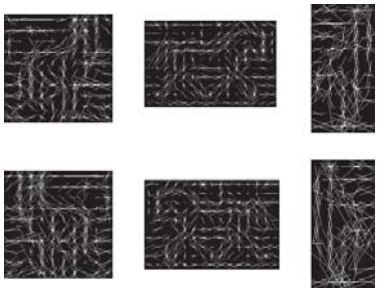


Fig. 5: The clustered horse category samples in the node of the foreign detection tree

4.2 Classification algorithm comparison

In order to compare the performance of proposed method to the method composed of multiple classifiers, we build four classifiers for four objects, (i.e. pedestrian, bicycle, vehicle and four-leg animal) for each categories. We made each of the multiple classifiers learn using the samples coming from the single category belonging to the same training data set and evaluate on the same test data set. For each classifier, the training error is set to 5*10e-4 and training detection rate is set to 0.95 which is the same as the FOD. Next, we evaluate the performance of 4 separated classifiers. We calculate the number of missing and false positive is from four categories. Fig. 6 plots the ROC curves.

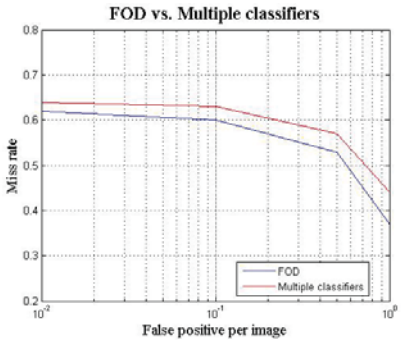


Fig. 6: Quantitative result of FOD vs. multiple classifiers

From the result, we observe that our FOD method outperforms the multiple classifier method. One main reason is that we use soft clustering algorithm which is based on the discriminating features while multiple classifier method use pre-defined category information which is hard clustering algorithm based on the domain knowledge. In this experiment, we notice that the merit of soft clustering algorithm,

which is allowed to be tolerant between similar categories. Another reason is that the extended vector boosting method makes use of the relation information of each feature among multiple classes which is totally ignored by the separated classifier method.

4.3 Multi-class classifier with single camera vs. distance map with two cameras

In this experiment, we use four data sets to test the correctness of the algorithm shown in Table 3. The data sets include two kinds of objects, i.e. pedestrian and vehicle. For each data set, the # frames means the number of frames in the video clips and the # object means the times the object shows up in the video clips. D means detected, M means missed and FA means false alarm. We make the experimental cases changeable and control the moving direction of the object for the purpose of all corner cases. The foreign object is designed to move on the designated path. The object could move from the far site to the near site. The object could move from the near site to the far site and move from left to right with the same distance. The pose of the objects can be arbitrary shape.

We compare our stereo-based algorithm based on the depth information to the previous work using the multi-classifier method based on the shape information. From the results, we observe that the detecting rate increased more than shape-based method. This can be contributed to the reason that the distance can be tolerated when the object is with shape that is not included in the model of the shape-based method. Besides that, the false alarm is also decreased than the shape-based method. The background may be clustered in most case, so the foreground and background may mix together to make the frame area much more like a target object. However, the distance is always not in the range. The area can be eliminated by the depth information.

Table 4: Experiment result comparison between stereo-based and shape-based method.

Datasets	Stereo-based			Shape-based		
	D	M	FA	D	M	FA
Dataset1 (42 objects/116 frames)	40	2	0	39	3	2
Dataset2 (15 objects/34 frames)	14	1	0	13	2	0
Dataset3 (12 objects/56 frames)	11	1	0	10	2	0
Dataset4 (22 objects/60 frames)	21	1	1	21	1	2

5. Conclusion

We describe a method to learn a foreign object detector which can detect many categories objects simultaneously, without knowing the label information in each instance. We divide the positive sample space by unsupervised clustering algorithm with tolerance ratio with the features learned by our proposed extended vector boosting algorithm. The extended vector boosting algorithm considers the relation of multiple classes instead of individual classes. In the research on detection of any foreign objects with no pre-set shape, we use stereo cameras configuration to generate depth map. The goal of our research is to design a FOD framework using a depth map to find block with certain area. Our fundamental approach is to use stereo matching algorithm to get the disparity information based on intensity images from stereo cameras and using the camera model to retrieve the distance information. From the result of our experiments, the proposed framework has a better performance with higher detection rate with lower false alarm. The processing speed is boosted significantly.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[2] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2895–2902.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 743–761, 2012.

[4] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, 2014, pp. 115–120.

[5] I. T. Ćirić, Ž. M. Ćojbašić, V. D. Nikolić, T. S. Igić, and B. A. Turšnek, "Intelligent optimal control of thermal vision-based person-following robot platform," *Thermal Science*, vol. 18, no. 3, pp. 957–966, 2014.

[6] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[7] T. Joachims, *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[8] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[9] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, vol. 37, no. 3, pp. 297–336, 1999.

[10] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 446–453.

[11] T. D. Sanger, "Stereo disparity computation using gabor filters," *Biological cybernetics*, vol. 59, no. 6, pp. 405–418, 1988.

[12] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3dtv services providing interoperability and scalability," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 217–234, 2007.

[13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 559–568.

PiQ: Perceptual Image Quantizer for Wavelet Image Coders

Jaime Moreno†‡, Oswaldo Morales‡, and Ricardo Tejeida†
†Superior School of Mechanical and Electrical Engineers, National Polytechnic Institute of Mexico, IPN Avenue, Lindavista, Mexico City, 07738, Mexico.
‡Signal, Image and Communications Department, University of Poitiers, Poitiers, 30179, France.
e-mail:jmorenoe@ipn.mx

Abstract— *The aim of this work is to explain how to apply perceptual criteria in order to define a perceptual forward and inverse quantizer. We present its application to the* Hi-SET *coder. Our approach consists in quantizing wavelet transform coefficients using some of the human visual system behavior properties. Taking in to account that noise is fatal to image compression performance, because it can be both annoying for the observer and consumes excessive bandwidth when the imagery is transmitted. Perceptual quantization reduces unperceivable details and thus improve both visual impression and transmission properties. The comparison between JPEG2000 coder and the combination of* Hi-SET *with the proposed perceptual quantizer (*χSET*) shows that the latter is not favorable in* PSNR *than the former, but the recovered image is more compressed (less bit-rate) at the same or even better visual quality measured with well-know image quality metrics, such as MSSIM, UQI or VIF, for instance.*

Keywords: Human Visual System, Contrast Sensitivity Function, Perceived Images, Wavelet Transform, Peak Signal-to-Noise Ratio, No-Reference Image Quality Assessment, JPEG2000.

1. Introduction

Digital image compression has been a research topic for many years and a number of image compression standards has been created for different applications. The JPEG2000 is intended to provide rate-distortion and subjective image quality performance superior to existing standards, as well as to supply functionality [1]. However, JPEG2000 does not provide the most relevant characteristics of the human visual system, since for removing information in order to compress the image mainly information theory criteria are applied. This information removal introduces artifacts to the image that are visible at high compression rates, because of many pixels with high perceptual significance have been discarded.

Hence, it is necessary an advanced model that removes information according to perceptual criteria, preserving the pixels with high perceptual relevance regardless of the numerical information. The Chromatic Induction Wavelet Model presents some perceptual concepts that can be suitable for it. Both CBPF and JPEG2000 use wavelet transform. CBPF uses it in order to generate an approximation to how every pixel is perceived from a certain distance taking into account the value of its neighboring pixels. By contrast, JPEG2000 applies a perceptual criteria for all coefficients in a certain spatial frequency independently of the values of its surrounding ones. In other words, JPEG2000 performs a global transformation of wavelet coefficients, while CBPF performs a local one.

CBPF attenuates the details that the human visual system is not able to perceive, enhances those that are perceptually relevant and produces an approximation of the image that the brain visual cortex perceives. At long distances the lack of information does not produce the well-known compression artifacts, rather it is presented as a softened version, where the details with high perceptual value remain (for example, some edges).

2. JPEG2000 Global Visual Frequency Weighting

In JPEG2000, only one set of weights is chosen and applied to wavelet coefficients according to a particular viewing condition (100, 200 or 400 dpi's) with fixed visual weighting[1, Annex J.8]. This viewing condition may be truncated depending on the stages of embedding, in other words at low bit rates, the quality of the compressed image is poor and the detailed features of the image are not available since at a relatively large distance the low frequencies are perceptually more important. The table 1 specifies a set of weights which was designed for the luminance component based on the CSF value at the mid-frequency of each spatial frequency. The viewing distance is supposed to be 4000 pixels, corresponding to 10 inches for 400 dpi print or display. The weight for *LL* is not included in the table, because it is always 1. Levels 1, 2, . . . , 5 denote the spatial frequency levels in low to high frequency order with three spatial orientations, *horizontal, vertical* and *diagonal*.

3. Perceptual Forward Quantization

3.1 Methodology

Quantization is the only cause that introduces distortion into a compression process. Since each transform sample

Table 1
RECOMMENDED JPEG2000 FREQUENCY (s) WEIGHTING FOR 400 DPI'S ($s = 1$ IS THE LOWEST FREQUENCY WAVELET PLANE).

s	*horizontal*	*vertical*	*diagonal*
1	1	1	1
2	1	1	0.731 668
3	0.564 344	0.564 344	0.285 968
4	0.179 609	0.179 609	0.043 903
5	0.014 774	0.014 774	0.000 573

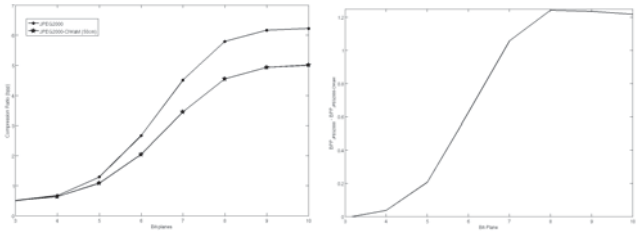
at the perceptual image $\mathcal{I}_\rho$ is mapped independently to a corresponding step size either Δ_s or Δ_n, thus $\mathcal{I}_\rho$ is associated with a specific interval on the real line. Then, the perceptually quantized coefficients $\mathcal{Q}$, from a known viewing distance d, are calculated as follows:

$$\mathcal{Q} = \sum_{s=1}^{n} \sum_{o=v,h,d} sign(\omega_{s,o}) \left\lfloor \frac{|\alpha(\nu, r) \cdot \omega_{s,o}|}{\Delta_s} \right\rfloor + \left\lfloor \frac{c_n}{\Delta_n} \right\rfloor \quad (1)$$

Unlike the classical techniques of Visual Frequency Weighting (VFW) on JPEG2000, which apply one CSF weight per sub-band [1, Annex J.8], Perceptual Quantization using CBPF (PiQ) applies one CSF weight per coefficient over all wavelet planes $\omega_{s,o}$. In this section we only explain Forward Perceptual Quantization using CBPF (F-PiQ). Thus, Equation 1 introduces the perceptual criteria of Perceptual Images to each quantized coefficient of Equation of Dead-zone Scalar Quantizer. A normalized quantization step size $\Delta = 1/128$ is used, namely the range between the minimal and maximal values at $\mathcal{I}_\rho$ is divided into 128 intervals. Finally, the perceptually quantized coefficients are entropy coded, before forming the output code stream or bitstream.

3.2 Experimental Results applied to JPEG2000

The Perceptual quantizer F-PiQ in JPEG2000 is tested on all the color images of the *Miscellaneous volume* of the University of Southern California Image Data Base[2]. The data sets are eight 256×256 pixel images and eight 512×512 pixel images, but only visual results of the well-known images *Lena*, *F-16* and *Baboon* are depicted, which are 24-bit color images and 512×512 of resolution. The CBPF model is performed for a 19 inch monitor with 1280 pixels of horizontal resolution at 50 centimeters of viewing distance. The software used to obtain a JPEG2000 compression for the experiment is *JJ2000*[3]. Figure 1(a) shows the assessment results of the average performance of color image compression for each bit-plane using a Dead-zone Uniform Scalar Quantizer (SQ, function with heavy dots), and it also depicts the results obtained when applying F-PiQ(function with heavy stars). Using CBPF as a method of forward quantization, achieves better compression ratios than SQ with the same threshold, obtaining better results at the highest bit-planes, since CBPF reduces unperceivable features. Figure 1(b) shows the contribution of F-PiQ in the JPEG2000 compression ratio, for example, at the eighth bit-plane, CBPF reduces 1.2423 bits per pixel than the bit rate obtained by SQ, namely in a 512×512 pixel color image, CBPF estimates that 39.75KB of information is perceptually irrelevant at 50 centimeters.



(a) (b)

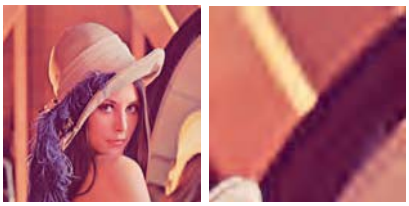
Fig. 1
(A)JPEG2000 COMPRESSION RATIO (BPP) AS A FUNCTION OF BIT-PLANE. FUNCTION WITH HEAVY DOTS SHOWS JPEG2000 ONLY QUANTIZED BY THE DEAD-ZONE UNIFORM SCALAR MANNER. WHILE FUNCTION WITH HEAVY STARS SHOWS JPEG2000 PERCEPTUALLY PRE-QUANTIZED BY F-PiQ. (B)THE BIT-RATE DECREASE BY EACH BIT-PLANE AFTER APPLYING F-PiQ ON THE JPEG2000 COMPRESSION.

Figure 2 depicts examples of recovered images compressed at 0.9 bits per pixel by means of JPEG2000 (a) without and (b) with F-PiQ. Also these figures show that the perceptual quality of images forward quantized by PiQ is better than the objective one. Also, figure 3 shows examples of recovered images of *Baboon* compressed at 0.59, and 0.45 bits per pixel by means of JPEG2000 (a) without and (b) with F-PiQ. In Fig. 3(a) PSNR=26.18 dB and in Fig. 3(b) PSNR=26.15 dB but a perceptual metrics like WSNR [4], for example, assesses that it is equal to 34.08 dB. Therefore, the recovered image Forward quantized by PiQ is perceptually better than the one only quantized by a SQ. Since the latter produces more compression artifacts, the PiQ result at 0.45 bpp (Fig. 3(b)) contains less artifacts than SQ at 0.59 bpp. For example the *Baboon*'s eye is softer and better defined using F-PiQ and it additionally saves 4.48 KB of information.

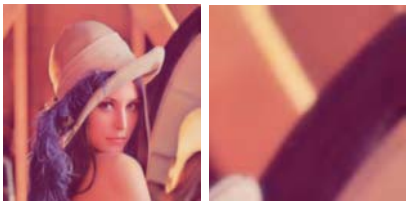
4. Perceptual Inverse Quantization

The proposed Perceptual Quantization is a generalized method, which can be applied to wavelet-transform-based image compression algorithms such as EZW, SPIHT, SPECK or JPEG2000. In this work, we introduce both forward (F-PiQ) and inverse perceptual quantization (I-PiQ) into the H*i*-SET coder. This process is shown in the green blocks of Fig. 4. An advantage of introducing PiQ is to maintain the embedded features not only of H*i*-SET algorithm but also of any wavelet-based image coder. Thus, we call CBPF Perceptual Quantization + H*i*-SET = cH*i*-SET or χSET.

Both JPEG2000 and χSET choose their VFWs according to a final viewing condition. When JPEG2000 modifies the



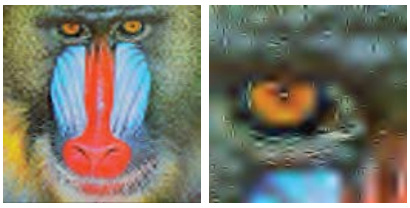
(a) JPEG2000 PSNR=31.19 dB.



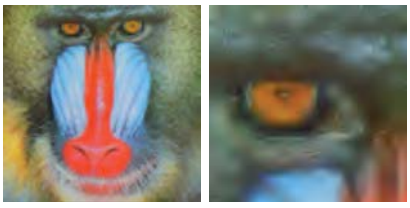
(b) JPEG2000-F-PiQ PSNR=27.57 dB.

