IJCTA, 9(39), 2016, pp. 127-141
© International Science Press

Estimation of Baseline of Single Camera Stereo Vision Based on an Inspiration from SFF

C.R. Srinivasan* Senthilnathan R. Subhasree P.*** Sivaramakrishnan R.*** Karthikeyan P.*** and Srividya.R******

Abstract : Computation Stereo Vision is a widely researched technique in the field of computer vision for scene reconstruction. One of the main issues to be addressed in stereo vision is the trade-off that needs to be achieved between accuracy and resolution. A wide-baseline offers better resolution in depth estimated, contrarily a narrow baseline though offers good accuracy but suffers from poor depth resolution. The proposed work in this paper presents a variable baseline stereo vision system which permits a knowledge and control of the depth resolution for a range of geometries that the system can handle. This is achieved using an algorithm inspired from Shape From Focus (SFF) technique in computer vision. In the current research work a new SFF-inspired algorithm is developed which utilizes images acquired with low focal length lenses in place of a telecentric lens. Based on the sparse and coarse depth map obtained an approach for determining the baseline of a single camera based stereo vision system for any desired depth resolution is presented in this paper.

Keywords : Shape From Focus (SFF).

1. INTRODUCTION

Computer vision being one of the most research area, has been increasingly finding its utilization in many areas like inspection, guidance, measurements, entertainment and other scientific and industrial applications. Scene reconstruction and pose estimation are two important problems in the field of 3-D computer vision has been the topic of interest for over three decades. Many passive, active and hybrid techniques have been presented in the open literature over the years. Passive methods for 3-D reconstruction aim at estimating the depth map of the scene with one or more cameras which records intensity information of the scene, while active vision techniques reconstructs a scene by purposefully releasing some form of energy into the scene. Hybrid methods such as the structured light technique combine the active and passive approach by throwing light patterns on the scene which is imaged by a passive camera. The methods for scene reconstruction are generally termed as Shape-from-X, where X denotes the technique or the cue used for the depth estimation. Out of various methods for scene reconstruction, Shape from Stereo is the most researched topic. One of the challenging problems in stereo vision is selecting the relative pose between the two or more vantage points from which images of the scene are acquired. The main issue is to achieve a trade-off between accuracy in matching and resolution in reconstructed depth. There are two main techniques which explicitly use focus as cue to estimate depth namely, Shape from Focus (SFF) [1] and Shape from Defocus [2] (SFD). SFF generally requires more number of images acquired from different

* Department of Instrumentation and Control Engineering, Manipal Institute of Technology, Manipal University, Manipal

** Department of Mechatronics Engineering, SRM University, Katankulathur, India

*** Department of Production Technology, MIT Campus, Anna University, Chennai, India

**** Department of Electrical and Electronics Engineering, Manipal Institute of Technology, Manipal University, Manipal
E-Mail– cr.srinivasan@manipal.edu

focal distances. This fact makes the SFF method computationally expensive and time consuming compared to SFD. Given this drawback, the precision of SFF is better than SFD. One of the inherent limitations of the SFF method of reconstruction of the scene is that they are highly sensitive to parallax. Many literature presented method which attempt to extend the applicability of SFF techniques by means of new image processing procedures [3]. The attempts made for increasing the applicability of the SFF technique is still bound to limitations of the conventional SFF in terms of the size of objects used. This paper presents a portion of the research work which deals with development of an algorithm inspired by conventional SFF which may be used for scenario involving structure dependent pixel motion. This increases the size of the objects that may be subjected for reconstruction at the cost of density and accuracy of reconstruction. The paper presents approaches which may use such sparse information for estimating the baseline of a single moving camera stereo vision system. A theoretical possibility is presented in the earlier work of the author [4]. Previous work [5, 6, 7] presented results of evaluation of focus measures based on image gradient, Laplacian and statistical measures, which is also part of the larger research work. Since the current paper is also part of a larger research work few of the common basic contents are carried over from the previous articles [5, 6, 7].

2. EXPERIMENTAL SETUP

The experimental hardware must basically support the requirements of the SFF-inspired algorithm and the stereo vision algorithms to achieve variable baseline. The setup required for SFF must be a precise translating mechanism along the optical axis of the camera. The photograph of the experimental setup is shown in Fig. 1.

Table 1
Imaging Specifications

Parameter	*Specification*
Light Source	White LED area light
Illuminance	160 lux
Lighting Technique	Partially diffused bright field incident lighting
Lens Type	Fixed focal length prime lens
Focal Length & f#	16 mm and 1.3
Camera Make and Model	Allied Vision Technologies Guppy F033b
Interface	IEEE 1394a – 400Mb/s, 1 port
Computer Interface	PCI – IEEE 1394a
Resolution	640 × 480
Aspect Ratio	4:3
Sensor	Sony ICX424
Sensor Type	CCD Progressive, Monochrome
Operating Frame Rate	45 Frames Per Second (fps)
Mode of operation	Mono8 mode
Trigger Type	Software trigger
Image Acquisition Time	180ms per image
Processor	Intel Core i5, 2.5GHz Quad Core
Memory	4 GB, 1300 MHz primary memory

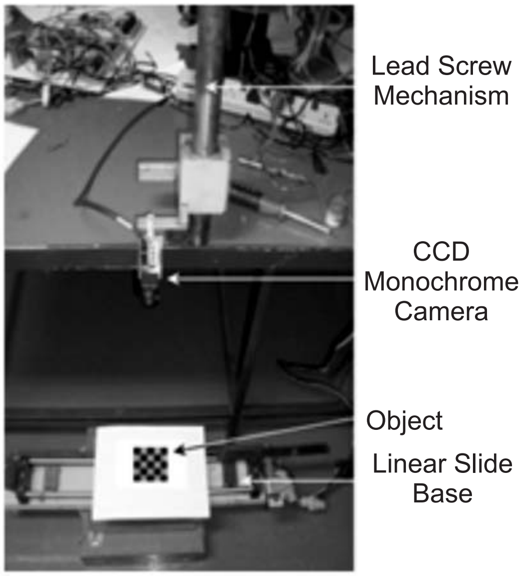

Figure 1: Experimental Setup

3. SFF-INSPIRED ALGORITHM

The following section of the paper presents the proposed SFF-inspired algorithm, which forms the central theme of the larger research work, a portion of which is presented in this paper. First a set of feature points present across the stack of images is detected using Speeded-Up Robust Features (SURF) feature detector [8]. The stack suffers from combined variations in focus and magnification because of the relative motion between the camera and the scene. The focus measure of only those particular pixels is computed. This is different from the conventional SFF route, where the focus measure is computed for all the pixels in all the images in the focal stack. Since in the current study, there is a finite pixel motion, focus measures cannot be directly applied to all the images in a direct manner. Conventional SFF uses a focus function, such as a Gaussian distribution, and interpolates the computed focus measures to obtain accurate depth estimates. Such a model is suitable only when a telecentric lens is used. This is because in the conventional SFF the depth of field is very limited, and no magnification changes occur due to the relative motion between the camera and the scene. In the current research, since a wide angle lens is used with a higher depth of field, in order to achieve a complete Gaussian distribution, large camera motion may be required. Large camera motion in turn causes extremely low magnification after which the spatial resolution of the image becomes too poor for any further measurements possible from the images. Because of these facts, a coarse method of depth estimate is adopted for the current work. The algorithm may be summarized as follows:

