STEREO VISION AUTO-ALIGNMENT AND THE UNSUPERVISED SEARCH FOR OBJECTS OF INTEREST WITH DEPTH ESTIMATION

Ling-Wei Lee
Neuramatix Sdn Bhd
No. 27-9 Level 9, Signature Office,
Bandar Mid-Valley,
59200 Kuala Lumpur, Malaysia.

Faeznor Diana binti Zainordin
Faculty of Computer Science and
Information Technology,
University of Malaya, Lembah Pantai,
50603 Kuala Lumpur, Malaysia.

KEYWORDS
Stereo vision alignment, objects segmentation, depth mapping, depth estimation.

ABSTRACT

Stereo vision is fast becoming a highly investigated area in the domain of image processing. Depth information may be obtained from stereo or multi-vision images for reconstructing objects in 3D based on 2D information. Robotic applications make use of stereo vision for navigation purposes, locking down targets, as well as simulating human-like behaviour. This paper presents an algorithm for the auto-alignment of stereo images followed by the self-extraction of objects of interest using an unsupervised search. Based on the understanding that different objects or regions are focused at different focal points, alignment between the two images is carried out to determine areas of high overlapping similarities. Objects are then identified in these selected areas with their estimated depth calculated based on the disparities between the stereo images. Results obtained for tests carried out on several experimental image pairs showed good extraction of the objects with close-to-real-world values obtained for the distances of the objects from the cameras.

INTRODUCTION

Human vision is capable of discerning objects from a background and enables a person to easily pick up a targeted object from a group of other entities. The estimation of distance from a person's hand to the target is remarkably accurate and of great precision. This is made possible via a combination of human vision, depth perception as well as input from other sensory organs. Human vision is akin to images captured on a camera – rods and cones in the retina capture scene information which is processed and transmitted to the brain for understanding and analysis. This then leads to questions such as: Are we able to derive useful information from images? Can we extract regions of interest while discarding the rest? Is it possible to execute certain algorithms for gathering further information from the image – information which may not be immediately visible but which requires the application of filtering kernels or functions? Such questions lead to the development of various techniques and algorithms in the domain of image processing.

Early algorithms placed emphasis on obtaining data from single images, and methods have been constantly reviewed and improved over the years thereby producing efficient solutions that are easily implementable. A single image consists of 2-dimensional data in the form of (x, y) spatial coordinates and pixels information. Human vision is able to perceive the 3-dimensional attribute of objects, hence the ability to grasp objects with precision. However, when translated to a 2D image, such characteristics become lost. The question one faces is – are we able to reconstruct 3-dimensional attributes based on captured 2D images? A single image may not provide sufficient information but two or more images may provide ample amount of details for reconstructing the object based on disparity maps. Stereo vision alignment and processing thus becomes a highly researched area applicable to the field of robotics and vision-based navigational systems. However, it is a "computationally expensive process … the accuracy of the disparity (depth) depends strongly on the calibration and rectification" (Sutton and Green, 2010).

Given a set of stereo images, one may notice that different objects in the images get aligned at different offsets when one image is overlaid onto the other. An attempt is made here to determine the regions of maximum overlap followed by procedures for identifying objects of potential interest. These objects are consequently extracted from the stereo images. The second part of this study is to determine if the depth (distance of the object from the cameras) can be estimated to a certain degree of accuracy based on the disparity between the images. A single set of stereo images containing a single object is used for calculation of the reference ratios. Reviews of past studies and implementations are first given in Background, followed by descriptions of the method in The Algorithm. The results of the study are presented in its respective section before the paper concludes with Discussion and Conclusion.