Fig. 2

EXAMPLES OF RECOVERED IMAGES OF LENNA COMPRESSED AT 0.9 BPP.



(a) JPEG2000 compressed at 0.59 bpp.



(b) JPEG2000-F-PiQ compressed at 0.45 bpp.

Fig. 3

EXAMPLES OF RECOVERED IMAGES OF BABOON.

Fig. 4

THE χSET IMAGE COMPRESSION ALGORITHM. GREEN BLOCKS ARE THE F-PiQ AND I-PiQ PROCEDURES.

quantization step size with a certain visual weight, it needs to explicitly specify the quantizer, which is not very suitable for embedded coding. While χSET neither needs to store the visual weights nor to necessarily specify a quantizer in order to keep its embedded coding properties.

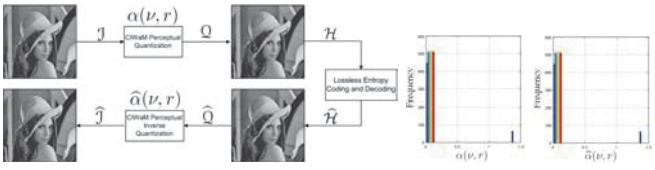
The main challenge underlies in to recover not only a good approximation of coefficients $\mathcal{Q}$ but also the visual weight $\alpha(\nu, r)$(Eq. 1) that weighted them. A recovered approximation $\widehat{\mathcal{Q}}$ with a certain distortion Λ is decoded from the bitstream by the entropy decoding process. The VFWs were not encoded during the entropy encoding process, since it would increase the amount of stored data. A possible solution is to embed these weights $\alpha(\nu, r)$ into $\widehat{\mathcal{Q}}$. Thus, our goal is to recover the $\alpha(\nu, r)$ weights only using the information from the bitstream, namely from the Forward quantized coefficients $\widehat{\mathcal{Q}}$.

Therefore, our hypothesis is that an approximation $\widehat{\alpha}(\nu, r)$ of $\alpha(\nu, r)$ can be recovered applying CBPF to $\widehat{\mathcal{Q}}$, with the same viewing conditions used in $\mathcal{I}$. That is, $\widehat{\alpha}(\nu, r)$ is the recovered e-CSF. Thus, the perceptual inverse quantizer or the recovered $\widehat{\alpha}(\nu, r)$ introduces perceptual criteria to Inverse Scalar Quantizer and is given by:

$$\widehat{\mathcal{I}} = \begin{cases} \sum_{s=1}^{n} \sum_{o=v,h,d} sign(\widehat{\omega_{s,o}}) \, \dfrac{\Delta_s \cdot (|\widehat{\omega_{s,o}}| + \delta)}{\widehat{\alpha}(\nu, r)} + (\widehat{c_n} + \delta) \cdot \Delta_n & |\widehat{\omega_{s,o}}| > 0 \\ 0, & \widehat{\omega_{s,o}} = 0 \end{cases} \tag{2}$$

For the sake of showing that the encoded VFWs are approximately equal to the decoded ones, that is $\alpha(\nu, r) \approx \widehat{\alpha}(\nu, r)$, we perform two experiments.

Experiment 1: Histogram of $\alpha(\nu, r)$ and $\widehat{\alpha}(\nu, r)$. The process of this short experiment is shown by Figure 5. Figure 5(a) depicts the process for obtaining losslessy both Encoded and Decoded visual weights for the 512×512 *Lena* image, channel Y at 10 meters. While Figures 5(b) and 5(c) shows the frequency histograms of $\alpha(\nu, r)$ and $\widehat{\alpha}(\nu, r)$, respectively. In both graphs, the horizontal axis represents the sort of VFW variations, whereas the vertical axis represents the number of repetitions in that particular VFW. The distribution in both histograms is similar and they have the same shape.



(a) (b) (c)

Fig. 5

(A) GRAPHICAL REPRESENTATION OF A WHOLE PROCESS OF COMPRESSION AND DECOMPRESSION. HISTOGRAMS OF (B) $\alpha(\nu, r)$ AND (C) $\widehat{\alpha}(\nu, r)$ VISUAL FREQUENCY WEIGHTS FOR THE 512×512 IMAGE *Lenna*, CHANNEL Y AT 10 METERS.

Experiment 2: Correlation analysis between $\alpha(\nu, r)$ and $\widehat{\alpha}(\nu, r)$. We employ the process shown in Fig. 5(a) for all the images of the CMU, CSIQ, and IVC Image Databases. In order to obtain $\widehat{\alpha}(\nu, r)$, we measure the lineal correlation between the original $\alpha(\nu, r)$ applied during the F-PiQ process and the recovered $\widehat{\alpha}(\nu, r)$. Table 2 shows that there is a

high similarity between the applied VFW and the recovered one, since their correlation is 0.9849, for gray-scale images, and 0.9840, for color images.

Table 2
CORRELATION BETWEEN $\alpha(\nu, r)$ AND $\widehat{\alpha}(\nu, r)$ ACROSS CMU, CSIQ, AND IVC IMAGE DATABASES.

Image Database	8 bpp gray-scale	24 bpp color
CMU	0.9840	0.9857
CSIQ	0.9857	0.9851
IVC	0.9840	0.9840
Overall	**0.9849**	**0.9844**

In this section, we only expose the results for the CMU image database.

Fig. 6 depicts the PSNR difference (dB) of each color image of the CMU database, that is, the gain in dB of image quality after applying $\widehat{\alpha}(\nu, r)$ at $d = 2000$ centimeters to the $\widehat{\mathcal{Q}}$ images. On average, this gain is about 15 dB. Visual examples of these results are shown by Fig. 7, where the right images are the original images, central images are perceptual quantized images after applying $\alpha(\nu, r)$ and left images are recovered images after applying $\widehat{\alpha}(\nu, r)$.

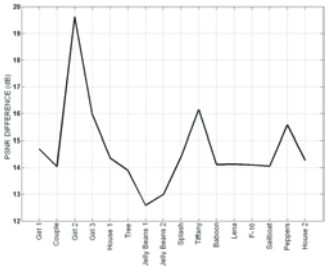
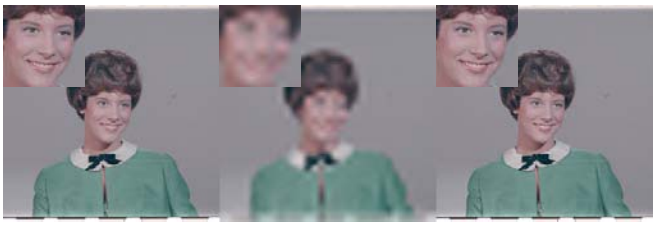


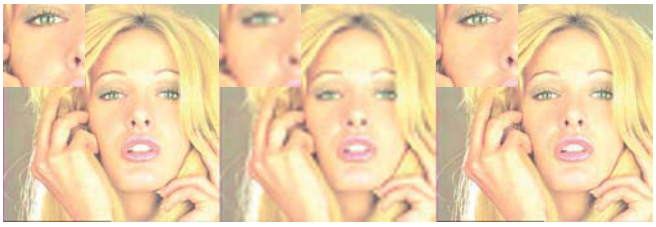
Fig. 6
PSNR DIFFERENCE BETWEEN $\widehat{\mathcal{Q}}$ IMAGE AFTER APPLYING $\alpha(\nu, r)$ AND RECOVERED $\widehat{\mathcal{I}}$ AFTER APPLYING $\widehat{\alpha}(\nu, r)$ FOR EVERY COLOR IMAGE OF THE CMU DATABASE.

After applying $\widehat{\alpha}(\nu, r)$, a visual inspection of these sixteen recovered images show a perceptually lossless quality. We perform the same experiment experiment for gray-scale and color images with $d =$ 20, 40, 60, 80, 100, 200, 400, 800, 1000 and 2000 centimeters, in addition to test their objective and subjective image quality by means of the PSNR and MSSIM metrics, respectively.

In Figs. 8 and 9, green functions denoted as F-PiQ are the quality metrics of perceptual quantized images after applying $\alpha(\nu, r)$, while blue functions denoted as I-PiQ are the quality metrics of recovered images after applying $\widehat{\alpha}(\nu, r)$. Thus,



(a) Girl 2



(b) Tiffany

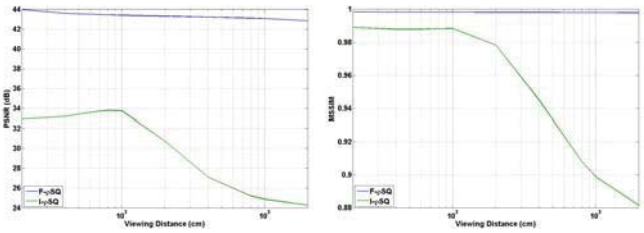
(c) Peppers

Fig. 7
VISUAL EXAMPLES OF PERCEPTUAL QUANTIZATION. LEFT IMAGES ARE THE ORIGINAL IMAGES, CENTRAL IMAGES ARE FORWARD PERCEPTUAL QUANTIZED IMAGES (F-PiQ) AFTER APPLYING $\alpha(\nu, r)$ AT $d = 2000$ CENTIMETERS AND RIGHT IMAGES ARE RECOVERED I-PiQ IMAGES AFTER APPLYING $\widehat{\alpha}(\nu, r)$.

either for gray-scale or color images, both PSNR and MSSIM estimations of the quantized image $\mathcal{Q}$ decrease regarding d, the longer d the greater the image quality decline. When the image decoder recovers $\widehat{\mathcal{Q}}$ and it is perceptually inverse quantized, the quality barely varies and is close to perceptually lossless, no matter the distance.

5. Conclusions

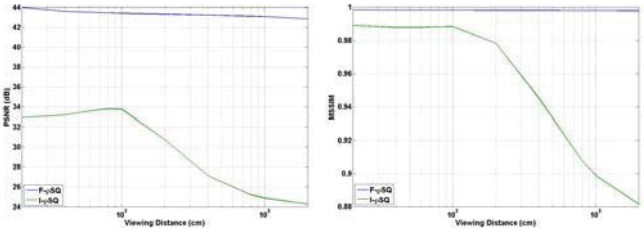
In this work, we defined both forward (F-PiQ) and inverse (I-PiQ) perceptual quantizer using CBPF. We incorporated it to Hi-SET, testing a perceptual image compression system χSET. In order to measure the effectiveness of the perceptual quantization, a performance analysis is done using thirteen assessments such as PSNR, MSSIM, VIF, WSNR or $\mathcal{NR}$PSNR, for instance, which measured the image quality between reconstructed and original images. The experimental results show that the solely usage of the Forward Perceptual Quantization improves the JPEG2000 compression and image perceptual quality. In addition, when both Forward and In-



(a) PSNR (b) MSSIM

Fig. 8

PSNR AND MSSIM ASSESSMENTS OF COMPRESSION OF GRAY-SCALE IMAGES (Y CHANNEL) OF THE CMU IMAGE DATABASE. GREEN FUNCTIONS DENOTED AS F-PiQ ARE THE QUALITY METRICS OF FORWARD PERCEPTUAL QUANTIZED IMAGES AFTER APPLYING $\alpha(\nu, r)$, WHILE BLUE FUNCTIONS DENOTED AS I-PiQ ARE THE QUALITY METRICS OF RECOVERED IMAGES AFTER APPLYING $\widehat{\alpha}(\nu, r)$.



(a) PSNR (b) MSSIM

Fig. 9

PSNR AND MSSIM ASSESSMENTS OF COMPRESSION OF COLOR IMAGES OF THE CMU IMAGE DATABASE. GREEN FUNCTIONS DENOTED AS F-PiQ ARE THE QUALITY METRICS OF FORWARD PERCEPTUAL QUANTIZED IMAGES AFTER APPLYING $\alpha(\nu, r)$, WHILE BLUE FUNCTIONS DENOTED AS I-PiQ ARE THE QUALITY METRICS OF RECOVERED IMAGES AFTER APPLYING $\widehat{\alpha}(\nu, r)$.

verse Quantization are applied into Hi-SET, it significatively improves the results regarding the JPEG2000 compression.

6. Acknowledgment

This work is supported by National Polytechnic Institute of Mexico by means of Project No. 20150508, the Academic Secretary and the Committee of Operation and Promotion of Academic Activities (COFAA), National Council of Science and Technology of Mexico by means of Project No. 204151/2013, and LABEX Σ-LIM France, Coimbra Group Scholarship Programme granted by University of Poitiers and Region of Poitou-Charentes, France.

References

[1] M. Boliek, C. Christopoulos, and E. Majani, *Information Technology: JPEG2000 Image Coding System*, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.

[2] S. I. P. I. of the University of Southern California. (1997) The USC-SIPI image database. Signal and Image Processing Institute of the University of Southern California. [Online]. Available: http://sipi.usc.edu/database/

[3] C. Research, École Polytechnique Fédérale de Lausanne, and Ericsson. (2001) JJ2000 implementation in Java. Cannon Research, École Polytechnique Fédérale de Lausanne and Ericsson. [Online]. Available: http://jj2000.epfl.ch/

[4] T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for formulation of quality measures incorporated in halftoning algorithms," *IEEE International Conference on Acustics, Speech and Signal Processing*, vol. 5, pp. 301–304, 1993.

Plankton Image Classification using Convolutional Neural Networks

Hussein A. Al-Barazanchi, Abhishek Verma, and Shawn Wang
Department of Computer Science, California State University, Fullerton, CA, USA

Abstract—*The study of plankton distribution is an important tool used for assessing the changes to marine ecosystem. Having a robust automated system for classification of plankton images plays an important role in advancing marine biology research. The images used in this study come from the SIPPER system. The challenges with SIPPER's plankton image dataset are the high degree of similarities between different classes, high variability within the same class, partial occlusion, and noise. Also, traditional computer vision techniques require tedious work to find suitable features to represent plankton. To overcome those issues, we propose the use of convolutional neural networks. Results of the experiments on SIPPER dataset show improvement in classification accuracy in comparison to other state of the art approaches. Another major advantage of our approach is the scalability for classification of new classes without the need for feature engineering.*

Keywords—plankton images, SIPPER system, convolutional neural networks, image search.

1 Introduction

The two main types of plankton: phytoplankton (drifting plants) and zooplankton (animal plankton) are considered as the main source of food for many aquatic animals. Also, carbon fixation by phytoplankton in the ocean plays an important role in the global carbon cycle. Due to their high ability to respond to changes in their environment: like pollution; plankton is considered as an alarm signal for detection of changes in marine ecosystem. Therefore, the fast mapping of plankton distribution is an important mission for oceanographic research.

In the early days, scientists were limited to the use of traditional techniques to investigate the distribution of plankton, such as Niskin bottles, towed nets, or pumps to collect samples. The counting and recognition of species was done by hand. As time progressed, use of cruise ships allowed researchers to collect bigger number of samples. However, the process of knowing the distribution of plankton remained laborious, time consuming and not elegant for real applications. Gradually, owing to the advancement in imaging technology several underwater devices for sampling were developed such as the HOLOMAR underwater holographic camera system [1], video plankton recorder (VPR) to [2], and the shadowed image particle profiling and evaluation recorder (SIPPER) [3]. With the use these instruments it became possible to perform continuous sampling of plankton. It was a major leap on the side of data collection, but the process of analysis remained tediously manual. In more recent times, the automated analysis of the pictures collected by these devices became feasible using sophisticated computer vision algorithms.

We can trace the first work on plankton image classification obtained by using VPR [4]. In 2005, Lue et al. [5] achieved 90% accuracy on plankton images recorded using SIPPER system. Their approach was based on classification with Support Vector Machine (SVM) and they did not make use of those image features that depend on the contour information. During the following year, a new shape descriptor was proposed by Tang et al [6] and the technique was named Normalized multilevel dominant eigenvector estimation, it achieved 91% recognition accuracy. Zhao et al. [7] extended the work in [6]; they make use of random sampling and multiple classifiers to achieve about 93% of recognition accuracy.

Regardless of the success shown by the aforementioned techniques, they suffer from one major drawback, which is total dependence on features engineering, i.e., the accuracy is determined by the quality of the used features. The process of feature engineering is difficult and requires much effort. Based on previous techniques, it requires extensive work to identify new classes of plankton; new features need to be identified, which could suitably represent those new classes. Hence, scaling up poses a challenge for those techniques.