- The initial location of the camera from the measurement plane is known a priori as sm, where $m = 1$ for the initial location of the camera.
- Accumulate the image sequences acquired at each step m where the stand-off distance (sm) increases in steps of Δd.
- Measure focus, Fm for each of the SURF feature points, across the stack at each step whose correspondences are matched using the SSD metric.
- Find the step number m in which the focus measure is the maximum for a point (x, y), such that Fm = Fmax, where Fmax is the maximum value of the focus measure for a particular pixel.
- Assign the value of the distance of the camera motion as the height of the object corresponding to the particular pixel, such that the height of the scene point $\overline{h} = m\,\Delta d$.
- Once the height of a point is computed, the depth of the point Cz may be computed as Cz = $s1 - \overline{h}$.

This algorithm, as may be observed, gives only a rough estimate of the depth. The performance of the algorithm is directly dependent on the selection of Δd. Lower values of Δd give better accuracy, albeit there is always a non-zero resolution error. Interestingly, at times the estimated depth becomes equal to the actual depth, depending on the particular scene point under consideration. In other words the depth error may be zero, although the system as a whole suffers from a non-zero resolution error. SFF requires a continuous range of images to be acquired, by having a linear translational motion of the camera along the optical axis with respect to the scene. By a continuous range is meant, that the distance of separation between the various vantage points of image acquisition must be as small as possible. A sequence of 15 images is acquired during the camera motion, though more number of images may help in reducing the resolution error. In any case, SFF always suffers from a non-zero resolution error due to the discrete number of steps. The total travel range is kept different for different scene geometries to ensure no loss of generality. This indicates that for different trials, the inherent resolution error is different, since the number of acquired frames is constant. The 16mm, f1.3 lens used for imaging is almost midway between a large aperture leading to a small depth of field, and a small aperture which causes low image brightness and a large depth of field. One of the important points in the present research is that the depth of field of the lens setting should always be less than at least half the height of the object. This is mandatory, since only for such a setting of aperture, a focus change may occur in the measurable order in the pixels of the image, for a relative motion between the camera and the object. The camera moves away from the scene or the measurement plane when an image sequence is acquired. This means that the first step in the sequence has the highest spatial resolution. As the camera moves away from the objects in the scene the magnification of the image gradually decreases depending on the focal length of the lens used. Table 2 shows the magnification variation across the image sequence for the 16mm lens.

Table 2

Magnification Factor

Step	*Magnification Factor*
1.	1
2.	0.9784
3.	0.9585
4.	0.9452
5.	0.9336
6.	0.9220
7.	0.9087
8.	0.8971
9.	0.8839
10.	0.8689
11.	0.8573
12.	0.8457
13.	0.8341
14.	0.8242
15.	0.8175

It is obvious that a low focal length lens exhibits faster changes in magnification whereas a higher focal length lens such as a telecentric lens experience minimum or no change in magnification which is exploited in conventional SFF. The feature detection is carried out using the SURF algorithm relies on the matrix, called the Hessian matrix, for scale and location. The following section presents only the basic necessary information about the implementation of the SURF algorithm. Detailed working of the algorithm is presented in many literatures [8]. In the SURF algorithm, the difference in scaling is detected, based on the number of octaves to be used, which is specified as an input to the algorithm. Each octave spans for a number of scales that are analyzed, using varying window sizes. For an octave value of three used in the current study, the window sizes range from an initial size of 27 × 27 to 99 × 99 constituting 4 scales for each octave. In general, at least three different levels are required to analyze the data in a single octave. The size of the octave is chosen based on the image size. Generally, for an image size of 50 × 50 not more than 2 octaves are required. In the implementation of SURF, a non-negative scalar called the metric threshold is chosen for selecting the strongest features. More blobs can be obtained for a smaller value of the metric threshold. In the current research, the metric threshold is chosen to be equal to 100. During the feature detection only the strongest features are considered as a priority, though during the execution of the SFF algorithm more features help in increasing the density of the depth maps.

Once the features and their descriptors are extracted from the images in the sequence, the next step is to establish some feature matches between the images. The feature points are identified across the sequence of images so that the focus can be measured for all the corresponding points. The identification of points is subject to the existence of a point in all the images in the focal stack. There are many matching strategies and metrics available in the computer vision literature. In the case of situations involving multi-image correspondence identification such as the one involved in this research study, the nearest neighbor distance to that of the second neighbor is a useful heuristic. In the current research this method is used in the form of Nearest Neighbor Distance Ratio (NNDR) which is defined as follows:

$$\text{NNDR} = \frac{d_1}{d_2} \tag{1}$$

where d_1 and d_2 are the nearest and second nearest neighbor distances respectively. Two feature vectors would match when the distance between them is less than the NNDR threshold of 0.6. The chief advantage of NNDR compared to other matching methods like the nearest neighborhood symmetry, and those based simply on the threshold, is its ability to eliminate ambiguous matches apart from using a match threshold. The match threshold is set at 10 in a possible range from 0 to 100. The higher the value used for the match threshold, the more the number of matches obtained with higher chance of false matches. The metric values obtained from a suitable function are actually matched in order to find the corresponding points. The match metric used in the current research is the Sum of Squared Differences (SSD) which is defined as follows:

$$\text{SSD}(r, c) = \frac{1}{n}\sum_{u,v \in \mathrm{I}_1, \mathrm{I}_2} \left(\mathrm{I}_1(x, y) - \mathrm{I}_2(r + x, c + y)\right)^2 \tag{2}$$

where I_1 and I_2 are the feature sets from the two images subjected to matching. The variables *x* and *r* are the row coordinates of the feature set and the variables y and c are the column coordinates of the feature set subjected to matching. The first step is the process of identifying the corresponding locations of the feature points in all the 15 images. Once the feature points are identified the features from the first image is taken as the reference, with which features from the rest of the fourteen images acquired from different focal distances, are matched. Fig. 2 shows the matching of the corresponding points between the first image and second image and between the first image and the fifteenth image in the focal stack.