Proceedings 27th European Conference on Modelling and Simulation ©ECMS Webjørn Rekdalsbakken, Robin T. Bye, Houxiang Zhang (Editors)
ISBN: 978-0-9564944-6-7 / ISBN: 978-0-9564944-7-4 (CD)

BACKGROUND

According to Weng et al (1992), "establishing correspondences between different perspective images of the same scene is one of the most challenging and critical steps in motion and scene analysis". One of the main difficulties lies in determining a suitable algorithm for matching attributes of the images involved. Objects may become occluded in one scene with discontinuities detected in the images. There is the challenge of segmenting meaningful regions or objects as well. Early methods employ the use of edge detection and matching for aligning stereo vision (Alwan and Naji, 1996). A method based on the combination of an improved isotropic edge detector and a fast entropic thresholding was developed for obtaining colour edges in input images (Fan et al, 2001). The centroids of all adjacent edge regions act as initial seeds for region growing, thus producing homogenous areas with closed boundaries.

Benera and Prokop (2012) proposed an image segmentation technique based on standard inter-pixel Euclidean distance enhanced by similarities of hues with limits imposed on the size of segments. The method delivered good performance and was shown to be robust. Due to the use of stereo images, algorithms have been designed to carry out stereo matching and object segmentation simultaneously. Bleyer et al (2011) presented an approach wherein a 3D scene consists of visually distinct and spatially coherent objects. Each of the objects may be represented in terms of its associated colour model, a novel 3D connectivity property, and a 3D plane which estimates the disparity distribution. The method is capable of retrieving depths of regions that are fully occluded in one of the stereo views, and is designed as an energy function.

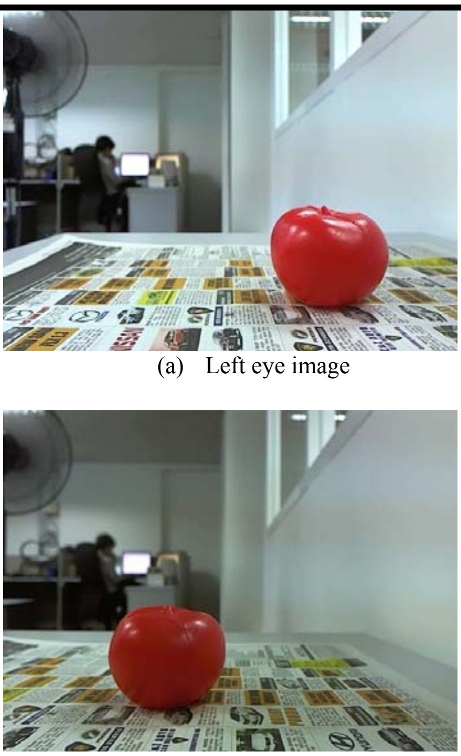
A stereo matching approach based on image segments was proposed by Kamencay et al (2012). The method is a hybrid segmentation algorithm using the Belief Propagation and Mean Shift algorithms for refining disparities and depth maps. Region segmentation is first carried out on the left image followed by a local window-based matching to estimate the disparities for the pixels. The results showed that the final depth map may be generated via application of segment disparities to the input images. In another work by Bleyer et al (2012) the authors combined an unsupervised object extraction method for a single image together with depth estimation from stereo images. 3D scene-consistency is emphasized with a series of "plausible object hypotheses" obtained that can be used as input for object recognition systems.

Depth estimation based on stereo images has been widely investigated. Kytö et al (2011) proposed a method for evaluating the accuracy of the derived depth in human centred applications. A multilevel test target was used and two crucial parameters – the focal length and baseline, were explored using different values. It was reported that the focal length is of larger influence to the accuracy of the estimated depth compared to the baseline. Lin and Setiawan (2008) attempted to identify the orientation of an object in space by using stereo cameras based on the Scale Invariant Feature Transform (SIFT) and Support Vector Machine (SVM). It was found that having an affine transform matrix was shown to outperform features matching in recognizing object orientations. Another method employing the SIFT feature extraction was presented by Lam et al (2009). The authors implemented virtual stereopsis in 3D modeling of human bodies without the use of multiple calibrated cameras. The position of virtual cameras were calculated using SIFT and motion estimation. In a similar study, Kouskouridas and Gasteratos (2011) proposed a novel implementation which used spatial information and a multi-camera system to estimate the location of objects in 3D space. The method was claimed to be simple and computationally efficient.