In this paper we propose the use of convolutional neural networks (CNN), which is end to end learning framework. One major advantage of convolutional neural networks is its easy scalability to classify new classes. Based on the experimental results our proposed CNN

Table 1: Plankton Types Distribution

Class No.	Class Name	Count
0	Acantharia	131
1	Calanoid	172
2	Chaetognath	450
3	Doliolid	485
4	Larvacean	529
5	Radiolaria	563
6	Trichodesmium	789

algorithm exceeds the performance of the previous methods.

This paper is organized as follows. In section 2, we describe the plankton image dataset obtained from SIPPER system. In section 3, we give details of our proposed CNN algorithm for this classification task. Section 4 discusses our implementation and gives experimental results Concluding remarks appear in Section 5.

2 Plankton image dataset

The plankton images that we used in our experiment are provided by the University of South Florida (Tampa, FL, USA). They are captured by the SIPPER system. The images were collected during the years 2010 to 2014 from the Gulf of Mexico. The dataset contains 81 plankton types with more than 750 thousand images. In order to compare our method with the previous studies [5], [6], [7] we choose the exact same 7 types from the 81 types. Table 1 gives the names of these seven types and their distribution.

There are many challenges with plankton images represented by the differences between the species of the same class and similar appearance between different classes. Besides that, occlusion and deformation add more difficulty. Figure 1 gives a randomly chosen sample from the SIPPER dataset. A major issue is the need to find extra features to represent any extra classes added to the dataset. The classic solution to this problem is to do features engineering and to find useful features to represent the new class. To overcome this problem, we need a robust scalable approach toward feature extraction without depending upon features engineering and followed by robust classification. Our proposed solution uses a convolutional neural network.

3 Convolutional neural network

Visual recognition tasks require the construction of a suitable and robust feature set to represent the world around us. Those features should be invariant to outside variations of objects and keep enough relevant information to be able to recognize objects. The challenge is how to automatically learn such features without the need for human intervention. One approach is to simulate the process by which animals perform the task of object recognition and classification. Convolutional neural networks are proved to be the best model that simulates the vision abilities in animals with end to end feature learning and classification [9].

Convolutional neural networks are models that can learn invariant features and they are inspired from the vision mechanism in animals. This mechanism discovered during Hubel et al. [10] work on cat's visual cortex. Fukushima's Neocognitron [11] was the first simulated program based on this architecture. LeCun et al. [12] showed a successful use of convolution networks for handwritten recognition. Figure 2 illustrates the architecture used by LeCun et al. [12]. The popularity of convolutional neural networks started after the impressive success achieved in ImageNet Competition [13].

The typical design of convolutional neural network is stacked stages one after the other. These stages are followed at the end by a fully connected neural network.

Class Name	Sample Images		
Acantharia			
Calanoid			
Chaetognath			
Doliolid			
Larvacean			
Radiolaria			
Trichodesmium			

Figure.1. Random samples from seven classes of the SIPPER dataset.

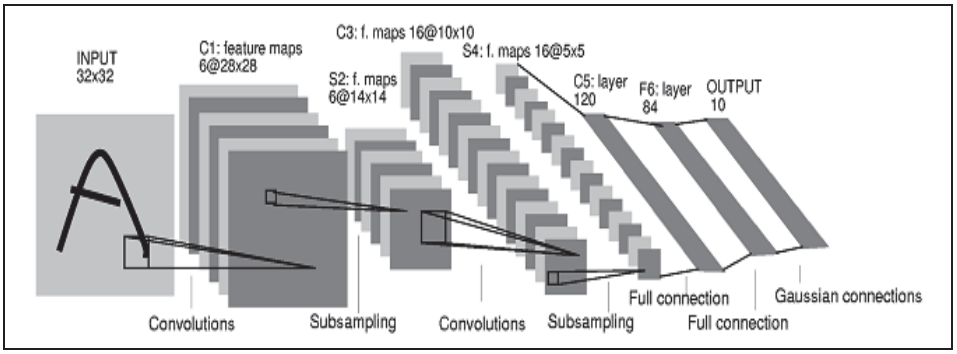


Figure 2. An example of convolutional neural network for handwritten recognition system [12].

The fully connected neural network works here as a classifier. At each level the convolutional neural networks consist typically of filters layer, non-linearity layer, and feature pooling layer [9] [12] [13]. The use of multi-level convolutional neural networks enables the system to learn the features' hierarchies. It starts from low-level features represented by the pixels, next it ascends to the mid-level features represented by edges and parts followed by the high-level features, which are objects.

3.1 Filter layer

The filter layer of the convolutional neural network is a variant form of neural networks in several aspects [15] [16]. First, neurons in convolutional neural network are sparsely connected to neurons in the next layer. On the other hand they are fully connected in regular neural networks. Second, convolutional neural networks' neurons follow a topographical layout. This means that connections are based on the related areas in the visual context. The regular neural networks do not make use of this feature. In our method, the images are fed to the convolutional layer in the format described in the equation (1). The symbols *h* and *w* refer to the height and width of the images while *c* refers to the number of color channels of the images.

$$h \times w \times c \tag{1}$$

$$y_j = b_j + \sum_i K_{ij} * x_i \tag{2}$$

We refer to each input to the layers as x_i. Where i is to indicate the filter number. Each component in the filters has the form x_{ijk}. The output will be computed by equation (2) [9]. The kernel (filter) K in the bank of filters has $K_n \times K_m$ dimensions depending upon the specified reception field; where n and m are the size of the reception field. Also, * indicates convolution operator while b is the network bias. Each kernel finds specific features at every place on the image. This means moving the kernel spatially will look for a particular feature in an image. As to which exact image feature a particular kernel will look for is decided dynamically by the algorithm [9].

3.2 Non-Linearity layer

The typical activations function for the output of neurons are *tanh()* and *sigmoid()* functions [9] [13], which are shown in equations (3) and (4). The problem with these activations is its slow speed when used with gradient descent. Using non saturating activation functions proves to be faster by many times of magnitude [13] [14]. We restrict our work to Rectified Linear Units (ReLUs), which is represented by equation (5).

$$sig(x) = 1 / (1 + e^{-x}) \tag{3}$$

$$\tanh(x) = (e^{2x} - 1) / (e^{2x} + 1) \tag{4}$$

$$f(x) = \max(0, x) \tag{5}$$

3.3 Pooling layer

Pooling is a technique for dimensionality reduction [16]. This layer aims to remove unrelated information and keeps only relevant ones [17]. The input to this layer is the output of the non-linearity layer. The output of this layer is the reduced version of the input [15]. This layer has pool units that are organized in topographical way and connect to local areas in the input coming from the non-linearity layer.

The replication of neurons' weights in the filter bank helps to detect features in the different regions of the image. The problem that arises in those features is that they are not translation invariant. The pooling is used to make the features invariant to translation in the input. Pooling helps to reduce the sensitivity of activations in neural network to the pixels' locations and the neural network structure [18]. The common functions used in pooling are maximum and average functions and they are usually named max-pooling and average-pooling. There are two different ways to feed the input to those functions,

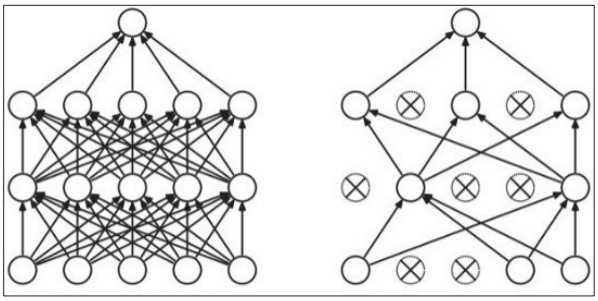


Figure 3. On the left fully connected neural network. On the right neural network after the dropout [19].

which could be either the separate or overlapping mode [15].

3.4 Dropout layer

Dropout is a recent technique developed by Srivastava et al. [19]. The purpose of this layer is to reduce the problem of overfitting and enhance generalization on the test data. This method works by removing random neurons with their connections during the process of learning. Fig 3 shows an example of the dropout layer.

3.5 Output layer

The output layer is different from all aforementioned layers. The output of this layer is in the form of probabilities that sum to one. The probability values indicate the confidence level about the chosen class where higher value means higher confidence. The common function used in this layer is the *softmax* function. This function is linear and it uses the log probability.

3.6 Learning algorithm

We used backpropagation algorithm for learning and

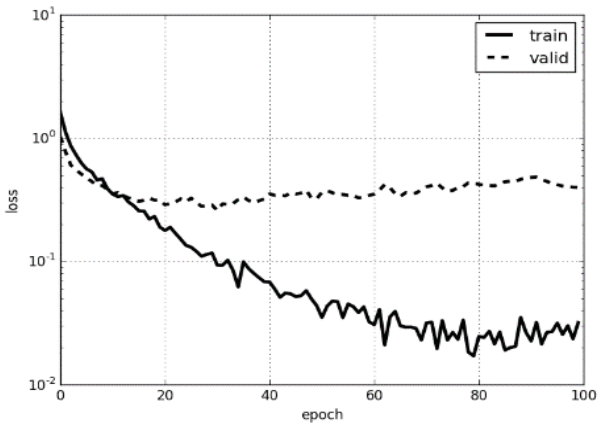


Figure 4. Training and validation loss for the highest accuracy in 1-layer convolutional neural network.

Table 2: Classification Accuracy for 1-Layer CNN

No of Filters	Reception Field	Accuracy (%)
8	2*2	88.90
8	3*3	90.71
8	4*4	90.18
8	**5*5**	**90.92**
16	2*2	89.86
16	3*3	89.75
16	**4*4**	**90.18**
16	5*5	90.07
32	2*2	88.68
32	3*3	89.75
32	4*4	92.38
32	**5*5**	**92.39**

stochastic gradient descent for optimization. We set the batch size to 32. Initially the momentum and learning rate are set to 0.9 and 0.01 respectively. The momentum and learning rate are continually updated as we get close to the minima. Also, we used a technique called early stopping [20] [21] to prevent overfitting problem in the neural network.

4 Implementation and experimental results

The total number of images in the seven class subset from the SIPPER dataset is 3119. The data used in the experiments is randomly divided into training, testing, and validation. Training consists of 56% of the images from each class, and testing and validation data is 30% and 14% respectively from each class. This gives us a total of 1745 samples for training, 437 samples for validation, and 937 samples for testing.

We divided our experiments into two phases. For the first phase we use only one convolutional layer and then extend the idea into the second phase with two convolutional layers. To standardize the testing results, we set many hyper parameters to fixed values. We used a fixed size classification layer, which is 2 fully connected layers. The first fully connected layer has 256 neurons while the second layer has 128 neurons. Each of them is followed by a 50% dropout layer. The output layer has the same number of classes which is 7 followed by the *softmax* activation function. All the convolutional layers and fully connected layers are followed by a ReLu activation function and pool layer. We limit the number of hyper-parameters to be tuned to the number of layers, number of filters in each layer, and the size of the reception field.

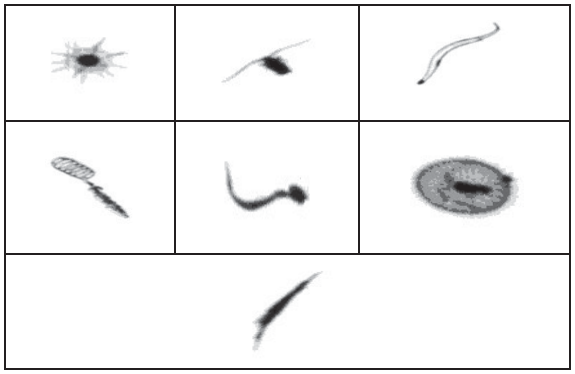


Figure 5. Correctly classified example images from 1-layer CNN, from left to right and top to bottom the images are from class number 0-6.

4.1 Phase one: 1-layer CNN

To simplify the experiments in this phase, we start with only one layer for the convolutional neural network. We set the number of filters to 8, 16, and 32. The reception field is set between 2 and 5. Table 2 shows the accuracy details related with each of our configuration. We highlight the best accuracy associated with different number of filters. Our results show that our algorithm performs better than what is achieved in [5] and [6].

Overfitting is a problem in the neural network. In this case, the neural network starts overfitting on the training data giving higher accuracies while the accuracy for validation and testing data start to drop. We utilize several techniques to stop overfitting such us pooling and dropping layers. In addition, we use the aforementioned early stopping technique in conjunction with pooling and dropping methods to achieve higher testing accuracy. Figure 4 shows the loss function values associated with the number of iterations. Figure 4 relates to the highest classification accuracy in 1-layer CNN. This figure explains that with higher number of iterations the loss function for the training continues to drop while the loss function for validation keeps on increasing.

Table 3: Confusion Matrix for 1-Layer CNN

Class No.	0	1	2	3	4	5	6	Classification Accuracy (%)
0	39	0	0	0	0	0	0	100.00
1	0	51	0	0	1	0	0	98.07
2	0	0	121	10	2	0	2	89.62
3	0	0	13	131	1	0	1	89.72
4	0	0	0	0	153	0	6	96.22
5	0	0	0	0	0	169	0	100.00
6	0	2	5	11	14	0	205	86.49
Overall Classification Accuracy					**92.74 %**			

Table 4: Classification Accuracy for 2-Layers CNN

No of Filters	Reception Field	Accuracy (%)
8	2*2	90.82
8	3*3	91.14
8	4*4	90.92
8	5*5	90.82
16	2*2	90.60
16	3*3	89.75
16	4*4	89.96
16	5*5	91.88
32	2*2	92.52
32	3*3	92.31
32	4*4	93.38
32	5*5	**94.26**

Figure 5 shows randomly chosen examples of correctly classified types with the 1-layer CNN. Those examples include all the seven plankton types. The confusion matrix for the 1-layer CNN based on 32 filters and 5*5 reception field for one particular cross fold with accuracy rate of 92.74% is shown in Table 3. We perform 3 cross validation for 1-layer CNN with 32 filters and 5*5 reception field to get the average accuracy rate of 92.39% as show in Table 2.

4.2 Phase two: 2-layer CNN

In this phase our focus is to add another convolutional layer. Depending upon the results we got from phase one, we chose the configuration with the best accuracy rate to be the setting for the first convolutional

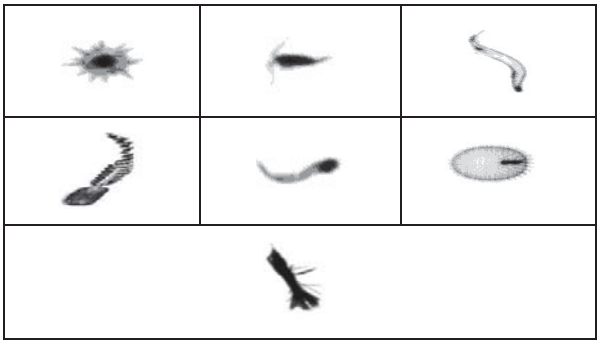


Figure 6. Correctly classified example images from 2-layers CNN, from left to right and top to bottom the images are from class number 0-6.

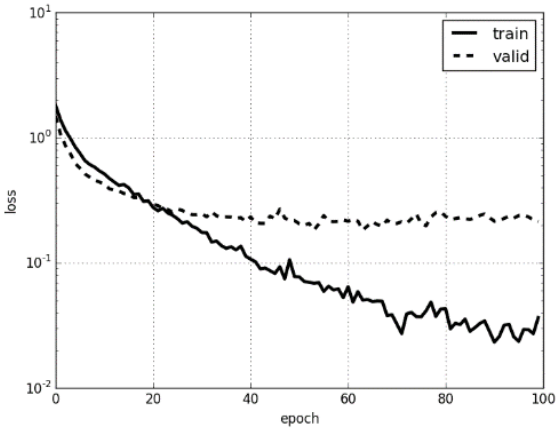


Figure 7. Training and validation loss for the highest accuracy in 2-layer CNN.

layer. In phase 2 we tune the second convolutional layer with different reception field and with different number of filters.

