Binocular Vision-Based Intelligent 3-D Perception for Robotics Application

R.T.H.S.K. Karunachandra
The Open University of Sri Lanka
H.M.K.K.M.B. Herath (kasunherathlive@gmail.com)
The Open University of Sri Lanka https://orcid.org/0000-0002-1873-768X

Research Article

Keywords: 3-D Depth Map, 3-D Reconstruction, Disparity Map, Feature Extraction, Stereo Matching, Stereo Vision

Posted Date: July 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-669187/v1

Version of Record: A version of this preprint was published at International Journal of Scientific and Research Publications (IJSRP) on September 18th, 2020. See the published version at https://doi.org/10.29322/IJSRP.10.09.2020.p10582.

Binocular Vision-Based Intelligent 3-D Perception for Robotics Application

R.T.H.S.K. Karunachandra [1], H.M.K.K.M.B. Herath [2]

Department of Mechanical Engineering, Faculty of Engineering Technology – The Open University of Sri Lanka [1, 2]

Abstract- Vision-based robotics has been the subject of several research contributions in the area of vision and control. Vision technology is becoming a pioneer in the most common applications such as localization, automated map creation, autonomous navigation, mapping analysis, or risk pattern prediction. The Stereo applications or programs use pairs of 2-D images as inputs and generate reconstructed 3-D imagery by locating the matching points. This paper introduces a method to the development of an algorithm of intelligent 3-D view reconstruction using the binocular vision for the robotic applications. The proposed system consists of two identical colour cameras and cameras were mounted as one stereo camera. 3-D reconstruction and visualization were performed according to the pair of 2-D images. Calibration, multi-view image acquisition, the stereo rectification process, and the disparity process were discussed in Section II. Real-time captured images and stereo image from Middlebury Stereo Datasets were used to test the system and verify results in Section III.

Index Terms- 3-D Depth Map, 3-D Reconstruction, Disparity Map, Feature Extraction, Stereo Matching, Stereo Vision

I. INTRODUCTION

Machine stereo vision, or also known as the stereoscopic vision has been an active area of robotics and engineering research for decades. It has been widely investigated before the emergence of event-based sensors. An autonomous robot needs to be aware of the three-dimensional state of the world to understand and think for its environment. However, the problem with vision is that the perceived image is a two-dimensional 3-D world projection [1]. Stereo vision must be viewed as a spatial integration of multiple viewpoints to recover depth, and a temporal integration is also possible.

Biology understands a scenario more easily than machines, even at smaller energy budgets *(Martin et al., 2018)*. Most animals have two eyes for a reason. Through eye's vision is combined in a stereoscopic reality that becomes a 3-D map of the human brain. Vision from each eye is only slightly different to a certain degree, and this variation is what helps us to perceive that anything is closer or farther away. In humans, stereopsis has become an attractive model system for understanding the neural activity-perception relationship *(Roe et al., 2007)*. Stereopsis has not been seen behaviorally in any non-human animal until 130 years after Wheatstone, with evidence of stereopsis by Bough in 1970 in macaque monkeys.

Modern machine's stereo algorithms are, to some extent, inspired by human stereopsis, which is powerful but also complicated and expensive [2]. Figure 1 illustrates the typical stereo vision system of a human. Stereo vision suggests numerous points of view and coordinating; thus, it gets profundity from a couple of pictures. Calculations for stereo system vision are additionally utilized prosperously in robotics [3].

Every visual sensor, whether artificial or biological, maps a 2-D representation of 3-D worlds. Depth sensors are the key to unlocking next-level machine vision applications in modern engineering. 3-D depth calculation and machine vision techniques are widely accepted in many applications, such as healthcare applications [4-8], autonomous navigation, teleoperation, or virtual/augmented reality modelling [9]. Google's Project Tango uses depth sensors to measure the real environment accurately and inform its graphics algorithms to virtual position content in the appropriate locations. Many warehouses are now using fully autonomous vehicles to carry items from one place to another. The vehicle's ability to travel on its own includes depth-sensing so it can know where it is in the world, where other important objects are, and most importantly how it can get from point to another point safely.

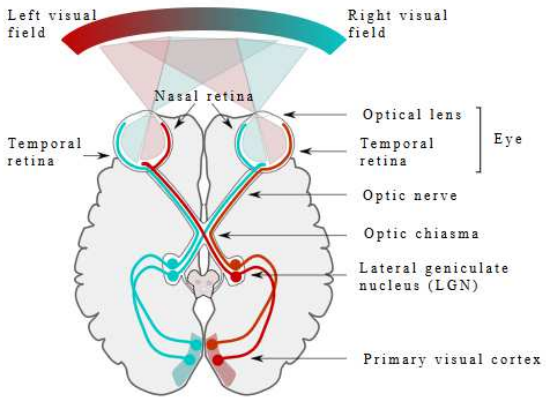

Figure 1. Stereo Vision System of Human [10]

Basic of Stereo Vision

Stereo technology takes 2-D picture stereo pairs as input and generates the replicated 3-D images by locating the respective positions. Most methods of stereo reconstruction are based on the use of model pinhole camera and parallel geometry.