One of the main issues to be addressed in the matching process is the problem of false match identification which may be observed in both the images. Since the images contain similar features in the form of textures of different scaling, they contribute to false matches. The problem of false matches is tackled by estimating the average distance between the corresponding matching points, and any points whose distance exceeds the threshold, selected based on the average distance, are suitably eliminated. From the figures it may also be observed, that the number of corresponding points of the first image in the fifteenth image were much less compared to the points obtained in the matching of the second image. This is mainly due to the scaling effect to which the matching algorithm is inert. It must be noted that the matching algorithms utilize a constant size of the region of interest, *i.e*., the feature set, but the spatial resolution of the images subjected to matching is different.

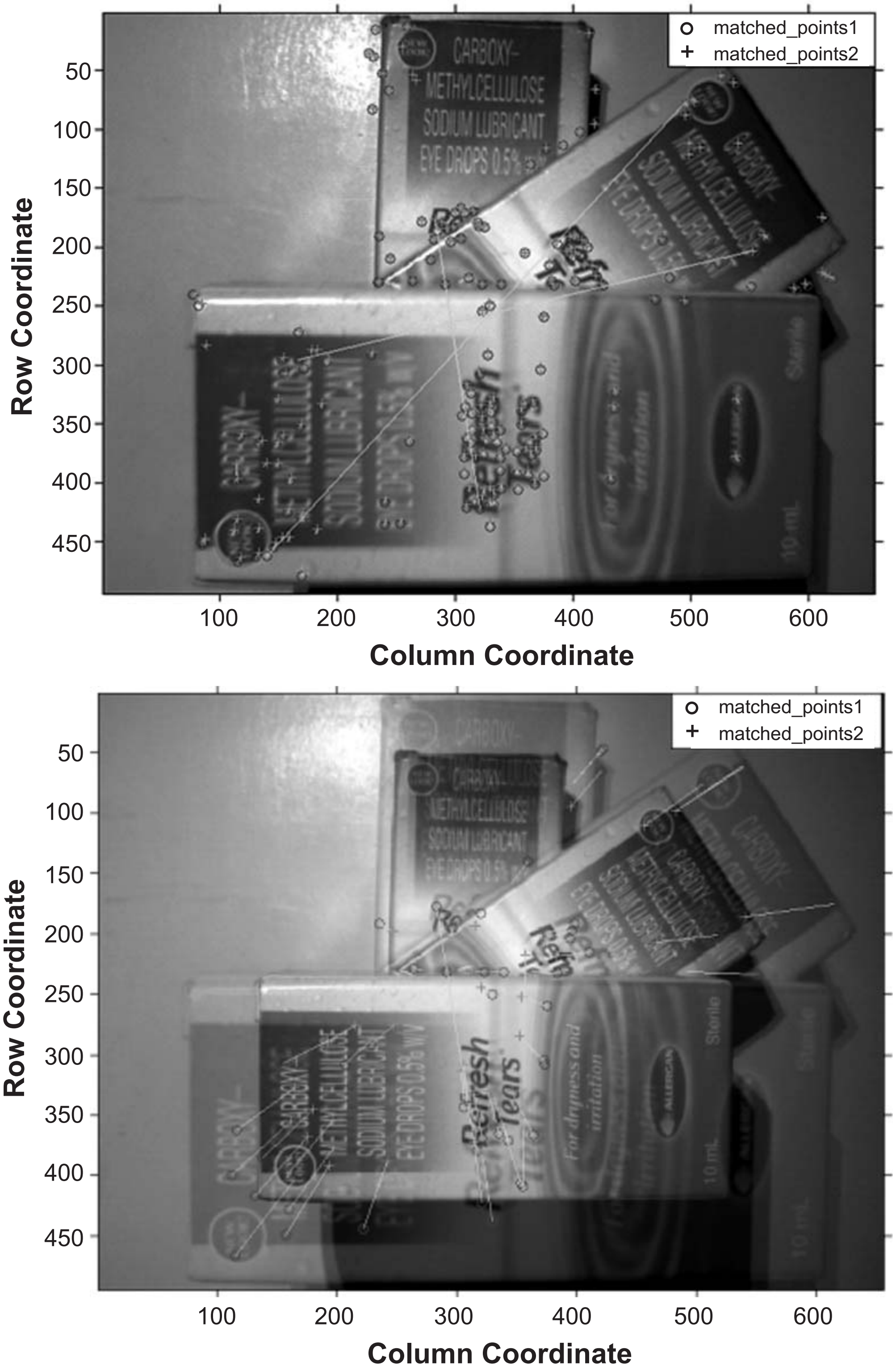

Figure 2: Corresponding Points in the First, Second Image and Fifteenth Image

3.1. Focus Measure

A focus measure is a mathematical function, which gives a measure of the focus of the image indirectly, by measuring the local grey-level differences in the image. It is generally computed in a small square window around the pixels in the image. Generally a high value for the focus measure indicates a sharply focussed region in the image, and a low value indicates blurred regions. Many focus measures were presented in the SFF and autofocus literature. Based on an evaluation carried out as part of the current research work a combined focus measure is used for the purpose. The focus measure is chosen after careful evaluation under a number of operating conditions such as different spatial resolution, window size, contrast changes, gray level saturation and camera noise. A portion of the earlier work is presented in [5, 6, 7] where gradient based focus measures are evaluated for study the performance in terms of accuracy and execution time. After the evaluation the following function is used as the focus measure:

$$\mathrm{F}_m = \mathrm{F}_{\mathrm{SG}} + \mathrm{F}_{\mathrm{EL}} + \mathrm{F}_{\mathrm{HR}} + \mathrm{F}_{\mathrm{B}} \tag{3}$$

where F_m is the summation of Squared Gradient [9], Energy of Laplacian [10], Histogram Range [11] and Brenner's Measure [9]. The individual functions are defines as follows:

$$\mathrm{F}_{\mathrm{SG}} = \sum_{(i,\,j)\in\Omega(x,y)} \left|\mathrm{I}(i,\,j+1)-\mathrm{I}(i,\,j)\right|^2 \geq \mathrm{T} \tag{4}$$

$$\mathrm{F}_{\mathrm{EL}} = \sum_{(i,j)\in\Omega(x,y)} \Delta\mathrm{I}(i,\,j)^2 \tag{5}$$

where ΔI is the image Laplacian obtained by convolving the image I with the Laplacian mask popularly used for edge detection.

$$\mathrm{F}_{\mathrm{HR}} = \max(k|\ \mathrm{H} > 0) - \min(k|\ \mathrm{H} > 0) \tag{6}$$

$$\mathrm{F}_{\mathrm{B}} = \sum_{(i,j)\in\Omega(x,y)} \left|\mathrm{I}(i,\,j)-\mathrm{I}(i+2\,j)\right|^2 \tag{7}$$

All these individual measures are computed in a small neighborhood around the feature points detected across the focal stack.