As shown in Table 4, overall we got better results with the 2-layers CNN in comparison with 1-layer CNN. Figure 6 shows randomly chosen examples of correctly classified types with the 2-layers CNN. Those examples include all the seven plankton types. The confusion matrix for the 2-layera CNN based on 32 filters and 5*5 reception field for one particular cross fold with accuracy rate of 94.55% is shown in Table 5. We perform 3 cross validation for 2-layers CNN with 32 filters and 5*5 reception field to get the average accuracy rate of 94.26% as show in Table 4.

Figure 7 relates to the highest classification accuracy in 2-layers CNN. This figure explains that with higher number of iterations the loss function for the training continues to drop while the loss function for validation keeps on increasing. It also suggests that the 2-layers CNN requires more training epochs to start converging than what is required by 1-layer CNN.

Table 5: Confusion Matrix for 2-Layers CNN

Class No.	0	1	2	3	4	5	6	Classification Accuracy (%)
0	39	0	0	0	0	0	0	100.00
1	0	51	0	0	1	0	0	98.07
2	0	0	125	9	1	0	0	92.59
3	0	0	15	129	1	0	1	88.35
4	0	1	1	0	155	0	2	97.48
5	0	0	0	0	0	168	1	99.40
6	0	3	2	6	7	0	219	92.40
Overall Classification Accuracy (%)					94.55 %			

Table 6: Comparison of the Classification Performance with other Methods on the SIPPER dataset

Method	Accuracy (%)
Normalized Multilevel Dominant Eigenvector Estimation [6]	91.70
Bagging Based [7]	93.04
Random Subspace [8]	93.27
Our proposed 2-layers CNN method	**94.26**

Table 6 gives a comparison of the classification performance with other methods on the SIPPER dataset. We obtain classification performance of 94.26%, which is better than other previous methods.

5 Conclusions and future work

The study of plankton distribution is an important tool used for assessing the changes to marine ecosystem. Efficient analysis and classification of huge amounts of plankton data requires robust algorithms. Traditional computer vision techniques require tedious work to find suitable features to represent plankton. In our paper, we proposed the use of convolutional neural networks. Results of the experiments using the SIPPER dataset show improvement in classification accuracy in comparison to the previous approaches from other research groups. Another major advantage of our approach is the scalability for classification of new classes without the need for feature engineering.

In the future we plan to further improve the performance of our method by expanding the number of layers in the convolutional neural network. We also plan to explore the combination of convolutional neural network and other classification algorithms such as SVM or Random Forest to improve the overall efficiency of the classification methodology.

Acknowledgment

We would like to thank Dr. Kendra L. Daly, Andrew Remsen, and Kurt Kramer (USF) for the validated SIPPER image dataset. The SIPPER imaging work was supported by a National Science Foundation grant OCE-0526545, a University of South Florida Sponsored Research Foundation grant, a Florida Institute of

Oceanography/BP grant, and a Gulf of Mexico Research Initiative (GOMRI) grant to K.L. Daly.

6 References

[1] J. Watson, G. Craig, V. Chalvidan, J. P. Chambard, A. Diard, G. L. Foresti, B. Forre, S. Gentili, P. R. Hobson, R. S. Lampitt, P. Maine, J. T. Malmo, H. Nareid, G. Pieroni, S. Serpico, K. Tipping, and A. Trucco, "High resolution in situ holographic recording and analysis of marine organisms and particles (Holomar)," in Proc. IEEE Int. Conf. OCEANS, 1998, pp. 1599–1604.

[2] S. Davis, S. M. Gallager, and A. R. Solow, "Microaggregations of oceanic plankton observed by towed video microscopy," Science, vol. 257, pp. 230–232, Jul. 1992.

[3] S. Samson, T. Hopkins, A. Remsen, L. Langebrake, T. Sutton, and J. Patten, "A system for high-resolution zooplankton imaging," IEEE J. Ocean. Eng., vol. 26, no. 4, pp. 671–676, Oct. 2001.

[4] X. Tang, W. K. Stewart, L. Vincent, H. Huang, M. Marty, S. M. Gallager, and C. S. Davis, "Automatic plankton image recognition," *Artif.Intell. Rev.*, vol. 12, pp. 177–199, 1998.

[5] T. Luo, K. Kramer, S. Samson, A. Remsen, D. B. Goldgof, L. O. Hall, and T. Hopkins, (2004, August). "Active learning to recognize multiple types of plankton". In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* (Vol. 3, pp. 478-481). IEEE.

[6] X. Tang, F. Lin, S. Samson, and A. Remsen, "Binary plankton image classification," *IEEE J. Ocean. Eng.*, vol. 31, no. 3, pp. 728–735, Jul. 2006.

[7] F. Zhao, F. Lin, and H. Seah, "Bagging based plankton image classification," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 2517–2520.

[8] L. Zhifeng, *Z.* Feng, L. Jianzhuang, *Q.* Yu, "Pairwise Nonparametric Discriminant Analysis for Binary Plankton Image Recognition," *IEEE J. Ocean. Eng.*, vol. 39, no. 4, pp. 695–701, 2014.

[9] Y. LeCun, K. Kavukcuoglu, and C. Farabet, (2010, May). Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on* (pp. 253-256). IEEE.

[10] D. Hubel, and T. Wiesel, (1968). Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology (London), 195, 215–243.

[11] K. Fukushima, (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36, 193–202.

[12] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[13] A. Krizhevsky, I. Sutskever and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Proc. Neural Information and Processing Systems,*2012.

[14] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proc. 27th International Conference on Machine Learning, 2010.

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[16] J. Van, "Analysis of Deep Convolutional Neural Network Architectures". 2014. Retrieved from http://referaat.cs.utwente.nl/conference/21/paper/7438/analysis-of-deep-convolutional-neural-network-architectures.pdf.

[17] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 111–118, 2010.

[18] M. D. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557, 2013.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, pages 1929–1958, 2014.

[20] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria", Neural Netw., 11 (1998), pp. 761–767

[21] L. Prechelt, (1998). "Early stopping-but when?". In *Neural Networks: Tricks of the trade* (pp. 55-69). Springer Berlin Heidelberg.

3D Segmentation of MRI Brain Tumor Using Fast Level Set Method

M.Abdelaziz [*], F.Alim.Ferhat-taleb[*], L.Ait mohamed [*], Y.Cherfa [**], B.Hachemi [*], F.Talbi [*],S.Seddiki[*]

[*]Centre de Dveloppement des Technologies Avances, Division Architecture des Systmes et Multimdia, Algiers, Algeria

[**]Department of Electronics, University Saad Dahlab Blida-1, Blida, Algeria

mabdelaziz@cdta.dz, falim@cdta.dz

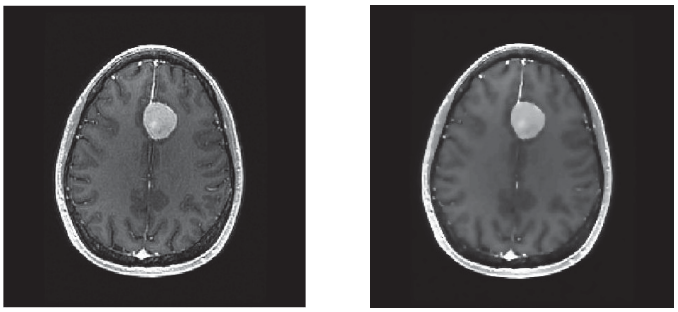
***Abstract*—The early diagnosis and treatment of tumors can make a second chance for life to more than 8.2 million people worldwide died from cancer juste in 2012 according to the world health organization, to help those people we work in a project aim to develop computer-aided detection/diagnosis (CAD) systeme, one of the crucial parts of this system is the segmentation of the tumor, a good tumor segmentation allow to calculate an accurate volume, localise and visualize it in 3D and also has a direct effect on the classification process, the modified level set by Fang dong is our chosen methode due its efficiency in the Intensity inhomogeneity regions wich is frequently happen in medical imaging and make difficulties for the physicians and Surgeons to analyse the images and take decisions especially in the surgical removal of the brain tumor where the mines error could make fatal consequences for the health and the life of the patient, we focus in this paper on the chosen segmentation methode theory and the results of the implementation on real 3D MRI data with tumor.**

I. Introduction

Primary brain tumors can be either malignant (contain cancer cells) or benign (do not contain cancer cells). A primary brain tumor is a tumor which begins in the brain. If a cancerous tumor which starts elsewhere in the body sends cells which end up growing in the brain, such tumors are then called secondary or metastatic brain tumors. In order to diagnose tumors the medical experts use different medical imaging techniques such as MRI (magnetic resonance imaging), CT(computed tomography) and PET (positron Such imaging techniques present the detailed anatomical structure into multiple 2D (two dimensional) images. But 2D images cannot accurately convey the complexities of human anatomy and interpretation of 2D complex anatomy requires special training. Although radiologists are trained to interpret these images, they often find difficulty in communicating their interpretations to a physician, who may have difficulty in imagining the 3D anatomy. However using image processing tasks like segmentation and visualization, 3D reconstruction of MR images is possible. In order to segment a tumor from a cancerous brain, a number of methods are proposed in the literature, the 3D segmentation of brain tumors is extensive and recently, we have Ria et al [1] using the 3D Slicer platform for localization of the cancerous tumor using then histogram and thresholding, Ana et al. [2] apply morphological operators and Otsu thresholding then they using bicubic interpolation for 3D reconstruction, Hou et al. [3] uses the segmentation of the tumor by classification K-means clustering then extract two futures: the circularity ratio and intensity NG to reduce the number of slice to reconstruct for better visualization, the other hand Kik Ron et al. [4] directly uses the platform of image processing 3D slicer for segmentation and 3D visualization we have also Yao Chen Tien et al. [5] that integrates Bayesian methods and level set for segmentation and calculation of tumor volume and the 3D Slicer for Reconstruction slices. Compared to other method, level set algorithm can attain better results for medical image segmentation. Taheri et al. [6] proposed a method for 3D brain tumor segmentation as threshold-based level set segmentation which use a speed function that does not need density function estimation. Bernard et al. [7] presented a formulation of active contours based on level set where the implicit function is modeled as a continuous parametric function expressed on a B-spline basis. Ben Ayed and Mitiche [8] advanced a curve evolution method anabling the effective number of regions to vary during curve evolution. This level set functional used a region merging prior embedding an implicit region merging in curve evolution. This work is a part of a project aim to create a computer- aided diagnosis system of human brain tumor, wich include two main part the segmentation and the classification, in this paper we will focus on the 3D segmentation part. We choose to use the modified Level set method by Fangfang Dong [9] for the 2D silces segmentations, especially when this method give us a good result with inhomogeneous images due to the use of the local intensity information, those results will be used to create the 3D reconstruction of the tumor by the use of an algorithm implemented with the visual toolkit (VTK).

II. Applied Methodologies

- **Pretreatment**: we applied an anisotropic Filter [10] to the 2D MRI Slices



(a) Original 2D Slice (b) The Filtred Slice

Fig. 1: Pretraitement

- **Segmentation**: We use a deformable model-Level set for the segmentation of the brain tumor

a) Level-set:
The geometric model of active contours implements a deforming curve in time and space to achieve the boundaries of an object to be detected in an image I (x,y) [11].The curve is deformed:
– According to its normal
– At a rate proportional to its curvature
Noting C the curve, $\vec{N}$ the inner normal of the curve and F a velocity term depending of curvature κ, the evolution equation is of the form:

$$\frac{\partial C}{\partial t} = F\vec{N} \tag{1}$$

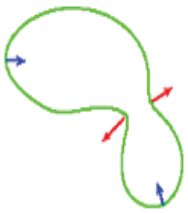


Fig. 2: A deforming curve according to its curvature

b) Level set formulation of Fangfang Dong [9] :
Using the variational level set method, the total energy $E(C)$ can be minimized by solving the following equation of the level set function ϕ

$$\frac{\partial \phi}{\partial t} = \left(w(X) \frac{I(X) - \frac{C_1+C_2}{2}}{C_1+C_2} + (1-w(X)) \frac{I_0(X) - \frac{m_1+m_2}{2}}{m_1-m_2} + \gamma div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right)\delta(\phi) \tag{2}$$

where:

$$c_1 = \frac{\int_\Omega IH(\phi)\,dx}{\int_\Omega H(\phi)\,dx}, c_2 = \frac{\int_\Omega I(1-IH(\phi))dx}{\int_\Omega (1-IH(\phi))dx} \tag{3}$$

$$m_1 = \frac{\int_\Omega I_0 H(\phi)\,dx}{\int_\Omega H(\phi)\,dx}, m_2 = \frac{\int_\Omega I_0(1-IH(\phi))dx}{\int_\Omega (1-IH(\phi))dx} \tag{4}$$

Because the evolution equations $\phi t = F\delta(\phi) and \Phi t = F|\Delta\phi|$
have the same steady states, we draw upon the following equation for the level set evolution

$$\frac{\partial \phi}{\partial t} = \left(w(X) \frac{I(X) - \frac{C_1+C_2}{2}}{C_1+C_2} + (1-w(X)) \frac{I_0(X) - \frac{m_1+m_2}{2}}{m_1-m_2} + \gamma div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right)|\nabla\phi| \tag{5}$$

However, the computations of nonlinear PDEs (2) and (5) are all time-consuming due to the divergence term, which is the mean curva- ture of the curve represented by the level set function. To avoid the de-nominator in the divergence term to be zero,$|\Delta\phi|$ always needs to be regularized, i.e., $|\nabla\phi| \approx \sqrt{|\nabla\phi|^2 + \varepsilon}$ then together with this condition, Eq. (5) can be simplied as

$$\frac{\partial \phi}{\partial t} = w(X) \frac{I(X) - \frac{C_1+C_2}{2}}{C_1+C_2} + (1 - w(X)) \frac{I_0(X) - \frac{m_1+m_2}{2}}{m_1-m_2} + \gamma\triangle\phi \tag{6}$$

or equivalently,

$$\frac{\partial \phi}{\partial t} = \frac{1}{\gamma}[w(X) \frac{I(X) - \frac{C_1+C_2}{2}}{C_1+C_2} + (1 - w(X)) \frac{I_0(X) - \frac{m_1+m_2}{2}}{m_1-m_2}] + \triangle\phi \tag{7}$$

c) Implementation:
According to above analysis, the detailed numerical schemes are summarized as follows

1. Initialize the level set function ϕ as:

$$\phi = \begin{cases} 1\ if\ X \in \Omega_1 \\ 0\ if\ X \in C \\ -1\ if\ X \in \Omega_2 \end{cases} \tag{8}$$

2. Update $c1$,$c2$,$m1$ and $m2$ by Eqs. (3) and (4)
3. Evolve the level set function ϕ by the Eq. (7)
4. Regularize the resulted level set function with a Gaussian fillter,i.e., $\phi_1 = G_\sigma * \phi$, where G_σ is a Gaussian kernel, and the parameter σ controls the smoothness of the contour

5. Let ϕ= 1 if $\phi_1 \geq 0$; Otherwise, ϕ= -1
6. Check whether the evolution of the level set function has converged.If not, return to step 2

In this method, the traditional level set function is substituted by a binary level set function. Furthermore, the level set function only needs to be simply initialized to constants, which have different signs inside and outside of the contour. It is very simple to implement in practice.