In the context of mobile robots navigation, studies have been carried out to obtain obstacle information using an omnidirectional stereo vision system (Su et al, 2006). By using triangulation, the 3D coordinates of a point are generated based on the given stereo images. Such systems have become essential features of autonomous mobile robots. Ben-Tzvi and Xu (2010) presented an embedded stereo vision system which catered to flexible baselines for use in compact-sized robots. In an earlier work (Oh and Lee, 2007), a general-purpose system was proposed for tasks associated with industrial robots. Functions like camera calibration, pattern registration and training etc were incorporated into the system.

Depth maps may be generated using just stereo images by applying a series of processes. A novel depth map generation approach using K-means clustering was proposed to classify objects into foreground or background entities (Meng and Jiang, 2012). Tong et al (2010) presented an object-oriented stereo matching algorithm using multi-scale superpixels for the generation of low-resolution depth maps. The approach was able to overcome downsampling-associated disadvantages such as merging of foreground objects to background, edge blurring etc. When tested on the Middlebury test-bed, the method was shown to outperform other low-resolutions approaches. Given sufficient information, 3D versions of the objects could be reconstructed using the stereo images. Ikeda (2005) employed a combination of "photometric stereo with colour segmentation and the binocular stereopsis to reconstruct accurate shapes from two colour images".

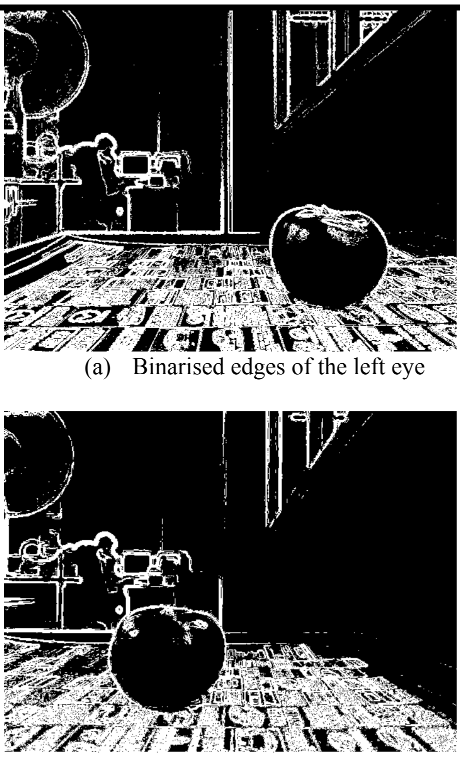
Chen et al (2012) employed the use of Markov Random Fields (MRF) to solve labeling problems in the domain of computer vision. In segmenting regions of interest the MRF model has proved to be robust against real scene complexities as well as noise corruption. In view of the many attempts made in aligning stereo vision



(a) Left eye image

(b) Right eye image

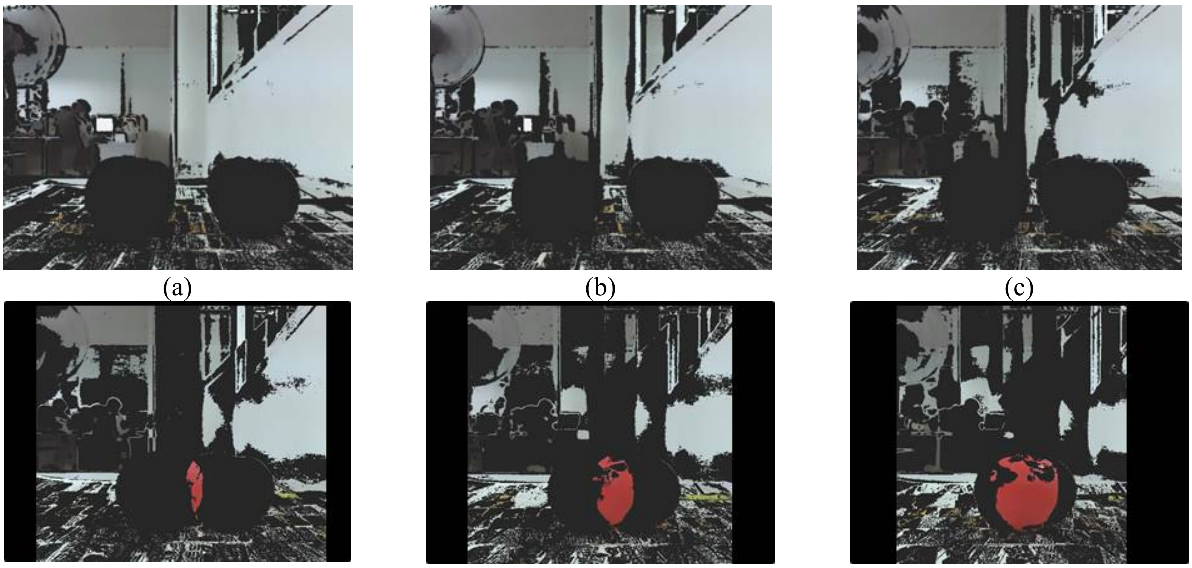
Figure 1. (a) The captured image from the left camera. (b) The captured image from the right camera.



