Relative Depth Estimation using a Rotating Camera System

Pallav Garg, Suresh Yerva and Krishnan Kutty
Centre for Research in Engineering Sciences and Technology (CREST)
KPIT Cummins Infosystems Ltd.
Pune, India

Abstract— In the proposed method, the relative depth of the objects present in a 3D scene is calculated, using 2D images captured from a camera placed on a rotating platform. Unlike the conventional stereo imaging system, the proposed method captures multiple views of a 3D scene, each view taken at a different camera positions while on the rotating platform. The approach adapted calculates the disparity between the corresponding pixels present in both the views to get the relative depth of the objects. The relative depth of the objects can be calculated for the objects which are in common FOV (Field of View) of both the views captured. By virtue of the rotating platform, the proposed system is capable of creating a depth map in full 360 degree field of view unlike its traditional counterpart. The approach presented is a thereby also cost effective way for depth estimation since only one camera is being used.

Keywords—stereo; image rectification; stereo matching; disparity; depth map

I. INTRODUCTION

Human beings tend to perceive depth by virtue of a powerful combination of a pair of eyes and a powerful neural network that estimates the depth based on the images captured on the retina by both the eyes. The same concept of stereo vision is employed in various robotic applications today- thanks to the active research happening in the field of Computer Vision. The research in computer vision started with simple analysis and interpretation of images and image data. Later, the development in the field of computer vision increased rapidly. The various developments in the field of computer vision include obstacle avoidance with stereo vision [1], automatic navigation, object recognition, pedestrian detection, medical imaging etc. Computer vision plays a vital role in automotive industry for driver assistance systems, which reduces the driver efforts. One of the important research areas in computer vision is stereo vision.

The existing stereo vision system comprises of two cameras placed at certain known distance to capture the two different views of a scene. The stereo vision system estimates the 3D point of the object in real world based on 2 images.

The following are some of the techniques used for calculating the depth of an object. The LASER (Light Amplification by Stimulated Emission of Radiation) Triangulation [2] projects LASER on an object and acquires the height profile using a camera. In [3], a known light pattern is projected on to an object. The depth information is calculated according to the distortion of the light pattern. Time of Flight (TOF) based Depth Sensor [4] synchronizes light source with image sensor in order to calculate distance based on the time between the pulse of light and the reflected light back onto the sensor. In the field of medical imaging, Optical Coherence Tomography (OCT) [5], uses infrared light to calculate depth information by measuring the reflections of light through the cross-section of the object. These are the some of the different technologies [6] to find the distance of an object from the source. Over the last decade, these techniques either got replaced or boosted up based on the performance of existing techniques. There are also some shortcomings of these methods. In the LASER triangulation method, the LASER sensor should be kept clean; else it may affect the accuracy of the system [7]. TOF based depth camera has low resolution which sometimes gives non homogeneous depth map, intensity based distance error and light interference effects [8]. These technologies are costly and take significant time to acquire data of an object. In order to optimize the cost and make the system work in various conditions, a rotating camera system is proposed.

Applications of scene reconstruction can be found not only in earth sciences but also in entertainment industry and in cultural heritage digital archival. The proposed system has common application in robotic vision, military (for spying), and to develop an autonomous vehicle. This system also finds applications where it can be used as a shape identifier, such as finding the shapes of bottle or coffee cup. It can also be used to enhance the accuracy of identification systems like facial recognition or other biometrics [6].

The rest of the paper is organized as follows: Section II describes our methodology; Section III shows the results and discussions; Section IV emphasis on concluding remarks thereafter appropriate references are provided.
II. METHODOLOGY

There are two approaches to obtain depth map of the scene. The first is a conventional approach, based on the camera calibration. This results in deriving the intrinsic and extrinsic parameters. This is followed by a process called Image Rectification (using intrinsic parameters), which aligns the row of both the views and further performing stereo matching to obtain depth information.