In our experiments, we employ a stopping criterion recently proposed in [12] for image segmentation. The iteration will be stopped automatically when the change of the curve length keeps smaller than a prescribed threshold η, i.e., if the following criterion

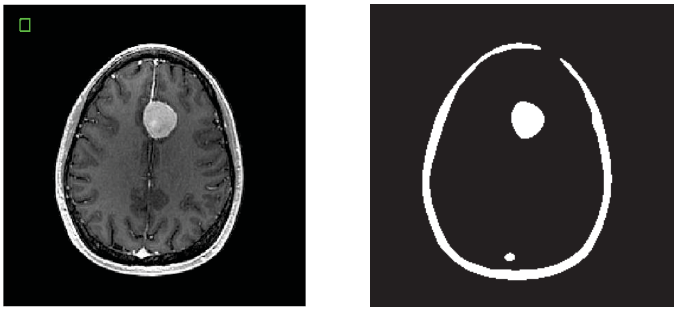
$$\left|Length\left(C^{\eta+1}\right) - Length\left(C^{\eta}\right)\right| < \eta \qquad (9)$$

or the approximate level set form

$$\left|\int_{\Omega} \delta\left(\phi^{\eta+1}\right)\left|\nabla\phi^{\eta+1}\right| dx - \int_{\Omega} \delta\left(\phi^{\eta}\right)\left|\nabla\phi^{\eta}\right| dx\right| < \eta \qquad (10)$$

is satisfied, the iteration will be stopped. Here η denotes the iteration number and $\delta(\phi)$ = H$'(\phi)$

The implementation of the original method developed by Fangfang make a complete segmentation of the MRI 2D Image as shown in Fig 3



(a) Original 2D Slice (b) Segmentation Result

Fig. 3: Segmentation Result with the original Algorithm

To avoid this situation we decide to take the initialization in one slice as ROI, but for the rest of slices our algorithm will automatically keep only the Tumor Segments

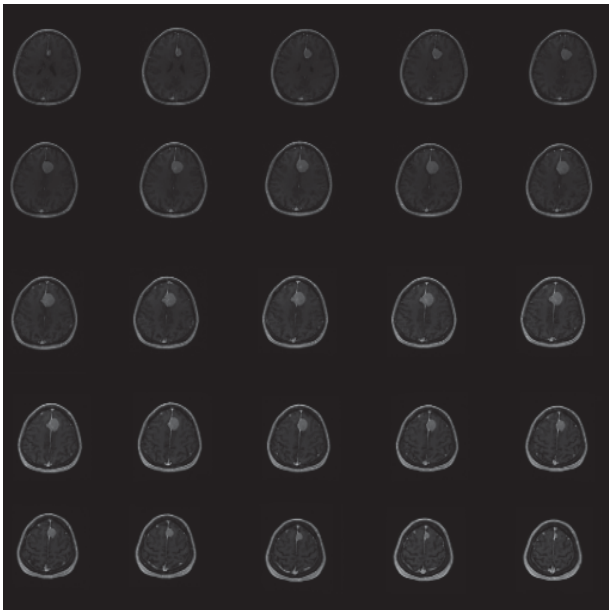


Fig. 4: Originals Slices

Fig. 5: Segmented Tumor

Our algorithme segment all the slices of the volume, at the end we get all the Tumor segments in 2D, now we can reconstruct the Tumor in 3D

d) **3D Reconstruction**: We choose to use The Visual Toolkit (VTK) library to reconstruct the Tumor in the Brain in 3D

Fig. 6: 3D Reconstruction of the Tumor

e) **Calcul of the Tumor's volume**
The 3D reconstructed Tumor in Fig.6 has 15286 mm^3 as a volume equivalent to 15.286 cm^3

f) **Localisation of the Tumor**
To Help the surgeons and the physicians to localise the tumor we decide to make the 3D reconstructed tumor in the 3D reconstructed brain which make a comprehensive and clear visualisation

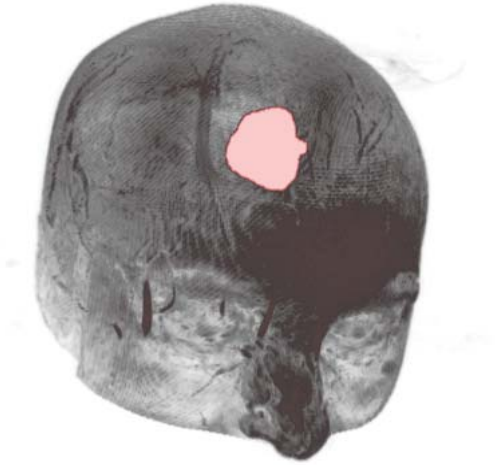


Fig. 7: 3D Reconstruction of the Tumor in the Brain

III. Conclusion and future work

This paper present the 3D segmentation using the fast level-set method and the visual toolkit for the reconstruction, which is a part of project aim to develop computer-aided diagnosis system of human brain tumor, the algorithm is tested on 3D MRI data, Our future work is to extend our approach to have a classication system for the tumor (benign,malignant), the complete system will be distributed for free to our Hospitals.

IV. Acknowledgment

We would like to thank Prof.Bakhti a neurosurgeon in Mustapha Bacha Hospital for providing us the necessary Data and expertise assistant in our research project (FNR 2014/2016).

References

[1] I. Riazul, A. M.Abdullah, M. H. Bhuiyan, and S. M. M. Rahman, "Segmentation and 3D Visualization of Volumetric Image for Detection of Tumor in Cancerous Brain," *IEEE International Conference on Electrical and Computer Engineering Dhaka, Banglades*, 2012.

[2] A.Anantatamukala, G.Abhijeet, and K.Yogesh, "A Systematic Algorithm for 3-D Reconstruction of MRI based Brain Tumors using Morphological Operators and Bicubic Interpolation," *International Conference on Computer Technology and Development (ICCTD)*, 2010.

[3] H. M. Moftah, A. E. Hassanien, and M. Shoman, "3D Brain Tumor Segmentation Scheme using K-mean Clustering and Connected Component Labeling Algorithms," *IEEE*, 2010.

[4] R. Kikinis and S. P. M.D, "3D Slicer as a Tool for Interactive Brain Tumor Segmentation. 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA," *IEEE*, 2011.

[5] Y.-T. Chen, "Brain Tumor Detection Using Three-Dimensional Bayesian Level Set Method With Volum Rendering," *IEEE International Conference on Wavelet Analysis and Pattern Recognition, Xian*, 2012.

[6] S. Taheri, S. Ong, and V. Chong, "Level-set segmentation of brain tumors using a threshold-based speed function," *Image and Vision Computing*, vol. 28, pp. 26–37, 2010.

[7] O. Bernard, D. Friboulet, P. Thevenaz, and M. Unser, "Variational B-spline level-set: A linear filtering approach for fast deformable model evolution," *IEEE Transactions on Image Processing*, vol. 18, pp. 1179–1191, 2009.

[8] I. B. Ayed and A. Mitiche, "A region merging prior for variational level set image segmentation," *IEEE Transactions on Image Processing*, vol. 17, pp. 2301–2311, 2008.

[9] F. Dong, Z. Chen, and J. Wang, "A new level set method for inhomogeneous image segmentation," *Image and Vision Computing*, vol. 31, pp. 809–822, 2013.

[10] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629–639, 1990.

[11] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 2, pp. 269–277, 1995.

[12] X.-F. Wang, D.-S. Huang, and H. Xua, "An efcient local ChanVese model for image segmentation," *Pattern Recognit*, vol. 43, p. 603618, 2010.

Enhanced Video Target Tracking using Kalman Filter Guided Covariance Descriptor with Gaussian Similarity Weighting

O. Akbulut[1], **S. Erturk**[2]
[1]Computer Engineering Department, Kocaeli University, Kocaeli, Turkey
[2]Electronics and Telecommunications Engineering Department, Kocaeli University, Kocaeli, Turkey

Abstract – *The region covariance descriptor, which includes statistical and spatial features as well as correlation between features, has been widely used for target representation in visual tracking. Robustness, enabling fusion of several features, low-computational load are powerful features of the region covariance descriptor for target representation. In this paper, we have proposed a novel approach in that isotropic Gaussian weighting and Kalman filtering is used together with the region covariance descriptor which increases performance of visual tracking in relatively complex situations such as occlusion, appearance changes etc. Experimental results demonstrate the effectiveness of this approach in terms of robust visual tracking.*

Keywords: Covariance Descriptor, Gaussian weighting, Kalman filter

1 Introduction

Target tracking is an important topic in computer vision applications. The performance of the tracking process depends on some special conditions such as semi/full occlusion, illumination conditions, noise factor, appearance changes and dynamic background in many cases. For example, using a fixed appearance model for a target region may degrade visual tracking performance.

In visual tracking, the selection of appropriate features, which represent different characteristic properties of the target, influences the quality of visual tracking. Conventionally, a single feature descriptor such as the color histogram or gradient based histogram has been used to represent the target for tracking purposes. In [1], color histogram based object representation is used for tracking. A kernel based search approach has been carried out via mean shift algorithm. However, performance degradation might occur if there is an abrupt motion in the scene. A more effective tracking performance based on particle filter has been realized in [2]. Color distributions are integrated into particle filtering to incorporate the scale changes and abrupt motion seen in tracking. Color histograms are robust to partial occlusion and they have rotation and scale invariant features. However, only statistical information is considered in most cases whereas spatial information is ignored. Furthermore, using color histogram is not actually effective for higher number of bins due to exponential size. The histogram of gradient (HOG) descriptor is well-known single feature descriptor used in visual tracking. In [3], the histogram of gradients has been utilized for multiple pedestrian tracking. The local binary pattern (LBP) that is a texture based alternative single feature descriptor has been used in visual tracking in [4]. Exploiting local binary codes for each pixel in target region, histogram based representation is computed. However, tracking performance of LBP based algorithm has not been investigated for challenging sequences. Instead of using a single feature descriptor, multi-feature descriptors have also been used for target representation in the literature. These approaches can improve the target tracking performance and achieve robustness at the expense of computational load as shown in [5-7]. In [5], intensity gradients and color histograms based elliptical head tracking approach has been proposed. Intensity gradients are calculated along the person's head boundary while color histograms are computed at the interior of the heads.

Tuzel et al. [6,7] proposed a region covariance descriptor (RCD) based appearance model for target tracking and object detection. RCD enables efficient fusion of multiple feature descriptors maintaining the computational load of a single feature descriptor. RCD has been used in different computer vision areas such as human detection [8,9] and face recognition [10]. Instead of using RCD based deterministic search in tracking problems, RCD based probabilistic search [11-13] has also been utilized to achieve better tracking performance. In [11], a particle filter based probabilistic target tracking has been proposed using RCD to represent target region. An incremental model-updating scheme has been integrated into the existing approach to handle appearance changes occurred in target by enabling low computational load. A Monte Carlo method, which is a special case of particle filter, combined with RCD has been proposed to track non-rigid objects in [12].They use multiple covariance matrices for a target region instead of single covariance matrix for partial occlusion. In [13], system dynamic models related with tracking process have been defined in Riemannian manifold and covariance based tracking has been performed. The state samples have been drawn on the manifold geodesics instead of vector space. The selection and number of different types of feature vectors used in RCD may affect the

performance of target tracking. In [14], performance evaluation for choosing feature vectors to construct the best covariance descriptor has been evaluated. It is emphasized that color features are more important than the other features to construct covariance matrix. The RCD approach has also been utilized in multi target tracking applications as in [15]. In [15], an easy model update approach, which is based on the mean of the last covariance matrix and current covariance matrix, is used to speed up the tracking process. In [16], covariance descriptors has been used for 3D shape matching and retrieval taking into account the geometry of the Riemannian manifold.

In this paper, a novel approach has been proposed to improve the tracking performance of the RCD based approach. There are three fundamental contributions of the proposed approach. The first one is to use isotropic Gaussian shaped weighting for the similarity metric used in covariance matrix matching to give more weight to locations with higher probability. Additionally, Kalman filtering [17] has been integrated to the proposed approach for determining the probable target region in the corresponding frame to guide the Gaussian weighting (and search) center. As a last contribution, occlusion detection has been integrated by detecting suddenly increasing distance errors of the target object.

2 Region Covariance Descriptor

RCD has a discriminative property to represent the appearance model of the object region for detection and tracking applications. RCD enables efficient fusion of multiple feature vectors. It is possible to construct multivariate data by extracting different features for each pixel of the image data such as intensity, color, texture etc. Each column and each row in the multivariate data represents an observation and a feature respectively. Note that, each observation corresponds to a pixel. The multivariate data can be denoted by

$$F = \left[\mathbf{f}_1 \quad .. \quad \mathbf{f}_i \quad .. \quad \mathbf{f}_n\right], \tag{1}$$

where $\{\mathbf{f}_i\}_{i=1}^{n}$ are d dimensional feature column vectors associated with the pixel index i, and n is the number of total pixels in the image. Feature vectors generated for each pixel are illustrated for a sample object region in Fig. 1.

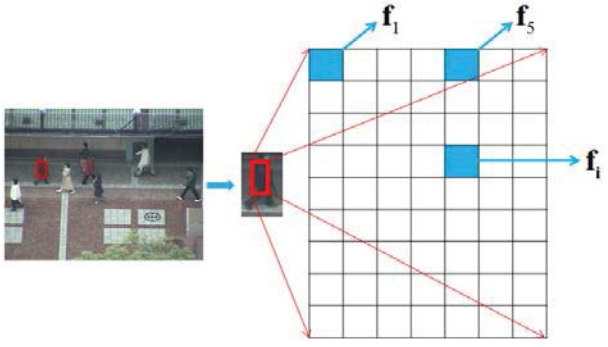


Fig. 1 Feature vectors for each pixel in an object region.

It is clear that, the distribution of the multivariate data can be interpreted by constructing the covariance matrix. From this point of view, a region covariance matrix can be defined for a sub-region R inside the image that represents the appearance descriptor of the target object. The representation of the region covariance matrix is given by

$$C_R = \frac{1}{s-1}\sum_{i=1}^{S}(\mathbf{f}_i - \boldsymbol{\mu}_R)(\mathbf{f}_i - \boldsymbol{\mu}_R)^T, \tag{2}$$

where C_R is $d \times d$ symmetric positive definite matrix (SPD), $\boldsymbol{\mu}_R$ is the mean of the corresponding feature vector and s is the number of pixels within the sub-region. The covariance descriptor encapsulates statistical information of various feature vectors and the size of the covariance descriptor only depends on the number of feature vectors regardless of the object region size.

The performance of visual object tracking is directly related to the similarity (or dissimilarity) metric and model update steps of the covariance matrix. Forstner et al. [18] have proposed a dissimilarity criterion to compare the two SPD covariance matrices in the form of

$$dist(C_1, C_2) = \sqrt{\sum_{i=1}^{d} \log^2 \lambda_i (C_1, C_2)}, \tag{3}$$

where λ_i are the generalized eigenvalues of matrices C_1 and C_2 given by $C_1 X = C_2 X \lambda$ with X being the generalized eigenvector matrix. The reader is referred to [6,7] for detailed information and the advantages of RCD.

In this paper, a single covariance matrix is used to model appearance of the target object and the model update method by means of Riemannian geometry is utilized as presented in [7], where T previous covariance matrices have been taken into consideration in order to obtain a sample mean covariance matrix by gradient descent.

3 Proposed Method

High-performance tracking of an object is not always possible in RCD based tracking methods in case of complex situations. Performance degradation might occur during the template matching or model update of covariance matrices in tracking. Different candidates of RCD may have similar feature descriptors/similar distributions according to the reference RCD. For example, most of the RCD based approaches do not take first order statistics into account over the distribution during the RCD based template matching. In cases of several local minimums with close values, determining the best/correct candidate RCD can be difficult.

Occlusion is also a crucial issue in target tracking problems. RCD based methods may demonstrate relatively low-tracking performance when full or semi occlusion occurs in video sequences. In general, RCD based methods use multiple covariance matrices to represent target region. Using multiple covariance matrices is useful when the reference target is occluded by another object. However, there is no adaptive control mechanism to check if there is full occlusion and to handle model update in the case of full occlusion. To deal with these shortcomings, three contributions are proposed in this paper. The flowchart of

proposed approach is shown in Fig. 2.The contributions of this paper are highlighted in this Figure.

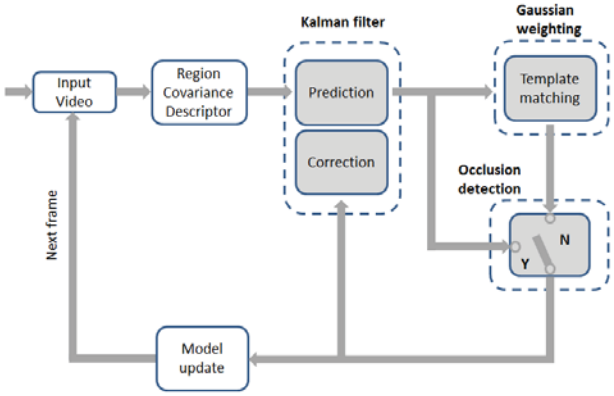


Fig. 2 Flowchart of proposed approach.