3.2. Window Size

The size of the window about which the focus measure is computed, is a vital parameter in the SFF method. Generally, the window size must be as small as possible to obtain accurate results. When the size of the window is large, a large neighbourhood is included to compute the focus. If the depth of the scene corresponding to different points in the window varies, it may lead to averaging of different focus levels caused by different depths of the scene points. Many authors have suggested a smaller sized window, particularly a 5 × 5 window, to be optimal. Window sizes lower than 5 × 5 may result in errors caused by random noises in grey levels. Larger mask sizes would result in averaging errors. In the current research, since the images suffer from magnification changes, a variable window size approach is developed. According to this method, the window size applied to a particular frame is scaled by the magnification factor corresponding to that frame. This means that in the current scenario, the window sizes would be reducing, starting from the first frame to the fifteenth image in the focal stack. Larger window sizes offer better results, but lead to averaging errors. It is justified, since the current work which uses the SFF-inspired algorithm only to obtain a sparse and coarse depth estimate. This issue may be considered as a drawback of the proposed method, as it inherently suffers from slightly higher averaging errors compared to the conventional SFF. The window size of the first image is chosen to be 15 × 15. When the window size is scaled by the magnification factor which are real numbers, it results in real values which are computationally not possible to be executed on a discrete domain; *i.e.*, the image's spatial domain. Hence the real numbers of the scaled window sizes are rounded off to the nearest integer.

3.3. Sparse Reconstruction

The following section of the paper presents the various results obtained from the application of the SFF-Inspired algorithm. Many different scene geometries were used to obtain their sparse and coarse depth estimate from the proposed SFF-inspired algorithm, though only one scene geometry are reported in this paper. Fig. 3 shows the images in the first, eighth and fifteenth step in the focal stack of scene.

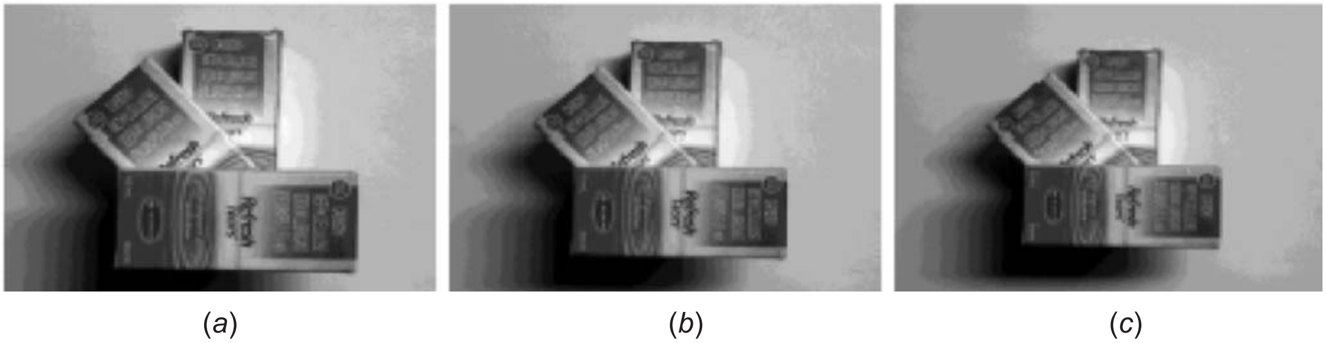

Figure 3: Image of Scene 1 (*a*) First Step (*b*) Eighth Step (*c*) Fifteenth Step

It may be observed, that both the focus and magnification change as a function of the depth for the various points in the image. Fig. 4 shows the focus levels of the images in the focal stack, computed by the same focus measure used for depth estimation.

It must be observed, that the focus measure falls after the fourth step. This behavior is caused due to two main factors. The first factor is that the focus levels of the majority of the pixels in the image are decreasing. The second understanding is that, since the field of view of the image is increasing for every step, more regions corresponding to the background become part of the images. As the background does not have texture variation, the focus measure will have small value for those regions. It must be noted that all the focus measures either directly or indirectly measure the difference in the pixel value. With all this information given, the usage of the overall focus measure of the images in the focal stack, to identify the plane of the best focus. Such an information is required when the proposed SFF-inspired approach is applied to determine the baseline of a stereo vision system. A stereo vision system with ideal baseline for a given scene geometry, may return dense and accurate results. Fig. 5 shows the common points in the stack overlaid on the image corresponding to the first step. It may be observed, that some feature points are identified in the junction between two different heights, result in averaging errors. Fig. 6 shows the frame number at which the focus measure attains the maximum value for all the feature points present across the focal stack.

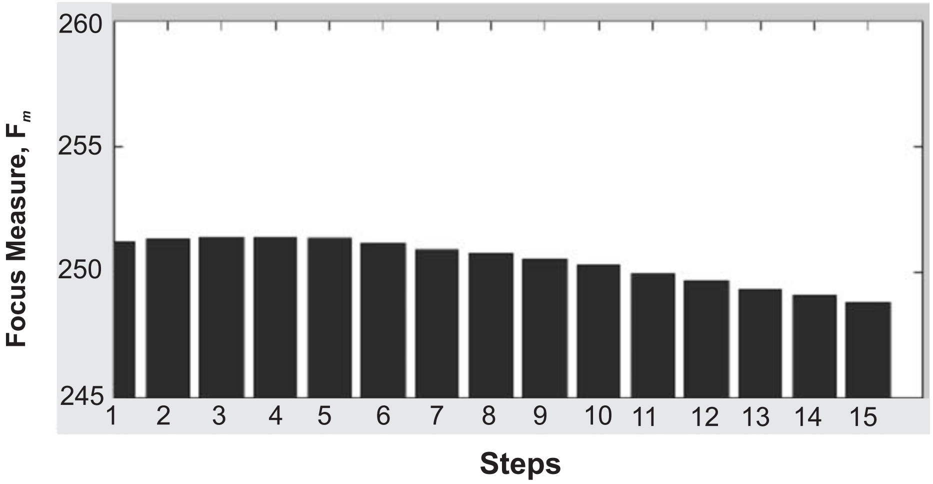

Figure 4: Focal Level of Images

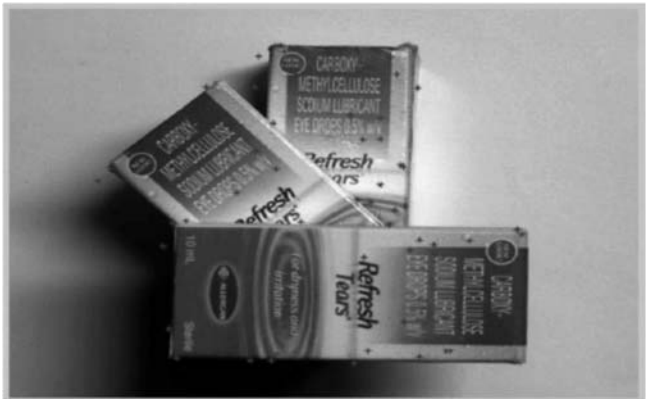

Figure 5: Common Points Overlay on the First Image of the Stack

Fig. 6 also shows the height estimate of Scene 1 obtained based on the information of the frame for which the maximum value of focus measure is obtained, which is displayed in the first part of the figure.