(a) Binarised edges of the left eye

(b) Binarised edges of the right eye

Figure 2. (a) The binarised edges for the left image. (b) The binarised edges for the right image.



(a) (b) (c)

(d) (e) (f)

Figure 3. The alignment process, with the right image overlaid onto the left and shifting the right image to the right. (a) Offset 10. (b) Offset 20. (c) Offset 50. (d) Offset 100. (e) Offset 150. (f) Offset 200. The images show the gradual focusing of the object in the centre.

images and the algorithms developed for extracting objects in recent years, it can be said that stereo vision research is fast gaining interest. This paper proposes a method for the auto-aligning of input images followed by an unsupervised search for objects ofinterest within the regions of different focal points.

THE ALGORITHM

Stereo images for this study are first obtained using two moderately-low resolution cameras arranged in parallel. The images are of dimensions 640x480. A test pair of left and right images is shown in Figure 1.

The images display shadows and regions of lower intensities. In order to increase the contrast, a brightness adjustment was carried out. This was to maximize the results obtained from using an edge detection method which was duly applied. Robert's Edge Detection was selected as the algorithm of choice due to the lower amount of noise produced in the images. A thresholding technique was then used to binarize the images thus segregating the edges and the background into separate groups. The outcome is given in Figure 2.

The next step was to align the stereo images for discovering regions of high correspondences. However, the use of only edges provided insufficient pixel information to determine matches. As such, alignments were carried out based on the brightness-adjusted images. The process is given in Figure 3. Matches were determined by checking the distances of all colour channels for each pixel in the left image to the overlayed pixel in the right image. The RGB values of a pixel were compared to the corresponding pixel values in the other image. A difference of ≤ 30 between the pixels of both images for all Red, Green and Blue channels resulted in the shortlisting of the pixel as a matching point. The right image was overlaid onto the left and shifted to the right pixel-by-pixel, and the total number of matching pixels found at each pixel shift was tabulated. For the above case, the maximum overlaps were found at offsets of 10 pixels and 230 pixels. By using these offset values, the edge-detected images from Figure 2 were then aligned. Figure 4 gives the aligned outputs at these positions.

From Figure 4 the regions of high correspondence were shown as black areas. At offset 10 the background had the highest matches between the left and right images while at offset 230 the object in the foreground was shown to be well correlated. The task now was to determine which region to extract from the alignments at each offset. A simple flood-fill algorithm was used to explore all possible regions, and the three highest-filled regions were automatically selected for the next step of processing. Examples of the flood-fill exploration are given in Figure 5.