3.1 Weighting of distance measures

In RCD based target tracking, the correct candidate position should provide a good similarity measure, i.e. a low distance measure with respect to the reference template. However, due to scene context it is possible that an incorrect candidate position also provides a low distance measure, which can sometimes be even below the correct candidate value.

In this paper, it is proposed to introduce a weight to the distance measures to give more emphasis to the candidates that are regarded to be more likely considering motion characteristics obtained from the previous frames. An isotropic and monotonically increasing Gaussian formed function (GFF) as defined in (4) is used for the weighting process with a predetermined σ value in the form of

$$GFF(x,y) = B - G_{2D}(x,y,\sigma) \times (B - S), \qquad (4)$$

where $G_{2D}(x,y,\sigma)$ is a 2D-Gaussian function normalized to unity height. The minimum and maximum amplitude of the upside down Gaussian function are denoted as S and B, respectively, which have been determined experimentally to achieve good tracking performance throughout all sequences. The GFF used in proposed approach is shown in Fig. 3.

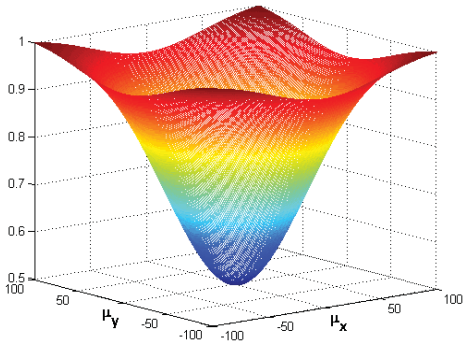


Fig. 3 Illustration of GFF.

The distance measures are weighted with the GFF so that more influence is given to the most likely target position. The weight of the similarity measure is adjusted so that the influence is reduced with distance to this point. Thanks to this approach, a candidate with similarity measure that has a slightly higher distance but is close to the most likely position will be preferred to another candidate with a slightly lower distance that is further away from the most likely position.

3.2 Kalman filter guidance

In order to determine the most likely position of the candidate target using previous motion characteristics, the KF is utilized in this paper, using a constant acceleration motion model. The best candidate template is thus searched around the center location determined by the KF using a fast search approach. Note that, just the prediction step of the KF has been taken into account to identify the center location (i.e. most likely position) of the target in the following form:

$$\hat{x}_k^- = A_k \hat{x}_{k-1} \qquad (5)$$

where $\hat{x}_k^-$ is a priori prediction at step k, A_k is the state-transition matrix, and $\hat{x}_{k-1}$ refers to a posteriori state estimation at step $k-1$. Note that, the prediction step is linear discrete time approximation from the continuous time system. The remaining parts of the KF in our approach have been used without any change.

In this paper, the predicted measurement of the KF is used as most likely target position, which is used for the weighting of RCD distance measures as explained in the previous sub-section.

3.3 Occlusion Detection

In order to increase the accuracy of the target tracking during occlusion, another contribution is also introduced by means of the KF. In this contribution, the priori state estimation is used instead of the actual best measurement of the target position when unexpected deviation from the distance measure is encountered. This restriction has been found to be very useful in case of occlusions. Hence, a threshold is adaptively calculated from the moving average value of the distance measures using a pre-determined number of M previous frames to detect unexpected deviation. The corresponding threshold at time t can be formulated as

$$TH_t = Coeff \times \left(1/M \sum_{i=1}^{M} dist_{t-i}(Cref_{t-i}, Ctarget_{t-i}) \right), \qquad (6)$$

where $Coeff$ is a preset coefficient.

4 Experimental Results

Several real-world video sequences with static and dynamic backgrounds have been evaluated in order to assess the performance of the proposed approach versus the reference COV [7], as well as recent state-of-the-art techniques ICTL [11], and MC [12]. Note that, the target within each sequence is manually initialized. Also the proposed framework considers only single target tracking. Visual tracking results corresponding to this work can be found at

http://kulis.kocaeli.edu.tr/RCD_comp_res.php.Video results can also be individually viewed from the given link.

The parameters for GFF are $\sigma = 50$, the minimum and maximum amplitudes are 0.5 and 1 respectively. A 7-dimensional feature vector, containing x and y pixel coordinates, RGB color and first derivative gradient at x and y direction, is used for each pixel of the target region. The best target template is searched around the most likely target candidate template location with different scales (large/small). The search range is restricted to ± 100 pixels and the number of search points is uniformly reduced by decimation, i.e. selecting every 4^{th} pixel location of the possible target region, to decrease computational load of the object detection stage. The number of previous covariance matrices (T) used to obtain the sample mean covariance matrix is set to 40, as in COV [7].

In order to assess tracking performance, importance of parameter selection has been investigated. For this purpose, at first, setting the diagonal elements of the priori error covariance matrix (PECM), measurement noise covariance matrix (MNCM) and number of previous frames M to a constant value, a sun-optimal *Coeff* parameter is determined within a plausible range according to the to the False Alarm Ratio (FAR). FAR shows the ratio of the number of false target detections to the number of total targets for the entire test sequences. Note that, an estimated target, that has a Euclidean distance more than 20 pixels apart from the ground truth, is marked as a false detection. In the next step, a sub-optimal M parameter is selected within a plausible range according to FAR using the *Coeff* parameter determined in the previous step and preserving PECM and MNCM values. Then, the PECM and MNCM values are selected within plausible ranges.

Fig. 4 shows the selection of the method parameters from *Coeff* to MNCM parameter respectively. It is clearly seen that the FAR values for each subfigure in Fig. 4 do not change very much through the neighborhood of the best parameter value. Through this process PECM, MNCM values, *Coeff* and M parameters are set to 12, 50, 2 and 30 respectively to obtain good tracking performance with lowest FAR. Finally, the *Coeff* parameter is recalculated using the previously estimated values and parameters to check robustness to parameter changes.

Fig. 5 shows the relation between the priori-estimated *Coeff* parameter and posteriori-estimated *Coeff* parameter. It is seen that the FAR results for each *Coeff* parameter given in Fig. 5 remain close throughout the pre-defined range. Note that, 75% of the total false target detections originate from "Woman" sequence. Therefore, the proposed method has about 80% tracking performance according to the FAR for the entire test sequences shown in Fig. 4 and Fig. 5.Note that tracking performance reaches to 95% when the "Woman" sequence is excluded. The ICTL approach has about 77% and 81% tracking performance for the entire test sequences when the "Woman" sequence is included and excluded, respectively. In addition, the MC approach has about 63% and 75% tracking performance for the same situation.

Table 1, Table 2 show average (mean), and standard deviation values of the tracking error results for several sequences recorded in stationary and moving camera, respectively. In Table 1, our method provides similar tracking performances compared to existing approaches in less challenging sequences such as the "Crowd" and "Subway" sequences. However, our method has generally better performance than the other approaches in terms of mean and standard values. As seen in Table 2, nearly all results of the proposed approach are better than the compared methods, with a clear success overall. Note that, the proposed method significantly outperforms the competition in complex sequences such as "Couple" and "Woman". Furthermore, for example MC and ICTL approaches lose the target due to full occlusion in the "Jogging1" and "Jogging2" sequences.

Table 1 Average (Mean) and Standard Deviation (SD) Values of Tracking Error (Pixel) in Sequences recorded in stationary camera.

Sequences \ Methods		COV [7]	ICTL [11]	MC [12]	**Proposed**
Crowd (440*360) 250 frames	Ave.	2.56	**1.88**	2.40	2.27
	SD	1.41	**1.07**	1.26	1.16
EnterExit1co (384*288) 195 frames	Ave.	11.05	13.31	11.61	**8.88**
	SD	9.18	20.71	19.07	**7.46**
Subway (352*288) 180 frames	Ave.	5.06	5.15	123.8	**3.99**
	SD	3.84	**3.59**	83.72	4.13
Renter1front (384*288) 239 frames	Ave.	**7.94**	9.78	11.05	8.08
	SD	6.90	8.65	10.29	**4.65**

Table 2 Average (Mean) and Standard Deviation (SD) Values of Tracking Error (Pixel) in Sequences recorded in moving camera.

Sequences \ Methods		COV [7]	ICTL [11]	MC [12]	**Proposed**
Couple (320*240) 140 frames	Ave.	20.26	14.35	16.46	**7.79**
	SD	39.95	16.85	19.82	**5.30**
Jogging1 (352*288) 306 frames	Ave.	8.58	8.15	90.10	**7.25**
	SD	12.40	5.52	51.82	**4.02**
Jogging2 (352*288) 306 frames	Ave.	6.76	124.5	6.02	**6.00**
	SD	12.36	66.31	6.39	7.19
Woman (352*288) 542 frames	Ave.	62.80	32.44	110.6	**26.42**
	SD	46.91	26.42	75.24	**13.80**

Fig. 6 shows quantitative evolution of the tracking performance using the proposed method and COV, ICTL, MC for all frames of the "EnterExit1cor" sequence. The tracking error is measured using the Euclidean distance between the center points of predicted targets and ground truth. It is seen that the proposed method has considerably lower tracking

errors, particularly for some frames where the other approaches perform poorly, thanks to the motion prediction and GFF introduced in this paper. Note that, ICTL and MC approaches show close tracking performance compared to the proposed method up to about the 165th frame, however, their performances decrease dramatically due to particle degeneracy and the lack of sufficient samples after this point.

Fig. 7 (a) shows visual tracking results of the proposed method and the other approaches for the "Couple" sequence which has abrupt motion and scale variations. Similar feature descriptors occurring on the background reduce the performance of the COV tracker. On the other hand, MC and ICTL trackers are unable to track the target efficiently because of the abrupt motion. However, this challenge does not affect the performance of the proposed method thanks to the introduced KF prediction and GFF weighting. Note that, multiple objects to be tracked are evaluated as a single object at the beginning of the sequence. Two different target region have been seen clearly and these regions are successfully detected by proposed method up to half of the sequence. After 60th frame, one of the target is occluded by another one. Note that the proposed approach is still able to track the targets accurately. However, tracking performance slightly decreases but still better than the compared approaches at the end of the sequence. This performance degradation is simply because of a split event occurred between two target regions and there is no information about the occluded target in the reference covariance matrix. Fig. 7 (b) shows visual tracking results for the "Woman" sequence, which has several long-term semi occlusion, and large-scale variations. It is clear that the reference appearance is affected over time by occlusion. However, the proposed method reduces background clutter and noise influence. Hence, the proposed method outperforms the other trackers thanks to the occlusion detection stage and KF prediction.

Fig. 8 shows performance evaluation of the proposed method using only GFF, only KF or both together. The proposed method combining GFF and KF has a better performance in almost all sequences, and clearly outperforms individual utilization cases in the overall. Moreover, the combined case is relatively more insensitive to occlusion and similar feature descriptors with respect to the target region. Using GFF and KF individually provides worse performance in cases of existing abrupt motion, large-scale variations and high number of similar feature descriptor regions.

The proposed algorithm takes about 0.74 ms per frame with a non-optimized code using MATLAB on a 3.4 GHz. PC. The ICTL and MC trackers take 0.30 and 0.36 ms per frame, respectively on the same platform. The number of search points related to RCD based tracking that is also valid for COV [7], has an effect on the relatively higher computational load. Note that, the proposed GFF and KF approaches that constitute the novel part of this paper have no significant effect on the complexity of the tracking process.

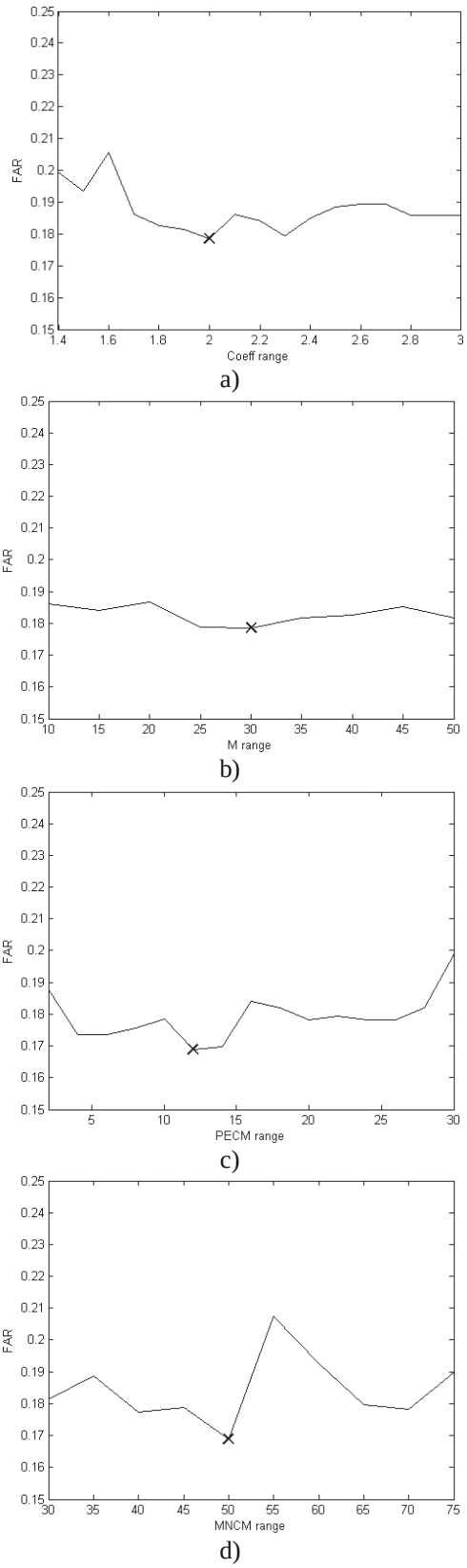


Fig. 4 Parameter optimization according to the FAR value a) sub-optimal Coeff selection b) sub-optimal M selection c) sub-optimal PECM selection d) sub-optimal MNCM selection

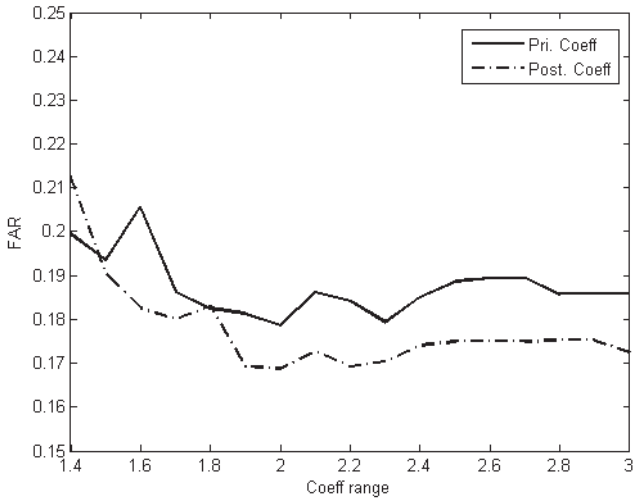


Fig. 5 The relation between priori estimated Coeff parameter and posteriori estimated Coeff parameter according to FAR value.

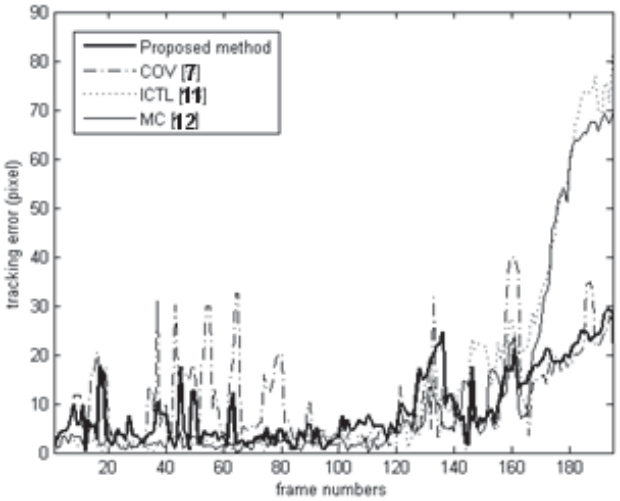


Fig. 6 Comparison of tracking error performance using proposed method and COV [7], ICTL [11], MC [12] approaches for EnterExit1cor sequence.