Therefore, given that any stereo matching process identifies two locations of points, the depth is calculated from the different points or the distinction of the two points in the picture pixel coordinates.

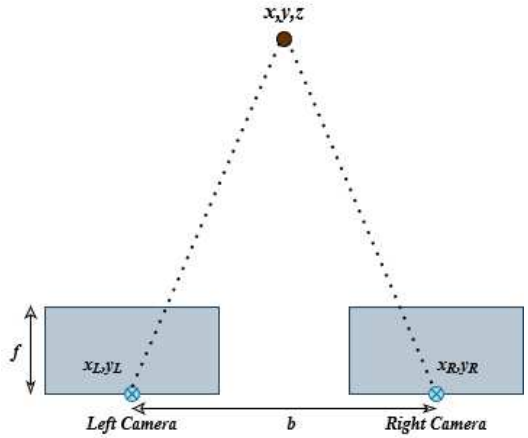

Figure 2. Illustration of the Stereo Geometry

Given the pixel coordinate (X_L, Y_L) at the left and (X_R, Y_R) at the right images, the coordinates of the 3-D universe (X, Y, Z) are determined as,

$$X = \frac{X_L b}{d}, Y = \frac{Y_L b}{d}, Z = \frac{fb}{d} \quad (1)$$

Where,

$d = (\mathrm{X_L} - \mathrm{X_R})$ in pixel, $f =$ Focal length of the camera,
$b =$ the parallax or interocular separation of camera (mm),

3-D performance of data reconstruction depends on the quality of the disparities, calibration, image rectification, and overall stereo system architecture. Figure 3 depicts the 3-D reconstruction model of the proposed system.

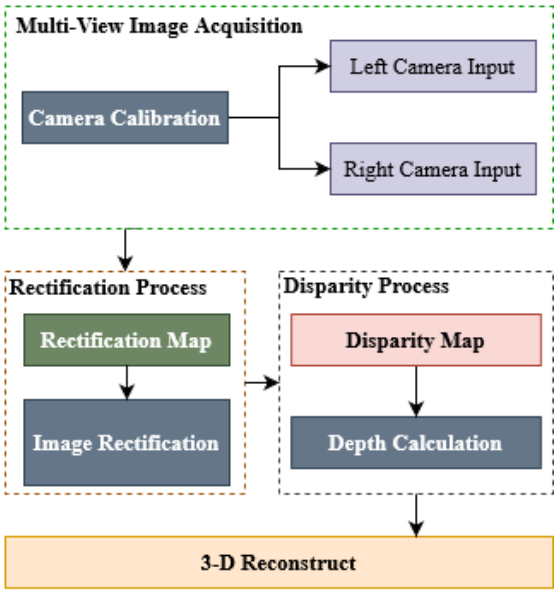

Figure 3. The Architecture of the 3-D Reconstruction Model

Image rectification

Methods of rectification are well known and have been studied extensively for years. These techniques are aimed at adjusting the captured images to simplify the problem of stereo correspondence. According to the optics, the resulting image varies from the real world geometry when the image is captured using an optical camera. Generally, there are two variables to be modified in stereo applications: image distortion and image epipolar geometry. This is known as the rectification process. If the stereo pair of images is fixed, then the problem of stereo correspondence simplifies and reduces from a 2-D search of order N^2 to a 1-D search of order N^1 for each matching pair of points on the same epipolar line [11].

Epipolar Geometry

Epipolar geometry is the geometry of the Stereo-Vision system. In case two cameras view a 3-D scene from two separate locations, there is an array of spatial relations between the 3-D focuses and their projections into the 2-D pictures resulting in imperatives between the focuses of seeing. These relations are established from the preface that the demonstrated pinhole device should surmise the cameras. Figure 4 illustrates the example of Epipolar Geometry.

Let us assume that the first camera is aligned with the world reference system with the second camera offset first by a rotation R and then by a translation T. This sets out the matrices for the image projection to be:

$$M = K[I \quad 0], M' = K'[R \quad T] \quad (2)$$

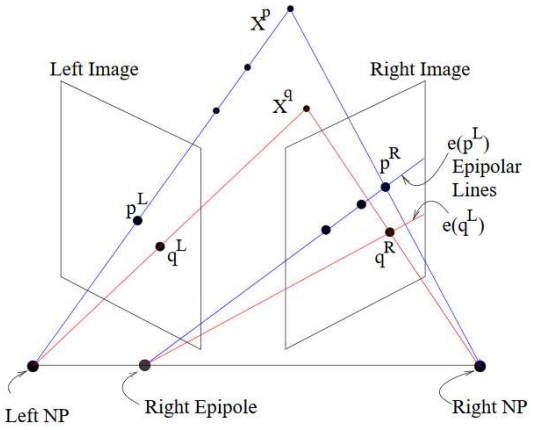

Figure 4. Epipolar Geometry

Let (u, v) be the pixel coordinate of the colour image; (X_C, Y_C, Z_C) is the corresponding coordinate in the colour camera coordinate system [12] Based on the principle of small hole imaging,

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 \\ 0 & \frac{1}{dy} \end{bmatrix} \begin{bmatrix} f & 0 \\ 0 & f \end{bmatrix} \begin{bmatrix} \frac{X_C}{Z_C} \\ \frac{Y_C}{Z_C} \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix} \quad (3)$$

Here, f is the focal length of the colour camera; $(c_{x,}c_y)$ are the coordinates of the principal point. d_x and d_y are physical sizes of the pixel in the horizontal and vertical directions, respectively.