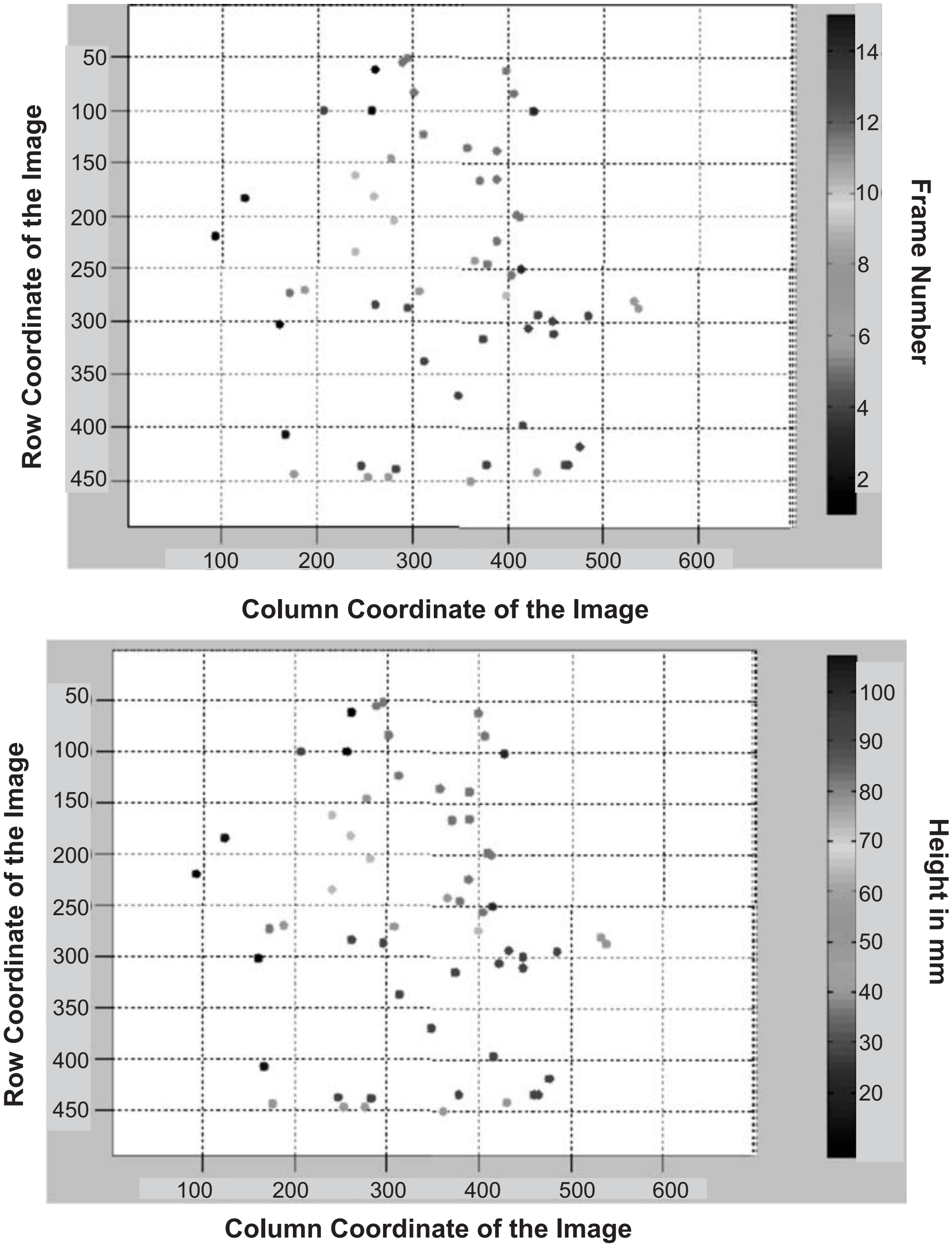

Figure 6: Frame Number of the Focus Function Maximization and Height of Scene Points

The frame number and the height estimate are encoded in a colour scale, as displayed in the figures. The averaging effects may be observed for pixels in the vicinity of the transition between the two different depths. The accuracy and density of reconstruction obtained from the SFF-inspired algorithm is definitely not comparable to the conventional SFF route. But the conventional SFF cannot reconstruct scene of the dimensionalities considered in the current research. The results published in [12] for the error in reconstruction using a variant of SFF method was reported to be around 0.37 step, where a lens step is 0.03 mm. The error for the conventional SFF is reported to be 0.40 step. The error in reconstruction in the current study was not more than 2 steps but here a lens step is around 6.67 mm. The error in reconstruction is different at different regions of the field of view, which is due to the averaging of different focus values arising due to different depths. These results are reasonable, considering the camera motion range and the size of objects under consideration.

4. BASELINE ESTIMATION AND SCENE RECONSTRUCTION

The sparse depth estimate obtained from the focus cue may be used as the preliminary information for estimating the baseline of the stereo vision system. This forms the central theme of the reported work. Such an approach is possible, since the relative pose between the camera and the scene may be altered both in the experimental setup, and in the eye-in-hand configuration considered in the simulation environment. The baseline is estimated using the following relationship [13]:

$$\Delta z = \frac{{}^{C}z^{2}}{f\,b}\Delta m \tag{8}$$

which may be rewritten as:

$$b = \frac{{}^{C}z^{2}}{f\,\Delta z}\Delta m \tag{9}$$

The equation relates the depth of the scene ${}^{C}z$, the associated distance resolution Δz, baseline of the stereo geometry b, the lens focal length f, and the error in disparity measurement Δm. The prefix superscript for z indicates that the depth is obtained with reference to the camera reference frame. This form of the equation allows estimating the baseline of the stereo vision system for a desired distance resolution. Since the scenes experimented with, have varying depth values, one unique value must be chosen by the algorithm. In the scope of the current study, the region with the least depth is chosen as the candidate information. This is because, assuming some prior information about the geometry of the scene may lead to loss of generality. The average value of ^{C}z obtained for scene 1 presented is 416 mm. In this research, the value of Δm is assumed to be 1 μm. This value is chosen, based on the values suggested in many literatures for similar specifications of the camera used [13]. The focal length of the lens used for all the trials of stereo vision is 16 mm lens. It may be observed that this expression allows choosing the Δz depending up on the need for an application. The distance resolution is the smallest measurable change in the distance of the origin of the camera coordinate frame from a scene point. A small value means a better resolution. This is the important advantage of the method proposed, as this allows the knowledge of the distance resolution, and also it can be preserved at a desired value invariant to the geometries dealt with. Moreover, the algorithm identifies the baseline based on the given situation automatically, without the need for the user defining one. This may act as a means of incorporating intelligence into systems that manipulate the environment such as robots.