A camera transforms a point in the real world onto a point in 2D coordinate system. This transformation can be sub-divided into two transformations: Extrinsic Transformation projects real world coordinate into camera coordinate and Intrinsic Transformation projects camera coordinate into image coordinate system. Extrinsic Transformation is depicted as given below:

\[
\begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix} = R \begin{bmatrix}
  x_w \\
  y_w \\
  z_w
\end{bmatrix} + T
\]

\[(x_w, y_w, z_w) = \text{Object world coordinate system} \]

\[(x, y, z) = \text{Camera 3D coordinate system} \]

The parameters to be calibrated are \( R \) and \( T \), where \( R \) is the \((3 \times 3)\) rotation matrix.

\[ R = \begin{bmatrix}
  r_1 & r_2 & r_3 \\
  r_4 & r_5 & r_6 \\
  r_7 & r_8 & r_9
\end{bmatrix} \]  

(2)

And \( T \) is the translation vector

\[ T = \begin{bmatrix}
  T_x \\
  T_y \\
  T_z
\end{bmatrix} \]  

(3)

Intrinsic Transformation is depicted as given below:

\[
\begin{bmatrix}
  x' \\
  y' \\
  1
\end{bmatrix} = s \begin{bmatrix}
  f_x & \gamma & o_x \\
  0 & f_y & o_y \\
  0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
  x \\
  y \\
  1
\end{bmatrix}
\]

Where, \( s = \) scaling factor

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} = \text{2D-image coordinates}
\]

\( f_x, f_y = \) Horizontal and Vertical focal lengths of camera

\( o_x, o_y = \) Center of image

\( \gamma = \) Skew coefficient

[9, 10], describes in depth, the details of camera calibration with lens distortion. The second approach computes disparity without performing camera calibration i.e., without using exact intrinsic parameters for image rectification. This is an uncalibrated rectification approach, which is then followed by stereo matching. The second approach is presented in this paper.

Fig. 1 depicts the process flow in estimating the depth information.

![Fig. 1. Flow of Rotating Camera System](image)

We have used the second approach since our set-up involves the use of a moving camera which in turn leads to the high calibration errors. Camera calibration is conventionally performed using images of a checker-board. However, due to the camera being placed on the rotating platform, many corners points of checker-board image are not captured simultaneously, in both views (at different camera position). This adds to the complexity for performing camera calibration. When the camera calibration was performed using image from the rotating platform, the ‘pixel error’ was 0.3167. This value is high when compared to the conventional stereo vision, when performed on normal stereo image pairs. [10] is used for the approximation of the pixel errors and distortion coefficient. Fig. 2 shows the error in detecting corners in various checker board images.

![Fig. 2. Re-Projection Error Analysis](image)
A. Setup for Capturing Images at Different Camera Position

In Fig. 3, the setup for capturing images using the proposed rotating platform is depicted. Fig. 3 also illustrates the direction of rotating camera; the light blue color shows the FOV of camera at Position ‘1’ and the FOV of camera at position ‘2’ after camera rotated by θ°. The region shown in dark blue shows the common FOV for both the views.

For the proposed method, we have taken the images at θ = 5°. We have to keep the angle of rotation as small as possible in order to contain the problem of quick disappearance of the objects in successive views. For larger angles say θ > 10°, there is a probability of missing the corresponding feature points in both the views.

B. Image Rectification

Image Rectification is defined as the alignment of epipolar lines of one image so that they become parallel with their corresponding epipolar lines in another image [11]. In other words, image rectification is a process of adjusting angles and distances between the views of two images. Image Rectification results in row aligned and rectified images. Fig. 4 and Fig. 5 depict the epipolar lines before and after alignment respectively.

There are not many algorithms that perform image rectification without camera calibration. In this paper, we have adopted Fusiello and Luca Irsara’s [12] approach, which performs image rectification without the need for calibrating the camera. This algorithm solves the problem of calculating camera projection matrices. This algorithm is best suited for rectifying images with two views. This method rotates the pair of projective images in order to make a pair of rectified images based on epipolar constraints. Other related algorithms that could also be used are Hartley’s approach [13] which makes use of fundamental matrix to calculate rigid transformation; Isgro and Trucco’s approach [14], which rectify images from feature points directly without calculating the fundamental matrix. There are few assumptions in our proposed approach. viz., intrinsic parameters are not known and corresponding points in both the views are available.

\[ c_1^i \leftrightarrow c_2^j \]  \hspace{2cm} (5)

Where, \( j \) is \( j^{th} \) correspondence of both the views of camera at position ‘1’ and position ‘2’.