Define $f_x = \frac{f}{d_x}$, $f_y = \frac{f}{d_y}$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{X_c}{Z_c} \\ \frac{Y_c}{Z_c} \\ 1 \end{bmatrix} \quad (4)$$

Here $f_{x,}\, f_y,\, c_x,\, c_y$ are the internal parameters of the colour camera.

II. METHODOLOGY

This study aims to develop an algorithm of intelligent 3-D view reconstruction using the binocular vision for machines and robotic applications. The methodology of the proposed system composed of calibration, image acquisition, pre-processing, stereo rectification, point-cloud generation, 3-D reconstruction, and visualization. Figure 5 illustrates the program flow chart of the proposed system.

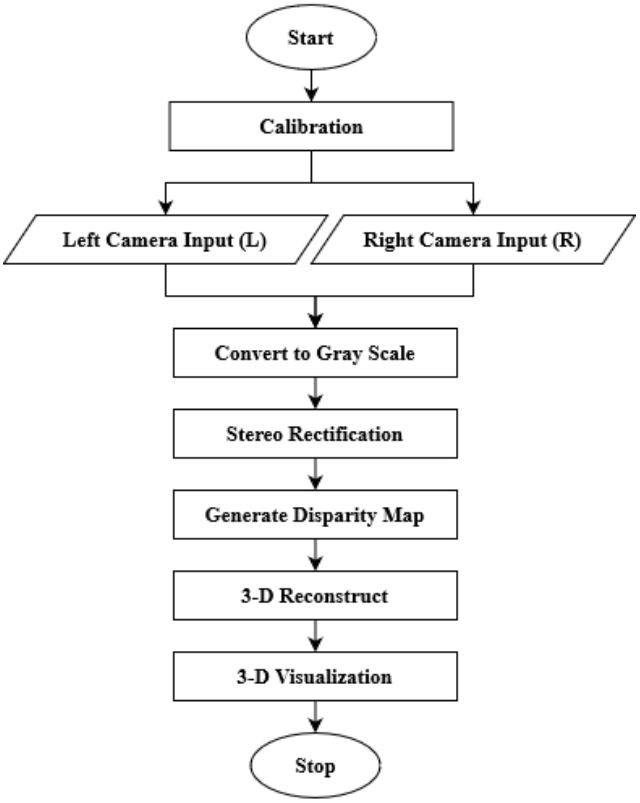

Figure 5. Algorithm Flow-Chart

Image Acquisition:

The proposed system was comprised of two optical cameras *(Raspberry Pi 5MP camera module, 60fps, 640×480 Pixels)*, positioned as a single stereo camera parallel to it. The system's most significant necessity is to guarantee that the frame of both cameras is recorded concurrently with the same brightness, exposure, shutter time, and parameters acquired. There are some methods of multiple-camera calibration which can overcome the adjustment of cameras for intrinsic and extrinsic parameters at the same time. Using pinhole cameras with nonlinear radial and tangential distortion compensation and the python language was used to develop calibration algorithms.

Grayscale Conversion:

In this study, acquired images were converted into grayscale and then passed results to the stereo rectification process. Grayscale is the set or range of monochrome *(gray)* shades that range from pure white on the lightest end to pure black on the other end. Grayscale includes only information about luminance *(brightness)* and no information about colours [13-14]. That's why the highest luminance is white, and the minimum luminance is black; the shade of gray is everywhere in between. Therefore, grayscale images contain only shades of gray and no colours. In this study, acquired stereo images were converted into the grayscale using the weighted grayscale method.

- **Unweighted:** We simply take on average the red, green, blue pixel data in this case. There's no bias, and there's no connection with human vision [15].

$$pixel\,[gray] = pixel\,([red] + [green] + [blue]) / 3 \quad (5)$$

- **Weighted:** In this scenario, we take into account the human eye's sensitivity factor in different colours and set the bias to the average as a result [15].

$$pixel\,[gray] = pixel\,([red] \times 0.299 + [green] \times 0.587 + [blue] \times 0.144\,) \quad (6)$$

Stereo Rectification:

Methods of rectification are well known and have been practised widely for years. These techniques are designed to adjust the captured images to quantify the analysis of stereo correspondence. Due to the lenses, the resulting image varies from the real world geometry as an optical system takes the image. For stereo implementations, there are essentially two variables that must be modified, such as image distortion and image epipolar geometry. In this study, Image rectification was performed in the mage calibration process.