4.1. Image Acquisition

Based on the baseline estimated from Equation (9) for different scene geometries, the images of the scene are acquired by moving the object suitably. The stereo algorithm actually assumes that the camera is moving, and not the object. As mentioned earlier, this assumption is valid when the motion is known, and moreover, the relative motion between the camera and the scene has the same effect on the images from which all the measurements are made. The selection of the distance resolution to be substituted in Equation (9) must be done properly. The value depends on many factors, as explained in the following section. From Equation (8) it may first be observed, that a higher distance resolution may be obtained by choosing a large baseline, but this value of the baseline must be realistic, since there are many problems associated with wide baseline stereo. When the baseline is too large the images would become too different, causing problems while attempting to match the corresponding features in the stereo image pair. This is the problem of occlusion in stereo vision, where features in the left image are not present in the right image and vice versa. This problem is generally solved by camera convergence. If the camera pose is converging towards the object of interest, the occlusion problem may be handled to some extent. In the current research, the camera vergence is not considered, since there is no provision for realizing changes in the relative orientation between the camera and the scene. Hence, the values of the distance resolution must be chosen in the realistic sense by considering the hardware restrictions. Even with this constraint, the generality of the algorithm is intact, since in a real world scenario such as in a real robot with an eye-in-hand camera, vergence may be easily achieved. For the trials carried out as part of the current study, the distance resolution is chosen to be equal to 0.2 mm. A common requirement for distance resolution comes from the application for which stereo vision is used. In the case of vision guided robots, the positioning repeatability may be used as a guideline for the selection of the distance resolution. This is because even if the distance resolution is better than the positioning repeatability of a robotic manipulator, which uses the information from the vision system, the robot may not be able to achieve that resolution. The value of 0.2 mm is quite better than the positioning repeatability of many commercially available industrial robots; hence it is justified for the actual application for which the proposed method is intended. On substitution of the respective values in Equation (8), the baseline was calculated to be approximately equal to 54 mm for the scene reported in this paper. It must be noted that the actual location of the camera is based on the focus level of the stack of images as shown in Fig. 3. As mentioned earlier, the computer vision algorithm assumes camera motion, instead of the motion of the object that actually takes place. Hence, the external camera parameters obtained from the camera calibration process for the initial camera location, are suitably altered for the two poses. The stereo images were taken from this distance of separation between the camera poses by moving the object for a distance of b/2 in the left direction from the current or initial position. Once the left image is acquired, the object is moved to a distance of b in the right direction with respect to the viewpoint of the camera. This is because the object is initially so located, that it is approximately in the centre of the field of view, during the series of image acquisition for the SFF-inspired algorithm. Since the object is only approximately close to the centre of the image, the left and the right stereo image will not look symmetrical with respect to the distance of the objects in the image from the corners of the images. This may be observed from the images of the various objects considered for trials. The lateral dimensions of the objects are chosen, such that the baseline distance will not cause the objects to move out of the field of view. This assumption is again attributed to the absence of the camera convergence facility in the experimental setup, as it is not considered within the scope of this study. The positioning accuracy of the linear slide base is around ±2 mm. The stereo image pair of the considered scene acquired from the baseline is shown in Fig. 7.

4.2. Scene Reconstruction

Once the stereo images are obtained, the rest of the process is part of the standard approach for triangulating a point based on stereo cues, as presented in many computer vision literatures. The first step in stereo vision is calibrating the camera, which in this research is pre-calibrated, and the model is regularly updated with new external camera parameters based on the relative motion between the camera and the scene. The second step is obtaining the set of corresponding features in the stereo image pairs. The correspondence algorithms may be broadly classified, as correlation-based and feature-based methods. In correlation based methods, the elements to match are image windows of fixed size, and the similarity criterion is a measure of correlation between the windows in the two images.

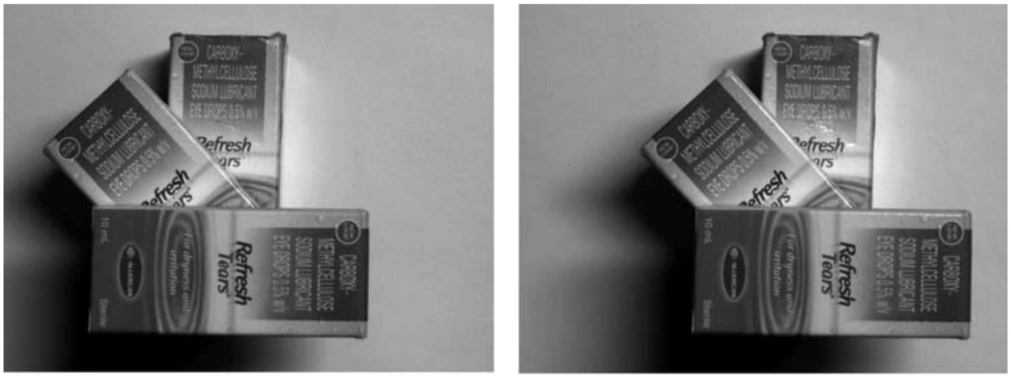

Figure 7: Stereo Image Pair of the Scene

The corresponding element is given by the window that maximizes the similarity criterion within a search region. Feature based methods restrict the search of features to a sparse set of features instead of image windows. They use numerical and symbolic properties of features, available from feature descriptors such as SURF, MSER, etc. In this study, a feature based approach for finding corresponding points is opted, due to its robustness, and the SURF features are used for the same. The SURF features are extracted and matched to find the corresponding points. Fig. 7 shows the corresponding features obtained in the stereo image pair. The image is basically a false colour overlay of the left image over the right image.