The fundamental matrix of the rectified image pair is considered to be a cross product of a vector, \( u_1 = (1,0,0) \) and a skew symmetric matrix which has \( a_{nk} = -a_{nk} \).
Here, ‘i’ denotes the row number in a matrix and ‘k’ denotes the column number in a matrix. Therefore, fundamental matrix \([15]\) of rectified image pair would be

\[
F = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & -1 \\
0 & 1 & 0
\end{bmatrix}
\] (6)

The geometric re-projection error is calculated by Simpson’s Error [16] which is a first order approximation of geometric error. The advantage of calculating Simpson’s error is to have an idea of rectification results on different pair of images. The Sampson error for \(j^{th}\) correspondence is

\[
E_s^j = \frac{\left(m^T_i F m_i^j\right)^2}{\left\|u_3 x^T F m_i^j\right\|^2 + \left(m^T_i F [u_3 x]\right)^2}
\] (7)

Where, \(u_3 = (0,0,1)\)

We take \(E_s^j = 0\) i.e., for all corresponding points the error is reduced to zero. The rectifying transformations of both the views considered are,

\[
H_1 = K_{n1} R_1 K_{o1}^{-1}
\] (8)

\[
H_2 = K_{n2} R_2 K_{o2}^{-1}
\] (9)

Where, \(K_{o2} = \text{Old Intrinsic Parameters (} z = 1 \text{ and } 2\) and \(R_1, R_2 = \text{Rotation matrices, are unknown. Therefore, } K_{n1}, K_{n2}\) are kept randomly with a condition that vertical focal length and vertical principal point are same. And to get the old intrinsic parameters, calculation is made easier by taking skew coefficient as zero, principal points at the center of the image.

\[
K_{o1} = K_{o2} = \begin{bmatrix}
f & 0 & w/2 \\
0 & f & h/2 \\
0 & 0 & 1
\end{bmatrix}
\] (10)

Where ‘\(f\)’ indicates the focal length of camera, ‘\(w\)’ indicates the width of the image captured, ‘\(h\)’ indicates the height of the image.

The focal length is expected to vary in the interval given by,

\[
\left[\frac{1}{3}(w + h), 3(w + h)\right]
\] (11)

C. Stereo Matching

The stereo matching is a process of matching features or finding corresponding points between two images of a scene which helps to compute disparity between the pixels.

The technique used here for correspondence matching is Block Matching [17]. Block matching is performed using SAD (Sum of Absolute Differences) algorithm [18]. SAD is a fast template matching algorithm, which is advantageous to be used for real time scenes. For block matching, \(n \times n\) pixel block is taken around every pixel of the reference image. This block slides on the other view in a same row of pixel as images are rectified. The matching is done on the basis of SAD result. Suppose, we have two blocks to be compared namely ‘\(X\)’ and ‘\(Y\)’, SAD is performed with (12).

The block is matched to the reference block when the result of SAD is minimum. For example, take one reference matrix and perform SAD with other image matrices. The one which gives sum as minimum is a respective matched block. The SAD can be represented by (12).

\[
\chi(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} |X(i,j) - Y(i,j)|
\] (12)

Where, ‘\(\chi\)’ indicates the Sum of Absolute Differences to the matrix ‘\(X\)’ of one view and matrix ‘\(Y\)’ of another view.

D. Disparity

The disparity is defined as a displacement of the pixel when the pixels are matched in two different views. The calculation of disparity of the pixel is shown in the below fig. 6 and (13)

\[
d = X_l - X_r
\] (13)

Where, \(d = \text{disparity}, X_l = \text{Pixel in first view}, X_r = \text{Pixel in second view}\)

![Diagram of Disparity](https://via.placeholder.com/150)

The relation between depth and disparity [19] is shown in (14). It can be inferred from (14) that higher the disparity between the pixels of an object, less will be the depth of that object.

\[
Z = \frac{f T}{x_l - x_r}
\] (14)

Here, ‘\(Z\)’ indicates the depth of object from the camera; ‘\(f\)’ indicates the focal length of camera; ‘\(T\)’ indicates the distance between two fixed cameras; and \(x_l - x_r = \text{disparity between pixels}\).
III. RESULTS AND DISCUSSIONS

The approach used in this paper is implemented on many sets of rotated images taken in different environments. One pair of images is shown in Fig. 7. In this paper, the uncalibrated rectification is considered as compared to the conventional calibrated rectification. The uncalibrated rectification approach gives quite good results as compared to calibrated rectification results. Fig. 7 shows the original images captured at 5° rotation of camera along rotating platform.