Finding the Disparity (Sum of Added Differences):

The Sum of Added Differences *(SAD)* is a way to determine the disparity. Since the images are represented as 2-D arrays, it will be associated with any block of *(m x n)* pixels. If all of the pixels match perfectly, then all of the colour values associated with each pixel is the same. Then the two blocks are going to be identical. But, in stereo pairs, these identical pairs don't occur so we need to search for the block that has the closest match. It will teach how to obtain measurements of the SAD.

$$m \times n\ Array\ for\ Right\ Block = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad (7)$$

$$m \times n\ Array\ for\ Left\ Block = \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{bmatrix} \quad (8)$$

$$SAD = \sum_{i=1}^{n} \sum_{j=1}^{m} (a-b)_{i,j} \quad (9)$$

$$SAD = \begin{bmatrix} (a-b)_{11} & \dots & (a-b)_{1n} \\ \vdots & \ddots & \vdots \\ (a-b)_{m1} & \dots & (a-b)_{mn} \end{bmatrix} \quad (10)$$

Mapping the Disparity in Three Dimensions:

To obtain measurement points, it is important to determine a disparity map before constructing an occupancy grid through stereo-vision. When the magnitude of the disparity increases, the warmth of colour grows proportionally. The height, width of the image are the x , y axes and each combination *(x , y)* represents every pixel in the image. Those pixels in the map displays the disparity which was simply a numerical representation of how close the pixel is to the camera. Finally, the distance information was plotted in the *z*-axis and plotted against the coordinates *x* and *y* (figure 6).

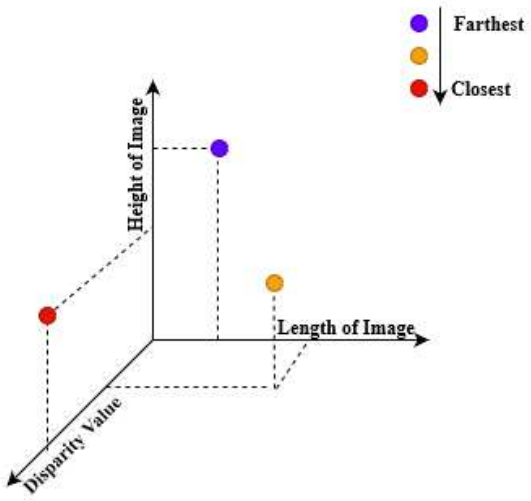

Figure 6. Disparity Visualization in Three Dimensions

Generating Point Cloud:

Point clouds are a means of assembling a significant number of single spatial measurements *(x, y, and z)* into a dataset that can then represent a whole *(object or space)*. Figure 7 illustrates a sample point cloud image of a Torus. Such points represent the geometric coordinates of a single point on a sampled surface underlying the *x, y,* and *z*. Several formats may be used to store a cloud of data. Essentially, any format which can store three numbers representing the coordinate *x, y,* and *z* can be used. Many formats are widely used for processing point clouds, however. These formats can be classified into Binary and ASCII types [16].

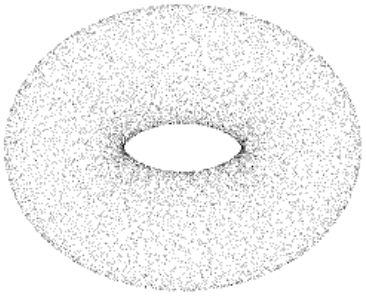

Figure 7. A Point Cloud Image of a Torus [17]

3-D Visualization:

MeshLab [18] is an open-source program that is widely used for 3-D triangular mesh creation and editing. It provides a set of tools that can be used to edit, clean, heal, inspect, render, texture, and convert meshes. In this study, we were used MeshLab to 3-D visualization based on the point cloud dataset, which generated in the point cloud generation process.

III. Results and Discussions

In this section, we were demonstrated depth map results using four image sets from two categories.

- **Category 1:** Experiment one was conducted in this segment, and the image was captured in real-time using two cameras that were functioned as a single stereo camera. The image shown in figure 8 and figure 9 is a man sitting on a chair with his left hand holding a helmet.

Experiment One

Figure 8 and figure 9 illustrate the left and right images acquired by two cameras in experiment one. The yellow horizontal line on the left and right image represents the epipolar line. A patch was marked on the left image's epipolar line, as shown in Figure 8.

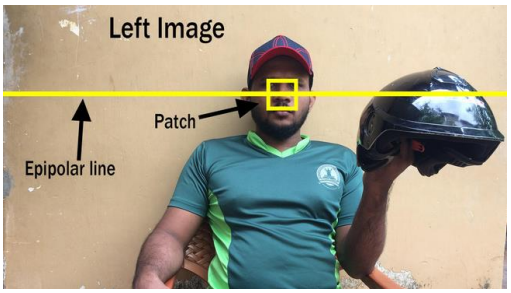

Figure 8. Acquired Left Image of Experiment 1

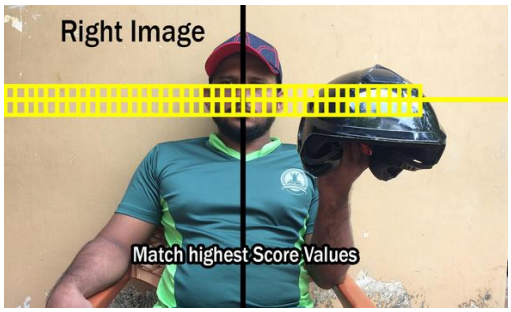

Figure 9. Acquired Right Image of Experiment 1

In the stereo matching process, the algorithm for a corresponding point is not searched for the entire 2-D right image. The "epipolar constraint" reduces the search space to a one-dimensional line. The patch in the left image was compared with the patches along the same row in the right image.