By referring to the filled regions, the pixels were consequently replaced with all colour-matching pixels (Figure 6). For example, Figure 5(a) shows a large filled area of the background. By referring to this filled region, pixels of similar colours between both the images were used to replace the region, thereby producing the patches of colours in Figure 6(a). This was due to the fact that not all pixels in the filled region were of similar values. Another check was performed to determine the area consisting of the highest count of colour-matching pixels after replacement, and this region was extracted and reconstructed based on the original images (Figure 7). The final step of the method was to determine the estimated distance of the object

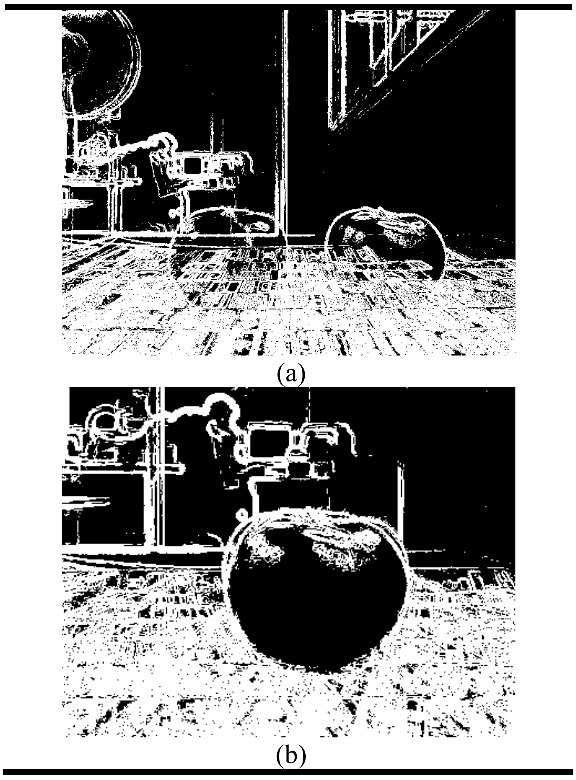


Figure 4. (a) Aligned output at offset 10. (b) Aligned output at offset 230 (image cropped).

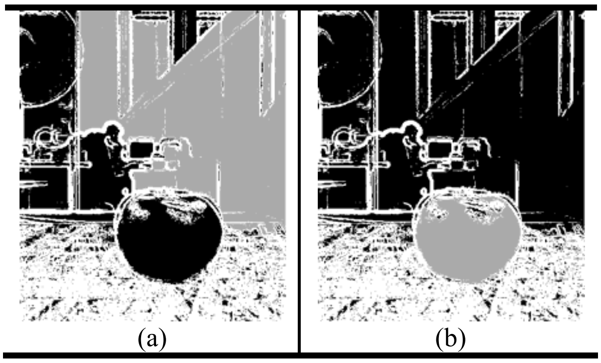


Figure 5. Alignment at offset 230. (a) Flood-fill of the background. (b) Flood-fill of the object.

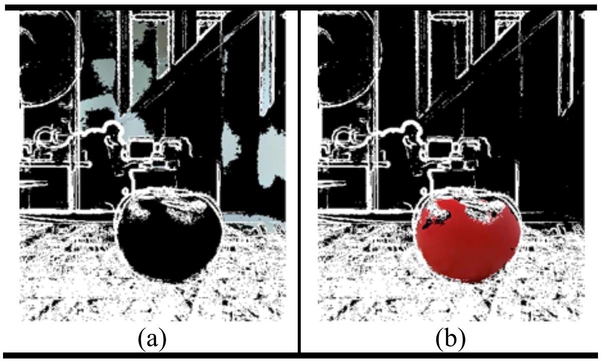


Figure 6. Alignment at offset 230. (a) Replacement of flood-filled region with colour-matching pixels between both the left and right images for the background as depicted in 5(a). (b) Replacement of flood-filled region with colour-matching pixels between both the left and right images for the object as depicted in 5(b). The object is picked over the background due to the high count of colour-matching pixels.

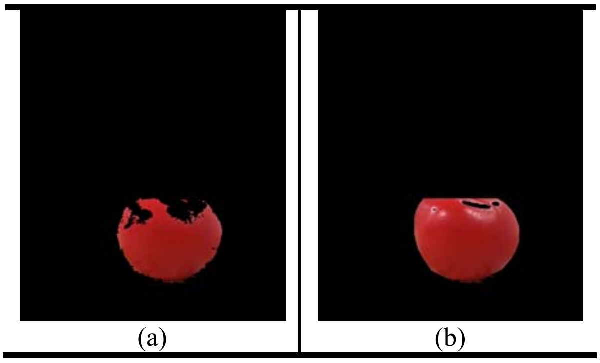


Figure 7. Alignment at offset 230. (a) Extracted object before reconstruction. (b) Extracted object after reconstruction using information from original image.