![Fig. 7. Original Images taken at 5° rotation of camera](image)

The rectified images corresponding to original images are shown in Fig. 8. The circles marked in Fig. 8 are some reference points taken to get the depth information of these marked points. Table I. shows the marked points, The door knob is represented by a black circle, ‘V’ point on side glass is represented by a red circle, and the upper corner of the side poster is represented by the light pink circle.

![Fig. 8. Rectified Images](image)

Table I shows the pixel positions from the rectified images and the corresponding disparity calculated with respect to the ‘objects’ marked with colored circles in Fig. 8. It can be noticed that the ‘y’ coordinate in first view is matched with the ‘y’ coordinate in second view. It clearly shows that the images are rectified and the disparity present is along ‘x’ direction only.

![Table I. Table showing disparity between matched points](image)

<table>
<thead>
<tr>
<th>Objects</th>
<th>First View (x, y) in pixels</th>
<th>Second View (x, y) in pixels</th>
<th>Disparity $d = X_1 - X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Door Knob</td>
<td>(232,216)</td>
<td>(186,216)</td>
<td>46 pixels</td>
</tr>
<tr>
<td>‘V’ Point on Side Glass</td>
<td>(406,203)</td>
<td>(372,203)</td>
<td>34 pixels</td>
</tr>
<tr>
<td>Side Poster Upper Corner</td>
<td>(512,146)</td>
<td>(482,146)</td>
<td>30 pixels</td>
</tr>
</tbody>
</table>

With the calculation of disparity, the relative depth can be inferred as depth is inversely proportional to the disparity. Likewise, in given Fig. 7 and Fig. 8, the door knob is having highest disparity i.e., 46 pixels as compared to ‘V’ point on side glass which have 34 pixels and upper corner of side poster which have 30 pixels. From 14,

$$Z \propto \frac{k}{d}$$

(15)

Where, ‘$k$’ indicates a constant value.

![Fig. 9. Graph showing the relation between Disparity and Depth](image)

In Fig. 9, the graph represents the depth corresponding to the disparity. It can be noticed that if disparity is decreasing or declining, the depth increases. Hence, it is clear that door knob is the closest point as depth comes to be the least according to the above relation i.e. and Side poster is the farthest point among other points taken in Fig. 8.

![Table II. Table shows object distance and average % error](image)

<table>
<thead>
<tr>
<th>Objects Distance (in centimeters)</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>320</td>
<td>0.08125</td>
</tr>
<tr>
<td>415</td>
<td>4.40</td>
</tr>
<tr>
<td>532</td>
<td>7.70</td>
</tr>
</tbody>
</table>
Table II gives the average error at objects distance respectively.

Fig. 10 shows the average percentage error of the proposed system. The accuracy of the system depends on the quality of the camera and the rotation speed of the camera. For the proposed setup, the camera used is an off-the-shelf web camera of VGA resolution (640 × 480).

![Graph between Distance versus % Error](image)

Fig. 10. Graph between Distance versus % Error

IV. CONCLUSION

The proposed single camera system on a rotating platform is a frugal method to estimate depth information a given object. The proposed system to find relative distance of the objects is theoretically and experimentally proven with the methodology adapted. It is difficult to find the depth of far objects as the disparity between the pixels of both views tends to move towards zero. The proposed stereo matching algorithm works better for images having noticeable pixel differences i.e., for the images with good texture, contrast etc. In future, we are planning to work more on the challenges faced. The major challenge of the system includes quick disappearance of the objects from the field of view of the camera as the camera is rotating. The future work includes estimating the absolute distance of an object from the axis of the rotating platform and to build a complete depth map of the environment around the camera set up.

REFERENCES