5 Conclusions

This paper proposes a novel visual tracking approach incorporating Gaussian formed function and Kalman filter with RCD. Robust target tracking is achieved using Kalman filter guided Gaussian similarity weighting. Additionally, an adaptive occlusion detection phase is integrated into the proposed approach to ensure more reliable tracking results. Detailed experimental results demonstrate robustness of the proposed method in relatively complex situations. The proposed approach is robust to abrupt motion as well as smooth motion in visual tracking. The proposed algorithm can easily be extended to RCD based multiple target tracking if desired.

6 References

[1] Comanicu, D., Ramesh, V., and Meer, P., "Real-time tracking of non-rigid objects using mean shift," IEEE Conf. on Computer Vision and Pattern Recognition, 142–149, 2000.

[2] Nummiaro, K., Koller, M. E., and Van Gool, L., "A color-based particle filter," In Proc.Of Workshop on Generative-Model Based Vision, 53-60, 2002.

[3] Sun, L., Liu,G., and Liu, Y., "Multiple pedestrians tracking algorithm by incorporating histogram of oriented gradient detections," IET Image Processing,7,7, 653-659 2013.

[4] Rami H., Hamri, M., Masmoudi L., "Object Tracking in Images Sequence using Local Binary Pattern," Int. Journal of Computer Applications, 63, 20, 19-23, 2013.

[5] Birchfield, S., "Elliptical head tracking using intensity gradients and colour histograms," In Proc. of IEEE Conf. on Comp. Vis. Pattern Recog., pp. 232-237, 1998.

[6] Tuzel, O., Porikli, F., and meer P., "Region Covariance: A fast descriptor for detection and classification," In Proc. 9th European Conf. on Computer Vision, pp. 18-25, 2005

[7] Porikli, F., Tuzel, O., and Meer, P., "Covariance Tracking using model update on lie algebra," In Proc. IEEE Conference Computer Vision and Pattern Recognition, 2006.

[8] Tuzel, O., Porikli, F., and Meer, P., "Pedestrian Detection via Classification on Riemannian Manifolds," IEEE Trans. on Pattern Analysis and Machine Intel, 30, 10, 1713-1727, 2008.

[9] Paisitkriangkrai, S., Shen, C., and Zhang, J., "Fast pedestrian detection using a cascade of boosted covariance features," IEEE Trans. Circuits Syst. Video Technol., 18, 8, 1140-1151, 2008.

[10] Pang, Y., Yuan, Y., and Li, X., "Gabor-based region covariance matrices for face recognition," IEEE Trans. Circuits Syst. Video Technol., 18, 7, 989-993, 2008.

[11] Wu, Y., Cheng, J., Wang, J., H. Lu, et al. "Real-Time Probabilistic Covariance Tracking with Efficient Model Update," IEEE Trans on Image Processing., 21, 5, 2824-2837, 2012.

[12] Ding , X., Huang, C., Huang, F., Xu, L., and Fang L. X., "Region covariance based object tracking using Monte Carlo method," 8th IEEE International Conf. on Control and Automation, pp. 1802-1805, 2010.

[13] Liu, Y., Li, G., and Shi, Z., "Covariance Tracking via Geometric Particle Filtering," EURASIP Journal on Advances in Signal Processing, 2010.

[14] Cortez, C. P., Undurraga, R. C., Mery, Q. D., and Alvaro, S., "Performance Evaluation of the Covariance Descriptor for Detection," International Conf. of the Chilean Computer Science Society, pp. 133-141, 2009.

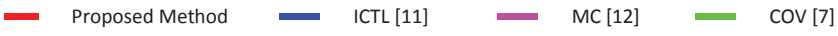
[15] Palaio, H., Batista, J., "A region covariance embedded in a particle filter for multi-objects tracking," 8th Int. Workshop on Visual Surveillance, France, 2008.

[16] Tabia, H., Laga, H., Picard, D., and Gosselin, P. H., "Covariance Descriptors for 3D Shape Matching and Retrieval," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4185-4192, 2014.

[17] Kalman, R. E., "A New Approach to Linear Filtering and Predictions Problems," Transactions of the ASME- Journal of Basic Engineering, 82, pp. 35-45, 1960.

[18] Forstner W., and Moonen, B., "A metric for covariance matrices," Technical report Dept. of Geodesy and Geoinformatics, 1999.

Fig. 7 Visual tracking results a) Couple b) Woman (Best view in color)



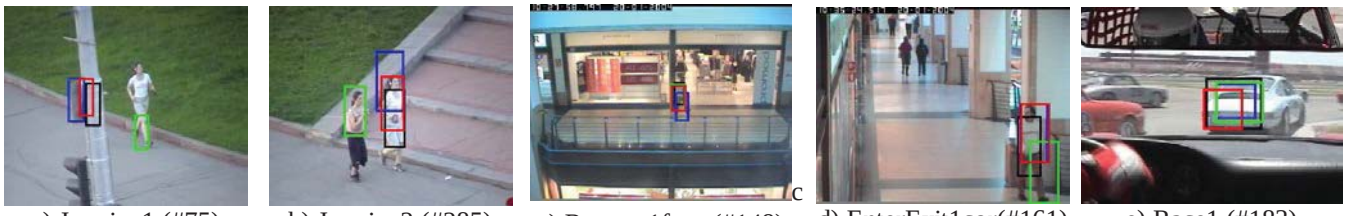


Fig. 8 Performance evolution of **t**he proposed method using only GFF, only KF and both together. (Best view in color)

Acknowledgments We would like to thank F. Porikli for valuable discussions and sharing corresponding video sequence databases. O. Akbulut would also like to thank O. Urhan for constructive discussions.

Toward Face Detection in 3D Data

Juan Paduano, Marcelo Romero and Vianney Muñoz
Department of Engineering
Autonomous University of the State of Mexico
{jpaduanos, mromeroh, vmunozj}@uaemex.mx

Abstract

In this paper, we present an experimental analysis on the face detection problem using 3D face data. Face detection is the first step in almost every face processing application, where the face is localized and extracted from an income image prior to specific analysis. Although its significant importance, an extended investigation on face detection in 3D data is still missed in the literature. Such absence might be associated to its well known difficulty, related to technical and natural complications, e.g. capturing sensors and conditions, processing algorithms and techniques. On the other hand, natural complications are present, because the human face is a complex no-rigid surface with a sophisticated anatomical structure and a fascinating behavior. Our main investigation is focused on the face detection problem in 3D data. In addressing that problem, we have identified three key published papers that use curvature analysis, a slicing approach and a segmentation technique respectively. Then, we have defined an experimental procedure to investigate those papers using the Face Recognition Grand Challenge database for a performance comparison and analysis. Results found so far are encouraging our main purpose, and they have illuminated different venues to attack the general problem when processing 3D images with more than one person on the scene.

Keywords— 3D face detection, 3D face processing, 3D feature descriptors.

1. Introduction

Face detection is one of the visual tasks that humans can do without effort. However, in terms of computational vision, this task is not easy. The growing need for automatic face recognition systems has encouraged the development of appropriate face detection algorithms. While 2D detection algorithms have reached an acceptable level, most of them assume impractical conditions, e.g. face recognition [1], face tracking [2], pose estimation [3], facial expressions and facial gestures recognition [4]. On the other hand, although 3D detection systems have been studied for decades with good progress, results are still limited to daily applications [5]. In this respect, any contribution that researchers can make in 3D face detection is very welcome.

The goal of face detection could be defined as: given an arbitrary image, you want to know if there are faces in the image, regardless of position, scale, guidelines and poses, if any, must return the location coordinates to these faces. Unfortunately, many factors delay or limit facial analysis on a surface image [5-8], despite face detection is a primary step for most face processing applications (e.g. monitoring and automated security systems, access control, personal identification and verification, facial reconstruction, design of man-machine interfaces, multimedia communication, medical diagnostics, surgery planning).

Chellappa et al. [18] and Bowyer et al. [5] have discussed the face detection problem, analyzing related topics as segmentation and feature extraction. Hence, types of automatic 3D face detection algorithms are identified: feature-based detection and holistic face detection.

Although face detection is essential for many face processing applications, one could see in the literature that it has received little attention, especially when using 3D data. Then, our overall research is aim to provide state of the art algorithms for face detection in 3D data. Following that objective, in this paper we are experimentally analyzing three key face detection papers from the literature which core is: curvature analysis [10], slicing [11] and segmentation [12]. We have replicated their experimental framework but using a common data corpus for comparison: the Face Recognition Grand Challenge (FRGC) database [9].

The rest of this paper is as follows. Section 2 describes three key face detection methods. Section 3 details our implementation procedure. Section 4 discusses our experimental results. Finally, Section 5 concludes this paper.

2. Face detection techniques

In this section we describe three key face detection techniques for experimental comparison, namely:

curvature analysis [10], slicing approach [11] and segmentation technique [12].

A. Curvature analysis

Colombo's technique [10] stars with a range image (see Figure 1-b), i.e. an image where for each location (i,j) the coordinates (x,y,z) of the 3D scene are expressed with respect to the camera reference system. Some acquisition devices return data in the form of a polygonal model, usually a triangular mesh. But, in his case, the range image is obtained using the z-buffer algorithm [15].

Once the range image is computed, surface curvature, which has the valuable characteristic of being viewpoint invariant, is exploited to segment candidate eyes and noses. In detail: (a) the mean *(H)* and Gaussian *(K)* curvature maps are first computed from a smoothed version of the original range image; (b) simple thresholding segments regions of high curvature which might correspond to eyes and noses are obtained; (c) a HK classification, based on the signs of Gaussian and Mean curvature, divides the segmented regions into four types: convex, concave, and two types of saddle regions.

Regions that may contain a nose and eyes are then characterized by their type and by some statistics of their curvature.

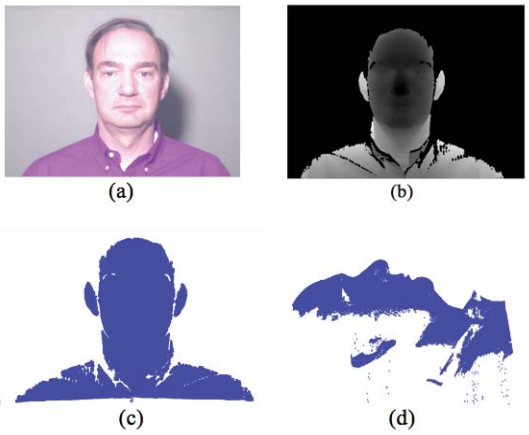


Figure 1. Example of 2D/3D images. (a) 2D intensity image. (b) Range image (c) 3D image front view (d) Profile of the 3D image, where peaks and holes are observed.

The output of the processing step may contain any number of candidate facial features. If no nose or less than two candidate eyes are detected, they assume that no faces are present in the acquired scene, while there are no upper bounds on the number of features that can be detected and further processed. The final output of the procedure is a list containing the location and extension of each detected face.

Analyzing the curvature of 3D face images, let S be the surface defined by a twice differentiable real valued function:

$$f : U \to R, \text{ defined on an open set } U \subseteq R^2: \quad (1)$$

$$S = \{(x,y,z) \mid (x,y) \in U; z \in R ; f(x,y) = z\} \quad (2)$$

For every point $\big(x,y,f(x,y)\big) \in S$ they consider two curvature measures, the mean *(H)* and the Gaussian *(K)* curvature [16].

$$H(x,y) = \frac{(1+f_y^2)f_{xx} - 2f_x f_y f_{xy} + (1+f_x^2)f_{yy}}{2(1+f_x^2+f_y^2)^{\frac{3}{2}}} \quad (3)$$

$$K(x,y) = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1+f_x^2+f_y^2)^2} \quad (4)$$

where $f_x, f_y, f_{xy}, f_{xx}, f_{yy}$ are the first and second derivatives of f in (x,y). A face is initially represented by a range image of $M \times N$ points. Since we have only a discrete representation of S, estimate the partial derivatives. For each point $\big(x_i, y_j\big)$ on the grid, they considered a biquadratic polynomial approximation of the surface:

$$\begin{aligned} g_{ij}(x,y) = {} & a_{ij} + b_{ij}(x - x_i) + c_{ij}\big(y - y_j\big) \\ & + d_{ij}(x - x_i)\big(y - y_j\big) \\ & + e_{ij}(x - x_i)^2 + f_{ij}(y - y_j)^2, \\ & i = 1 \ldots N, j = 1 \ldots M \end{aligned} \quad (5)$$

Where coefficients $a_{ij}, b_{ij}, c_{ij}, d_{ij}, e_{ij}, f_{ij}$ are obtained by least squares fitting of the points in a neighborhood of $\big(x_i, y_j\big)$. The derivatives of f in $\big(x_i, y_j\big)$ are then estimated by the derivatives of g_{ij}:

$$\begin{aligned} f_x\big(x_i, y_j\big) = b_{ij}, \quad & f_y\big(x_i, y_j\big) = c_{ij}, \\ f_{xy}\big(x_i, y_j\big) = d_{ij}, \quad & f_{xx}\big(x_i, y_j\big) = 2e_{ij}, \\ & f_{yy}\big(x_i, y_j\big) = 2f_{ij}. \end{aligned} \quad (6)$$

Since the second derivative is very sensitive to noise, a smoothing filter must be applied to the surface. Before computing the curvature, they apply a Gaussian filter to the depth image, discarding high-frequency fluctuations of the surface, while the salient facial features, such as the eyes and nose, are still clearly distinguishable. By analyzing the signs of the mean and the Gaussian curvature, they perform what is called an HK classification of the points on the surface to obtain a concise description of the local behavior of the surface (see Figure 2).

HK classification was introduced by Besl et al. [17]. Image points can be labelled as belonging to a viewpoint-independent surface shape class type based on the combination of the signs from the Gaussian and mean curvatures as shown in Table 1.

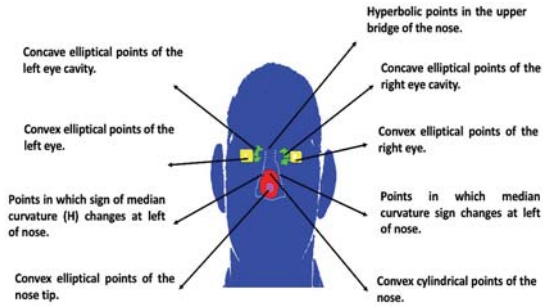


Figure 2. Segmented regions of a face from which facial features are extracted.

Table 1. HK classification.

	$K < 0$	$K = 0$	$K > 0$
$H < 0$	Hyperbolic concave	Cylindrical concave	Elliptical concave
$H = 0$	Hyperbolic symmetric	Planar	Impossible
$H > 0$	Hyperbolic convex	Cylindrical convex	Elliptical convex

They use a thresholding process to isolate regions of high curvature. Points with low curvature values are discarded:

$$|H(u,v)| \geq T_h \qquad |K(u,v)| \geq T_k \tag{7}$$

where T_h and T_k are predefined thresholds.

Consequently, they reduce the number of candidates (see Figure 3) by filtering each candidate region i, considering average Mean and Gaussian curvature ($\overline{Hi}$ and $\overline{Ki}$, respectively): for noses and for eyes.

$$\overline{Hi} \geq \overline{H}_{min} \tag{8}$$
$$\overline{Ki} \geq \overline{K}_{min} \tag{9}$$

B. Slicing analysis

The second method is present for Mian et al. [11]. This method detect the nose tip in the first step in order to crop out the required facial area from the 3D face. The nose tip is detected using a coarse to fine approach as follows. Each 3D face is horizontally sliced at multiple steps *dv*. Initially a large value is selected for *dv* to improve speed and once the nose is coarsely located the search is repeated in the neighboring region with a smaller value of *dv*. The data points of each slice are interpolated at uniform intervals to fill in any holes. Next, circles centered at multiple horizontal intervals *dh* on the slice are used to select a segment from the slice and a triangle is inscribed using the center of the circle and the points of intersection of the slice with the circle as shown in Figure 4-a. Once again a coarse to fine approach is used for selecting the value of *dh* for performance reasons. The point which has the maximum altitude triangle associated with it is considered to be a potential nose tip on the slice and is assigned a confidence value equal to the altitude.