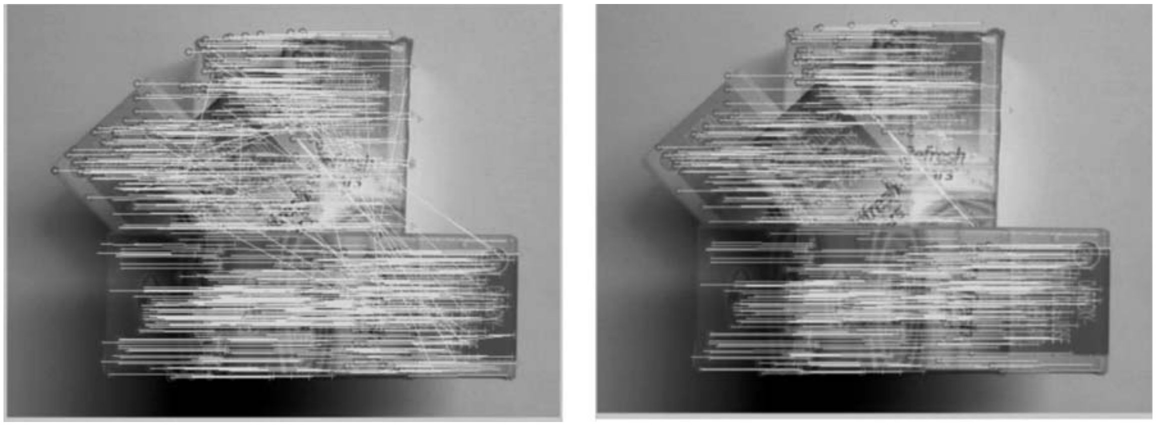

Figure 8: Point Correspondences Before and After Elimination of Outliers

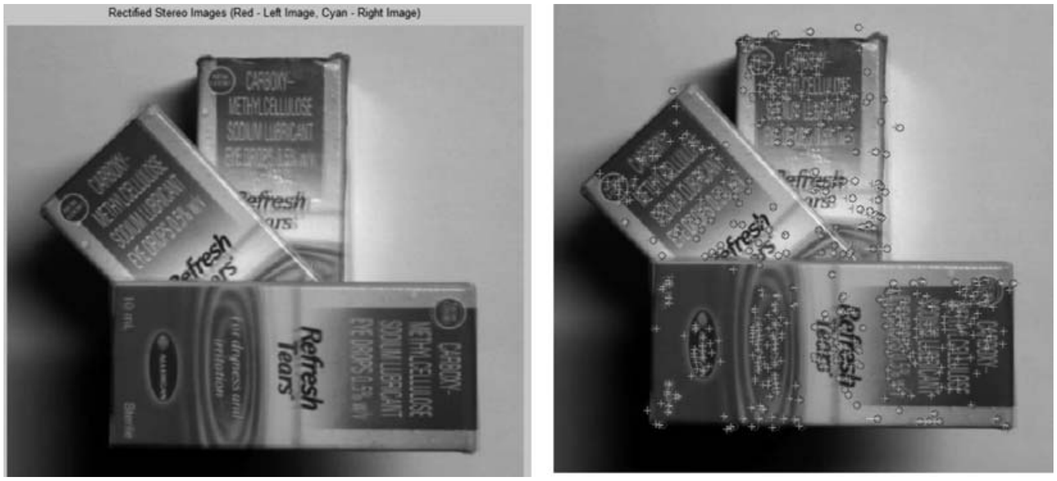
Rectified Stereo Images (Red - Left Image, Cyan - Right Image)

Figure 9: Rectified Stereo Image Pair and Inlier Points in the Stereo Image Pair

Similar to most feature based methods, the number of inliers is narrowed down by enforcing the epipolar constraint on feasible matches. It may be observed from Fig. 7, that there are many false matches, and it is for this purpose that a constraint is applied to eliminate the outliers. In the current study, the geometrical constraint based on the fundamental matrix is used. The fundamental matrices of the images are computed based on the corresponding points, and the false matches (outliers) were eliminated suitably. The second image in Fig. 7 shows the inlier points obtained after the application of the constraint. Once the correct correspondences are obtained, the stereo images are rectified to reduce the dimensionality of the search problem involved in disparity computation. Fig. 8 shows the false colour overlay of the right and left rectified images and the epipolar inliers plotted over the respective images. Since rectification involves image interpolation, some locations would contain artifacts due to zero padding. Such zero padded regions are manually eliminated in the figure.

From the figure it may be observed, that almost all the inliers are part of the rich texture, and only such points may lead to correct correspondence. Any regions with a repetitive texture or no texture would lead to ambiguous results. Once the rectified images are obtained, all the points in the left image have a unique row to search for. On establishing the correspondences for all the pixels, the disparity map for the entire image is obtained. The disparity map is basically the difference in the pixel locations of the points in the left image, and the corresponding points in the right image. Fig. 9 shows the disparity map of the stereo image pair of Scene 1 and Scene 2.

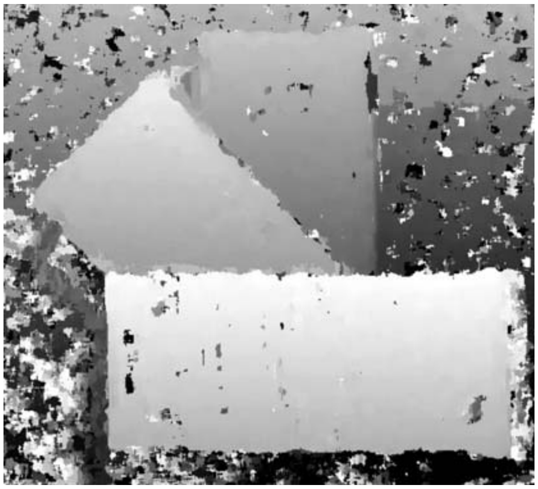

Figure 10: Disparity Map

After obtaining the disparity map the process of reconstruction is straight forward from the relation between the depth with focal length, baseline and disparity. This is possible, since the baseline of the stereo geometry, the external and the internal camera parameters are known in this research. This way of stereo vision based reconstruction is the unambiguous case, generally termed as metric or absolute reconstruction. The metric reconstructed information is shown in Fig. 10.

The depth values are encoded in the color scale adjacent to the plot. Minute variations in depth is also captured as may be observed in the figures in the regions of the stacked objects. It may also be observed from the figure, that regions around the stacked object show noisy information of depth. This is due to the lack of any visual business. As already mentioned, for performing stereo vision a rich, distinguishable texture is mandatory. The problems of erroneous depth estimate in texture-less regions and occlusion are not within the scope of the present research. Many researchers have presented methods to tackle these problems, which may be found in stereo vision literatures.