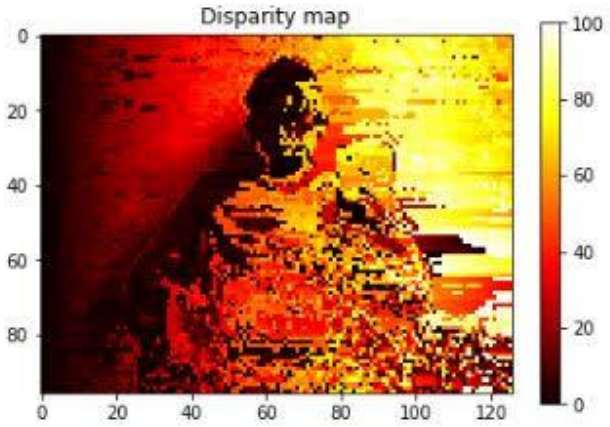

Figure 10. Disparity Map of Experiment 1

To achieve measurement points, it is important to determine a disparity map before constructing an occupancy grid through stereo-vision. Figure 10 illustrates the disparity map of the test one.

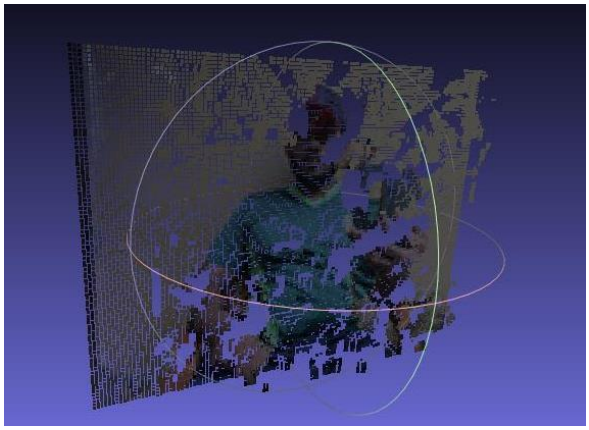

Figure 11. 3-D Visualization of Experiment 1

Figure 11 illustrates the 3-D reconstruction of the stereo image acquired by experiment one. 3-D visualization was archived through the point cloud generation.

- **Category 2:** Three stereo images from 2005[19] and 2014 [20] Middlebury Stereo Datasets, which is publically available for researchers in *vision.middlebury.edu* [21-22] was used for this category.

Experiment Two

Experiment two was carried out by using a Middlebury stereo image, as shown in figure 12. The depth map for the second experiment is shown in figure 13.

Figure 12. Stereo Image of Experiment Two

Figure 13. Depth Map of Experiment 2

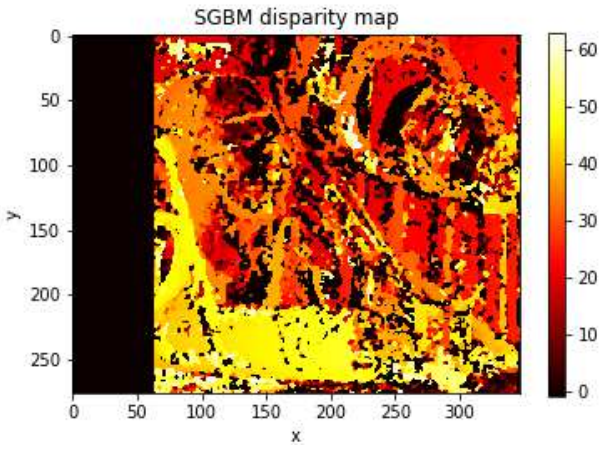

Figure 14. Disparity Map of Experiment 2

As shown in figure 13, the differences between the two images give depth information. This depth information is visualized as the depth map. Due to the low patch values disparity map for the second test was reduced it's detailed as shown in figure 14. Close objects were resulted in a large disparity value. This is translated into light greyscale values and objects further away will appear darker.