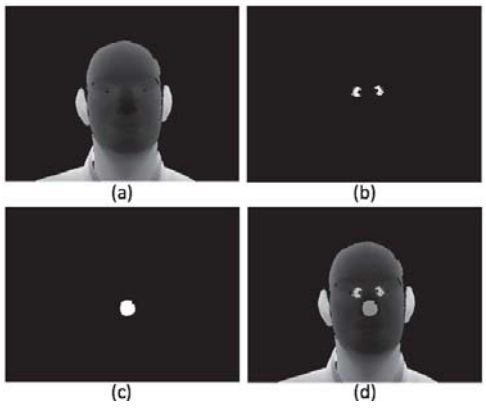


Figure 3. Curvature analysis; (a) Rage image, (b) Binary image used to localize eye candidates, (c) Binary image used to localize nose candidates, (d) selected regions from the rage image.

This process is repeated for all slices resulting in one candidate point per slice along with its confidence value. These candidate points form the nose ridge. Points that do not correspond to the nose ridge are outliers and are removed using RANSAC [16]. Out of the remaining points, the one which has the maximum confidence is taken as the nose tip and the above process is repeated at smaller values of dv and dh in the neighboring region of the nose tip for a more accurate localization.

A sphere of radius *r* centered at the nose tip (see Figure 4-b) is then used to crop the 3D face and its corresponding registered 2D face. A constant value of *r* = 80 *mm* was selected in this experiment.

This process crops an elliptical region (when viewed in the xy plane) from the face with vertical major axis and horizontal minor axis. The ellipse varies with the curvature of the face. For example, the more narrow a face is, the greater is the major axis to minor axis ratio.

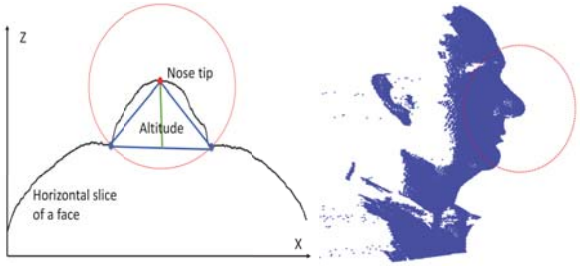
Once the face is cropped, outlier points causing spikes in the 3D face are removed. They defined outlier points as the ones whose distance is greater than a threshold *dt* from any one of its 8-connected neighbors. The *dt* is automatically calculated as:

$$dt = \mu + 0.6\,\sigma \tag{10}$$

where μ is the mean distance between eight neighboring points and σ is its standard deviation.

After removing spikes the 3D face and its corresponding 2D face are resampled on a uniform square grid at 1 mm resolution. Removal of spikes may

result in holes in the 3D face which are filled using cubic interpolation. Resampling the 2D face on a similar grid as the 3D face, ensures a one-to-one correspondence between the two images. Since noise in 3D data generally occurs along the viewing direction (z-axis) of the sensor, the z-component of the 3D face is denoised using a median filter.



(a) **(b)**

Figure 4. An image-slice to detect the nose tip. Profile view when centering a sphere at the nose tip level.

C. Face segmentation

Segundo et al. [12] present a method to extract the face region from an input range image containing only one subject. Their face segmentation algorithm is composed basically by two main stages: (a) locating homogeneous regions in the input image by using clustering combined with edge data; and (b) identifying candidate regions that belong to the face region by an ellipse detection method based on the Hough Transform [14]. They apply the *K-Means* algorithm [13] by setting $k = 3$ (i.e., the number of clusters) to segment the image in three main regions: background, body, and face (see Figure 5). However, this step alone is not enough to correctly extract the face region because some characteristics may interfere, i.e., hairstyle, ears, neck. Then, the goal of the edge detection process is to help in eliminating these parts from the resulting segmented face region after applying *K-Means*.

To perform edge detection they apply the *Sobel* operator on the depth information from the range image and define a threshold. The gradient threshold is based only on spatial information and is easily obtained by analyzing the histogram of the gradient image. They computed the histograms for all gradient images and observed that they have similar behavior. Thus, we defined an automatic threshold based on the inflection point of the histogram curve for each gradient image to detect the face boundaries. In general, the computed threshold has a value between 8 and 12 for the tested databases. Finally, to obtain the final edge map they applied a closing process (dilation followed by erosion of neighbor pixels) to link some broken lines generated by the threshold operation. After performing region and edge detection, which can be made in parallel, they combine the two resulting images by using an AND operation. The combined image is processed to obtain the region of interest (ROI), i.e. the face region, without irrelevant, disturbing parts. Figure 5 shows the complete process.

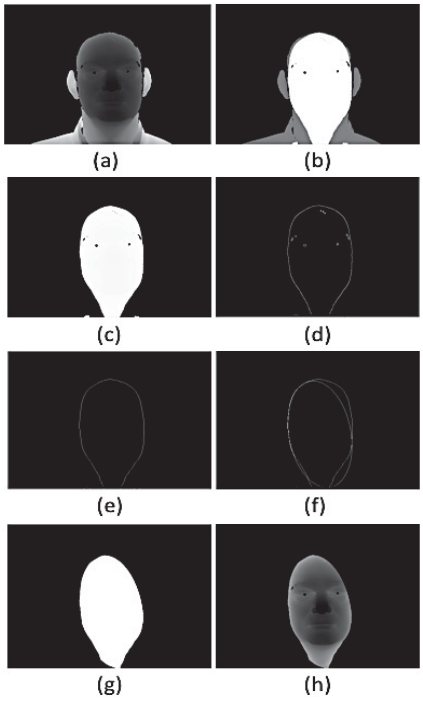


Figure 5. Examples of (a) range image, (b) K-Means algorithm result, (c) face region extracting, (d) Sobel operator result, (e) closing process result (f) ellipse detection with Hough Transform (g) binary image after normalization (h) final segmentation.

In this stage, they used the new edge map to precisely locate the face region in the input image. To perform shape detection they used the Hough Transform [14], looking for an ellipse, which is considered to be the geometric shape most similar to the face boundary. After the ellipse detection, they select the labeled regions that belong to the face region, by selecting regions inside of the detected ellipse (see Figure 5-f). The region selected as the ROI is represented as a binary image that indicates the location of the face region in the input image. However, this binary image can have some holes which are fulfilled by a closing process.

After that, they perform a logical AND between the resulting binary image and the input range image, obtaining the final segmentation of the face region.

3. Experimental analysis

This section describes our implementations of the three face detection techniques described in Section 2.

For this investigation, we use a state of the art database from the Face Recognition Grand Challenge (FRGC) program [9], which is the most widely used data set available for the research community. In total, this database contains 4950 2D/3D face images from different subjects acquired on different conditions. For this publications we are reporting performance over the complete set.

Now we describe every face detection technique implemented and compared for this publication. For each case, we follow the author's pipeline and in case that no detail is given we have used an alternate implementation or threshold value.

A. Curvature analysis

Particularities on our replication of Colombo's curvature analysis technique [10] are as follows:

1. Depth images are computed from respective 3D point cloud using equations (1) and (2).
2. We apply a Gaussian filter (10) to the depth image, discarding high-frequency fluctuations of the surface, while the salient facial features, such as the eyes and nose, are still clearly distinguishable.

$$\left(\frac{1}{16}\right) * \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \quad (11)$$

3. For every point on the depth map, Mean (3) and Gaussian (4) curvature are obtained by calculating the first and second derivatives as [16]:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + O(h^2) \quad (12)$$

$$f''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} + O(h^2) \quad (13)$$

where x_0 is (x, y) in the range image, h is an integer greater than zero, and $O(h^2)$ is the *truncation error*, caused by stopping the polynomial approximation to second order, which tends to zero.

4. Then, by analyzing the signs of the mean and the Gaussian curvature, we perform what is called an HK classification (see Table 1).
5. We use a thresholding process (7) to isolate regions of high curvature as shown in Figure 6, while areas with low curvature values are discarded.
6. *Our* T_h and T_k values are estimated by:

$$T_h = (h_{max} - h_{min})/4 \quad (14)$$

$$T_k = (k_{max} - k_{min})/4 \quad (15)$$

8. Finally, we reduce the number of candidates as can be saw in Figure 7, by filtering each candidate region *i*, considering average Mean (8) and Gaussian curvature (9).

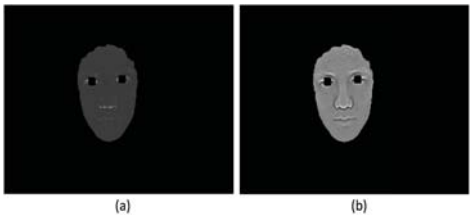


Figure 6. A thresholding process, analyzing (a) Gaussian curvature values and (b) Mean curvature values.

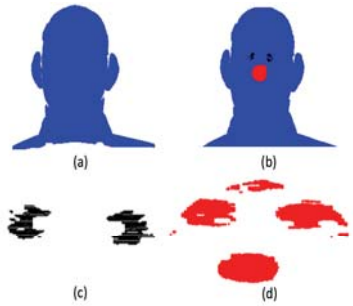


Figure 7. (a) 3D data, (b) final candidate points region, (c) candidate points of the eyes cavity, (d) candidate points of the nose tip.

B. Slicing analysis

Particularities on our replication of Mian's slicing technique [11] are as follows:

1. Each 3D image is horizontally sliced at multiple steps *dv*. Next, removing spikes of each sliced using the *dt.* The *dt* is automatically calculated using equation (10).
2. Removal of spikes may result in holes in the 3D image which are filled using cubic interpolation. Then, circles centered at multiple horizontal intervals *dh* on the slice are used to select a segment from the slice and a triangle is inscribed using the center of the circle and the points of intersection of the slice with the circle as shown in Figure 4.
3. The point which has the maximum altitude triangle associated with it is considered to be a

potential nose tip on the slice and is assigned a confidence value equal to the altitude.

4. Step 3 is repeated for all slices, resulting in one candidate point per slice. These candidate points form the nose ridge. Points that do not correspond to the nose ridge are outliers and are removed using RANSAC [16].

C. Face segmentation

Particularities on our implementation of Segundo's technique [12] are as follows:

1. Depth images from 3D point cloud are computed using the equations (1) and (2).
2. Then, we literally follow the sequel shown in Figure 5.
3. *K-Means* algorithm [13] is applied by setting k = 3 to segment the image in three main regions: background, body, and face (see Figure 5-b).
4. For this publication, we have made a Matlab implementation for the *K-Means* algorithm, which in detail is doing as follows:

   Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

   a. Randomly select *c* cluster centers.
   b. Calculate the distance between each data point and cluster centers.
   c. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
   d. Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{i=1}^{c_i} x_i \qquad (16)$$

   where, *ci* represents the number of data points in *i* cluster.

   e. Recalculate the distance between each data point and obtain new cluster centers.
   f. If no data point was reassigned, stop, otherwise go to step *c*.

5. *Sobel* operator is applied on the face region of the range image.
6. An elliptical are is detected using Hough Transform [14].
7. To obtain the final edge map we applied a closing process and combine the two resulting images by using an AND operation.

4. Results

In this section we report our experimental results. As we have mentioned before, three key papers on face detection have been investigated using the Face Recognition Grand Challenge database.

Our results include a performance comparison (Table 2) and a processing time summary (Table 3).

First, we implement every face detection approach by developing specific purpose functions or using standard available libraries in Matlab version 14. Those experiments were executed on a Mac BookPro laptop core i7 2.3GHz, 16GB in RAM.

For the three face detection algorithms, their performance is calculated by counting those facial areas that contained the fourteen facial landmarks in the reference ground truth [19].

After experimentation, a decrease in performance for each of the three techniques is observed, in comparison to those results reported by their authors. However, as summarized in Table 2, Segundo et al. [12] still scores the best performance among the three experimented approaches with an 87.90% successful face detection.

Table 2. Performance summary.

Technique	*Overall performance*	*Comments*
Curvature analysis [10].	85.35%	Performance corresponds to those cases where the localized facial area contains the inner-eye-corners and the tip of the nose.
Slicing analysis [11].	81.45%	Performance corresponds to those cases where the localized facial area contains the tip of the nose.
Face segmentation [12].	87.90%	Performance corresponds to those cases where the localized facial area contains 14 landmarks [19].

Table 3. Processing time summary.

Technique	*Average processing time [s]*
Curvature analysis [10].	2.94
Slicing analysis [11].	30.11
Face segmentation [12].	4.78

Discussing now the processing time. As listed in Table 3, one can observe that Colombo's technique [10] is faster than Segundo's [12] and Mian's techniques [11]. In addition to that, we believe that Colombo's technique is also more robust to pose variations, as it is based on curvature analysis. Contrarily, Segundo's technique would suffer when computed with extreme pose variation 3D images. However, further research is needed to validate our assumption.

5. Conclusion

In this paper we have presented an experimental analysis on the face detection problem using 3D data.

Three key approaches in related literature have been investigated. The first approach uses a curvature analysis to detect the most rigid facial area where the inner-eyes corners and the tip of the nose are expected. The second approach horizontally slices an income image to localize the most likely facial area that contains the tip of the nose. The third approach uses a segmentation procedure to detect an elliptical area where a human face is expected.

We have used the Face Recognition Grand Challenge database to replicate those identified face detection approaches for our experimental analysis and comparison.

In our performance evaluation we have used state of the art ground-truth of fourteen facial landmarks [19]. Then, we count every facial landmark contained within every detected facial region to decide whether or not a human face has been detected.

We notice that our experimental results are lower than those reported by their authors. However, among the three investigated papers, Segundo et al. [12] still scores the best face detection performance (87.90%).

In addition to their performance, two important notes are observed. The first note is that those three approaches are pose-dependent. Most importantly, the second note is that their application is limited for 3D images that contain more than one person. The good thing is that some ideas from those papers are illuminating in addressing our main objective, but further research is needed.

Acknowledgements

Authors thank the Research Department of the Autonomous University of the State of Mexico and the National Council for Science and Technology (CONACYT) for their financial support, grants 3720/2014/CID and 634473, respectively.

References

[1] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 2, pp. 372–386, 2012.

[2] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1442–1468, 2014.

[3] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation," *arXiv preprint arXiv*:1502.05908, 2015.

[4] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Metaanalysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics, Part B: Cybernetics*, IEEE Transactions on, vol. 42, no. 4, pp. 966–979, 2012.

[5] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1 – 15, 2006.

[6] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *Pattern Analysis and Machine Intelligence,* IEEE Transactions on, vol. 24, no. 1, pp. 34–58, Jan 2002.

[7] P. B. Nick Pears, Yonghuai Liu, Ed., 3D *Imaging, Analysis and Applications*. Springer London, 2012.

[8] M. Romero, J. Paduano, and V. Munoz, "Point-triplet spin-images for landmark localisation in 3d face data," *in Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings*, 2014 IEEE Workshop on, Oct 2014, pp. 8–14.

[9] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang,K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition*, 2005. CVPR 2005. IEEE computer society conference on, vol. 1. IEEE, 2005, pp. 947–954.

[10] A. Colombo, C. Cusano, and R. Schettini, "3d face detection using curvature analysis," Pattern Recognition, vol. 39, no. 3, pp. 444 – 455, 2006.

[11] A. Mian, M. Bennamoun, and R. Owens, "Automatic 3d face detection, normalization and recognition," in 3D Data Processing, Visualization, and Transmission, Third International Symposium on, June 2006, pp. 735–742.

[12] M. Segundo, C. Queirolo, O. Bellon, and L. Silva, "Automatic 3d facial segmentation and landmark detection," in Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on, Sept 2007, pp. 431–436.

[13] K. Krishna and M. N. Murty, "Genetic k-means algorithm," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 29, no. 3, pp. 433–439, 1999.

[14] R. A. McLaughlin, "Randomized hough transform: better ellipse detection," in IEEE TENCON-Digital Signal Processing Applications, vol. 1, 1996, pp. 409–414.

[15] A. Watt and M. Watt, "*Advanced animation and rendering techniques theory and practice*, 1994."

[16] E. Trucco and A. Verri, *Introductory techniques for 3-D computer vision.* Prentice Hall Englewood Cliffs, 1998, vol. 201.

[17] P. J. Besl and R. C. Jain, "Invariant surface characteristics for 3d object recognition in range images," *Computer vision, graphics, and image processing,* vol. 33, no. 1, pp. 33–80, 1986.

[18] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, 1995.

[19] C. Creusot, N. Pears, and J. Austin, "3d face landmark labelling," in *Proceedings of the ACM workshop on 3D object retrieval*, 2010.