In order to demonstrate that the selection of a proper baseline has resulted in the desired distance resolution, the reconstructed images are subjected to some analysis. The reconstructed depth maps are again analyzed for the actual distance resolution, considering the known baseline, by substituting them in Equation (8). Since the scene has varying levels of depth, the depth of the top most object in the stack is used in the current study for the purpose of analysis. It must be noted, that even while finding the baseline from Equation (9), the depth of the top most object is used. This is again due to the same reason that no other information is possible to be automatically extracted, since no prior information about the scene is known. Hence, the region of the least depth from the camera is chosen as ${}^{C}z$. The average values of ${}^{C}z$ obtained from the reconstructed information for the region of least depth is 408 mm. It must be noted that the difference between the depth estimate obtained from the focus cue and stereo vision is 8 mm. Hence, though the magnitude is reasonably large, the application is not affected by the magnitude. The intention of presenting the depth error is only to highlight the differences to identify the factors that contribute to the error. On substituting these values in Equation (8), the distance resolutions obtained is 0.192 mm. It must be noticed that the desired values of the distance resolution is set at 0.2 mm. The absolute error in distance resolution is 0.008 mm. The chief contributing elements for the error may be one or more factors from the following list.

- The resolution error in the sparse estimate obtained from the focus cue.
- Errors in camera calibration.
- Positioning errors in the linear slide base.

These numbers are satisfactory enough for the level of information required for a robotic manipulator, though for some other applications, such as the estimation of accurate geometrical information of the scene for reverse engineering, it may not be satisfactory enough.

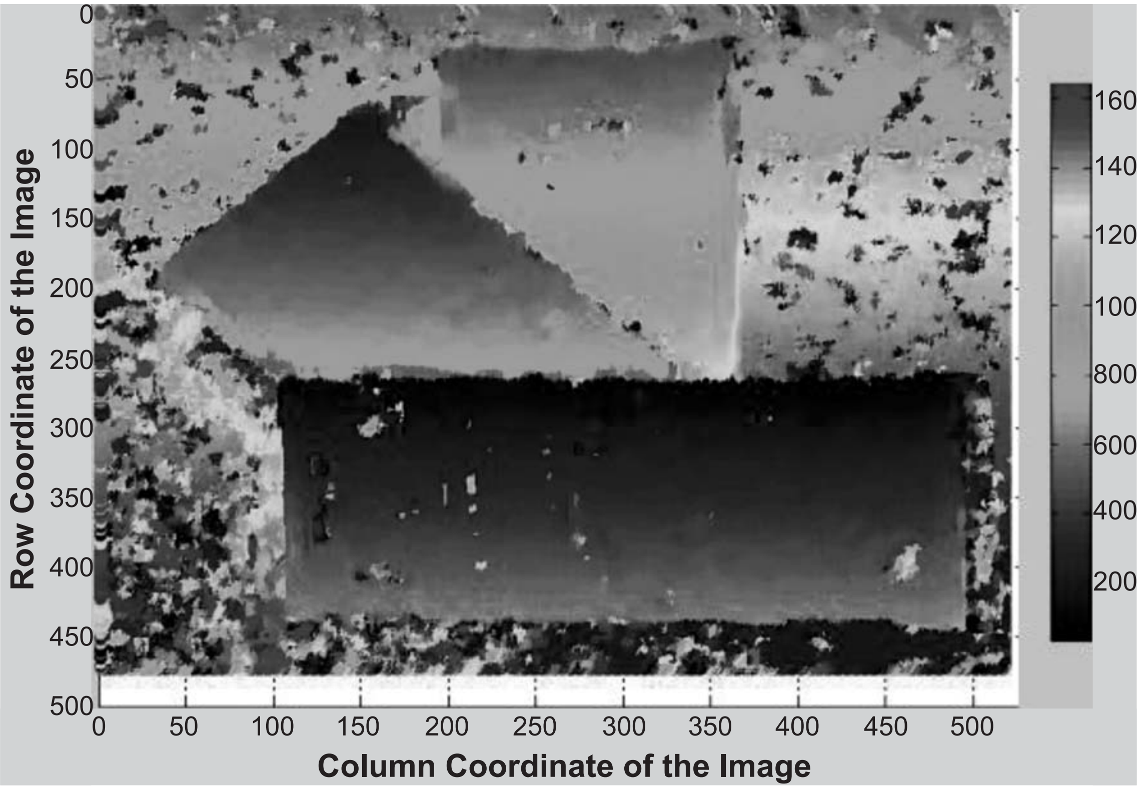

Figure 11: Metric Reconstructed Scene

5. REFERENCES

1. S.K. Nayar, Y. Nakagawa, “Shape from Focus: An Effective Approach for Rough Surfaces”, CRA90, 1990, pp. 218-225.
2. Y. Xiong, S.A. Shafer, “Depth from focusing and defocusing”, IEEE Computer Vision Pattern Recognition, 1993, pp. 68-73.
3. Rajiv Ranjan Sahay and A. N. Rajagopalan, “Dealing with Parallax In Shape-From-Focus”, IEEE Transactions on Image Processing, vol. 20, no. 2, 2011, pp. 558-569.
4. Senthilnathan, R., and R. Sivaramakrishnan. “Estimation of relative depth in the scene using SFF-inspired focus cue”, IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2012.

5. R. Senthilnathan, P. Subhasree and R. Sivaramakrishnan, “Performance Analysis of Gradient-Based Focus Measures in a Parallax Affected SFF Scenario”, International Journal of Computer Aided Manufacturing, vol. 1, no. 1, 2015, pp. 1-12

6. R. Senthilnathan, P. Subhasree, R. Sivaramakrishnan, C.R. Srinivasan, R. Srividhya, “Performance Analysis of Focus Measures in a SFF-Inspired Approach for Sparse Scene Reconstruction”, International Journal of Control Theory and Applications, vol. 8, no. 3, 2015, pp. 1153-1160.

7. C.R. Srinivasan, R Senthilnathan, P. Subhasree, R. Sivaramakrishnan, R. Srividhya, “Evaluation of statistical focus measures in a parallax affected SFF-inspired approach”, International Journal of Control Theory and Applications, vol. 8, no. 3, 2015, pp. 847-855.

8. H. Bay, T. Tuytelaars and L. Van Gool, “SURF: Speeded up robust features”, Proceedings of European Conference on Computer Vision, 2006, pp. 404-417.

9. Y. Sun, S. Duthaler, and B. J. Nelson, “Autofocusing in computer microscopy: selecting the optimal focus algorithm” Microscopy Research and Technique, vol. 65, 2004, pp. 139-149.

10. M.B.Ahmad and T.S. Choi, “A heuristic approach for finding best focused shape”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 4, 2007, pp. 566–574.

11. L. Firestone, K. Cook, K. Culp, N. Talsania and K. Preston Jr., “Comparison of autofocus methods for automated microscopy”, Cytometry, vol. 12, 1991, pp.195-206.

12. M. Subbarao and T.S. Choi, “Accurate recovery of three dimensional shape from image focus”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 3, 1995, pp. 266–274.

13. D. Gallup, J.M. Frahm, P. Mordohai and M. Pollefeys, “IEEE Conference on Computer Vision and Pattern Recognition”, 2008, pp. 1-8.