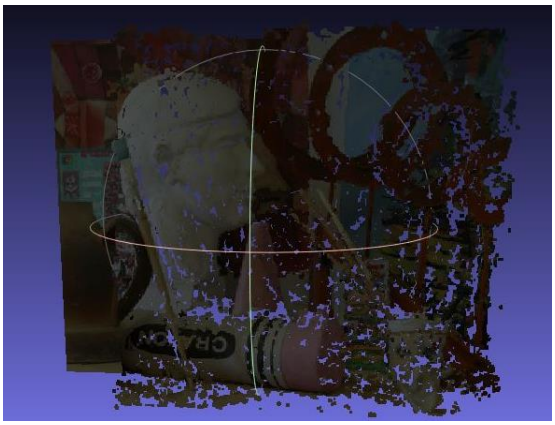

Figure 15. 3-D Visualization of Experiment 2

Experiment Three

Figure 16 to Figure 19 is a representation of the stereo image, depth map, disparity map, and 3-D visualization of experiment three.

Figure 16. Stereo Image of Experiment 3

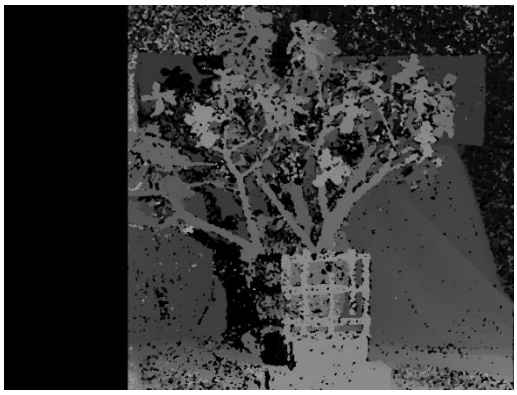

Figure 17. Depth Map of Experiment 3

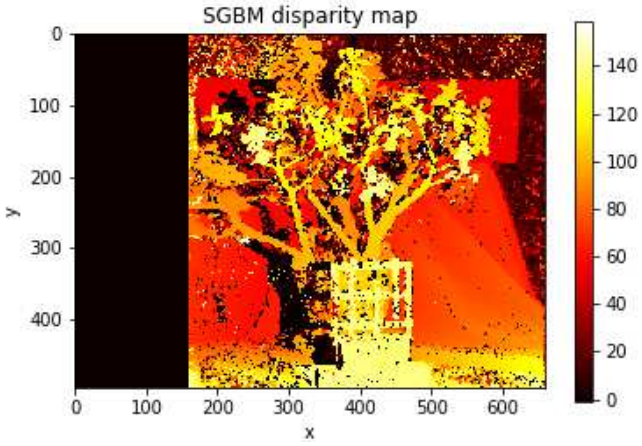

Figure 18. Disparity Map of Experiment 3

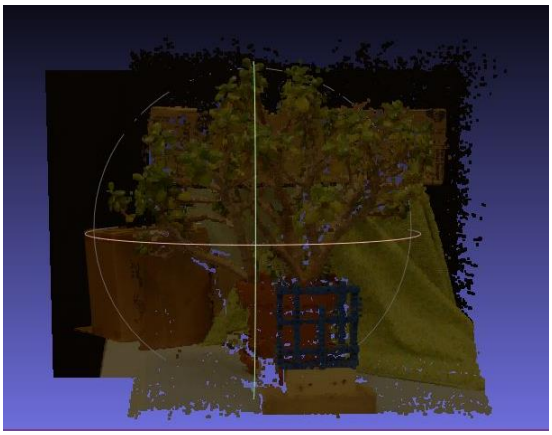

Figure 19. 3-D Visualization of Experiment 3

Figure 17 depicts the depth map of the third experiment. The image shows more and precise detail. The noise in the image was more prominent at the first observation. After applying more significant patch values, the noise of the image was reduced. But this action was lead to a decrease in the precision and details of the image being constructed.

Experiment Four

Figure 20 to Figure 23 represents the results for experiment four.

Figure 20. Stereo Image of Experiment 4

Figure 21. Depth Map of Experiment 4

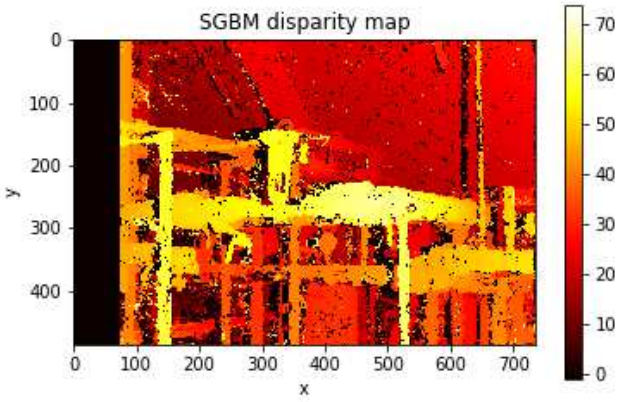

Figure 22. Disparity Map of Experiment 4

Figure 22 illustrates the disparity map for the fourth experiment. Due to the small patch value, large amounts of noise were observed.

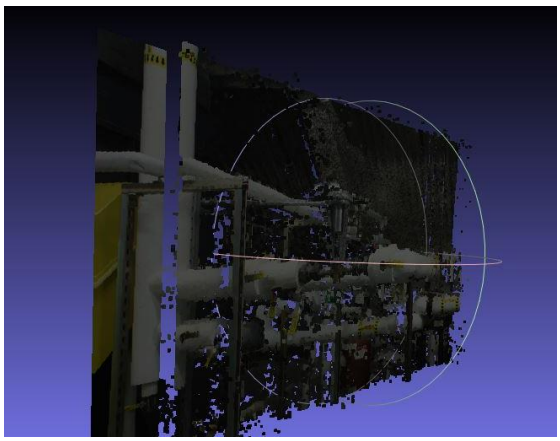

Figure 23. 3-D Visualization of Experiment 4

IV. CONCLUSIONS

In this paper, we presented the development of an algorithm of an intelligent 3-D perception for robotic applications based on the binocular vision using two identical cameras. Stereo imaging is a passive technique that can restore the environmental structure by comparing the features observed in different photographs of the same scene. This algorithm can be used utilizing robotic hands which are guided by visual perception for instruments equipped for handling devices.

These experiments are used two identical cameras which were mounted as a single camera to get a standard epipolar line. Stereo matching is the most crucial step in binocular vision reconstruction. Due to the lack of perfect stereo image acquisition from both cameras, the 3-D visualization was not correctly completed in experiment one. The cloud point development failed because of the errors caused in the process of calibration and image acquisition.

According to the results, the disparity was larger (brighter) for closer surfaces. Figure 24, 25, and 26 represents the depth (z) against disparity (px) for experiment 2, 3, and 4.

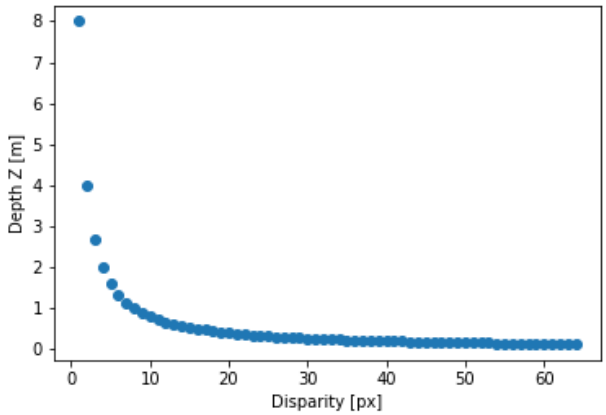

Figure 24. Disparity Vs. Depth of Experiment 2

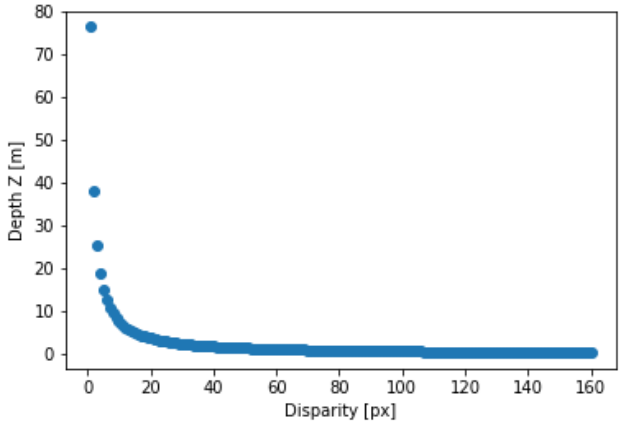

Figure 25. Disparity Vs. Depth of Experiment 3

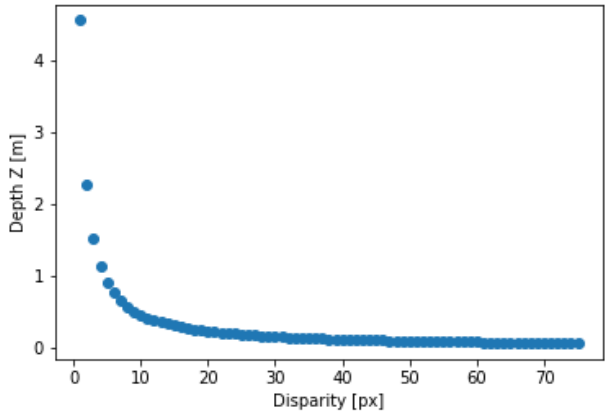

Figure 26. Disparity Vs. Depth of Experiment 4

Table 1 shows the focal length, baseline, and maximum disparity values for experiments 2, 3, and 4, which were observed during the experiments. Each experiment was associated with the stereo images that were taken from the Middlebury stereo dataset.

Table 1. Focal Length, Baseline and Maximum Disparity Values for Experiment 2, 3, and 4

Experiment No	Focal Length /(mm)	Baseline Value /(mm)	Maximum Disparity /(px)
2	50	160	64
3	201	380	160
4	19	237	75

According to the results, some errors were identified due to the camera calibration, low image resolution, occlusion, violations of brightness constancy, large motions, and low-contrast image regions. For this experiment, the use of a single stereo camera is highly recommended. Using two cameras that work as a single camera causes errors in epipolar line development. Hence the 3-D reconstruction development was also not perfect. The author's next step is to minimize the errors and reconstruct 3-D visualization for the development of a laboratory environment for remote laboratory.

REFERENCES

[1] De Cubber, Geert & Nalpantidis, Lazaros & Ch, Georgios & Sirakoulis, & Gasteratos, Antonios. (2008). Intelligent Robots need Intelligent Vision: Visual 3D Perception.

[2] V, Nityananda and J. C. A., Read. (2017). Stereopsis in animals: evolution, function and mechanisms, The Journal of Experimental Biology, vol. 220, no. 14, pp. 2502–2512.

[3] L, Steffen & D, Reichard & J, Weinland & J, Kaiser & A, Roennau & R, Dillmann. (2019). Neuromorphic Stereo Vision: A Survey of Bio-Inspired Sensors and Algorithms. Frontiers in Neurorobotics, vol. 13.

[4] Silva, Tharindu & Madhusanka, Achintha & Priyankara, Hapuarachchige. (2020). Vision Based System for "Free-Weight Back Squat" Angle Assessment. 10.13140/RG.2.2.35805.23523.

[5] Sanjeewa, E.D.G. & Herath, K K L & Madhusanka, Achintha & Priyankara, Hapuarachchige. (2020). Visual Attention Model for Mobile Robot Navigation In Domestic Environment. 10.13140/RG.2.2.32638.20804.

[6] Herath, K K L & Sanjeewa, E.D.G. & Madhusanka, Achintha & Priyankara, Hapuarachchige. (2020). Hand Gesture Command to Understanding of Human-Robot Interaction. 10.13140/RG.2.2.18115.43049.

[7] Madhusanka, Achintha & Ramadass, Sureswaran. (2020). Recognition of Daily Living Activities Using Convolutional Neural Network Based Support Vector Machine. 10.13140/RG.2.2.18910.05444.

[8] Madhusanka, Achintha & Jayasekara, Buddhika. (2016). Design and Development of Adaptive Vision Attentive Robot Eye for Service Robot in Domestic Environment. 10.1109/ICIAFS.2016.7946529.

[9] C. L. R., Team. Depth sensors are the key to unlocking next level computer vision applications. Medium, 27-Jul-2017. [Online]. Available: https://blog.cometlabs.io/depth-sensors-are-the-key-to-unlocking-next-level-computer-vision-applications-3499533d3246. [Accessed: 16-Aug-2020].

[10] Visual system. Wikipedia, 31-Jul-2020. [Online]. Available: https://en.wikipedia.org/wiki/Visual_system. [Accessed: 15-Aug-2020].

[11] A. Lipnickas and A. Knyš. (2019). A Stereovision System for 3-D Perception, Elektronika ir Elektrotechnika, pp. 99–102.

[12] Y. Yang, X. Meng, and M. Gao, "Vision System of Mobile Robot Combining Binocular and Depth Cameras," Journal of Sensors, 24-Sep-2017.

[13] Karunasena, G.M.K.B. & Priyankara, Hapuarachchige. (2020). Tea Bud Leaf Identification by Using Machine Learning and Image Processing Techniques. International Journal of Scientific & Engineering Research. 10.14299/ijser.2020.08.02.

[14] Herath, K., & de Mel, W. R. (2017). Rice Grains Classification Using Image Precessing Technics. International Journal of Scientific and Engineering Reseach, 10-14.

[15] S, Paul. (2019). Stereoscopic Depth Sensing – A Python Approach. Artificial Intelligence - The Key To The Next Door of Evolution.

[16] Landa, Jaromir & Procházka, David & Stastny, Jiri. (2013). Point cloud processing for smart systems. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. 61. 2415-2421. 10.11118/actaun201361072415.

[17] L. Vieira , "A point cloud image of a torus," Wikipedia, 01-Aug-2020. [Online]. Available: https://en.wikipedia.org/wiki/Point_cloud. [Accessed: 16-Aug-2020].

[18] MeshLab. [Online]. Available: https://www.meshlab.net/. [Accessed: 16-Aug-2020].

[19] D, Scharstein and R, Szeliski. (2003). High-accuracy stereo depth maps using structured light. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), volume 1, pages 195-202, Madison.

[20] D, Scharstein & H, Hirschmüller & Y, Kitajima & G, Krathwohl & N, Nesic & X, Wang and P, Westling. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In German Conference on Pattern Recognition (GCPR 2014), Münster, Germany.

[21] D, Scharstein and C, Pal. (2007). Learning conditional random fields for stereo. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN.

[22] H, Hirschmüller and D, Scharstein. (2007). Evaluation of cost functions for stereo matching. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN.

AUTHORS

First Author – Mr. R.T.H.S.K. Karunachandra
Mechatronics Engineering Student at The Open University of Sri Lanka
Faculty of Engineering Technology, The Open University of Sri Lanka
E-mail: karunachandra1993@gmail.com

Second Author – Eng. H.M.K.K.M.B. Herath
B.Tech Eng. Mechatronics, AMIE(SL), AEng EC(SL), MIEEE
Faculty of Engineering Technology, The Open University of Sri Lanka
E-mail:kasunherathlive@gmail.com