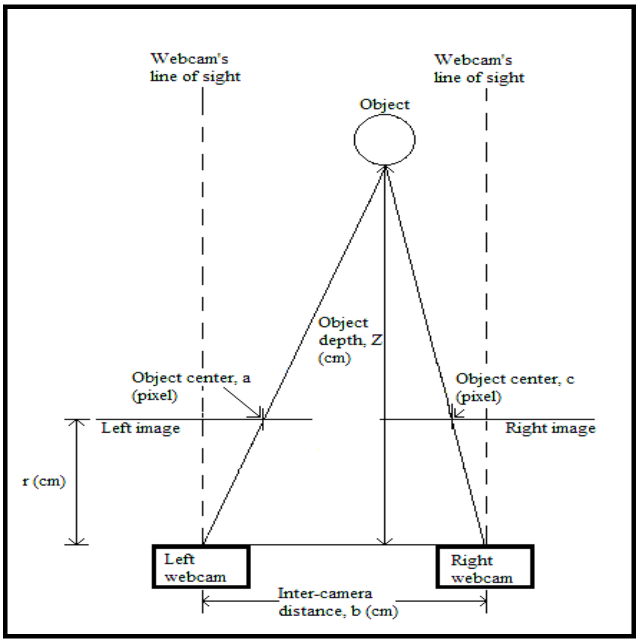


Figure 8. Set-up for obtaining the initial reference ratios for calculating the estimated distance of each identified object.

from the camera. Figure 8 shows the set-up for obtaining the initial measurements and reference ratios for calculation of the estimated distance. The formula is given by Equation (1).

$$Z = \frac{b\,r}{|(a-c)|\,p} \qquad (1)$$

$$p = \text{pixel width (cm)}$$

RESULTS

Experiments were carried out on several sets of images and the results are shown as follows. The left and right images were first shown followed by all extracted objects. The output showed that the objects of interest had been successfully extracted from the original images. The reconstruction process based on information from the input images produced a close-to-original representation of the objects.

Table 1 gives the estimated distances of the objects to the camera based on disparity values obtained from the stereo images. The estimated values are shown to be close to the actual depth, thus indicating the robustness of the algorithm in aligning and selection of objects.

DISCUSSION

The implementation has shown to be capable of auto-extracting objects without prior information from the user. The method takes as input a pair of stereo images and the algorithm proceeds to determine the regions of high matches followed by auto-selection of highly-correlated areas. Objects were shown to be successfully extracted and the reconstruction process produces depictions that are close to the original images.

An analysis of the estimated depth (distances of the objects to the camera) shows promising results. This is useful for robotic applications, especially when the objective of the robot is to identify an object, navigate towards the object and pick up or avoid the object. The method, however, takes approximately 10s to 40s on a Core 2 Duo PC with 2GB RAM to auto-align and identify the regions of interest, depending on the number of alignments obtained. This needs to be substantially improved to achieve real-time processing of stereo vision, and may benefit from using the generalised belief propagation (Chen and Wang, 2012) in which stereo matching was shown to achieve a highly significant speed-up using the reported approach.

The advantage of this method is the fact that it does not require prior knowledge about the objects or the environment to achieve the results. A future work is the application of this method to an actual drone or robot to simulate human-like vision and behaviour. Also included as part of the future work is the use of textured entities as objects of interest, which should pose higher challenges in the identification and extraction process. Research work is currently underway and it is hoped that a robot capable of self-navigating while learning about its environment and new objects can be attained. This work represents the initial studies for such vision-based robotic systems. Factors such as changes in environmental settings as well as the use of complex objects are included into consideration in upcoming experimental studies.

CONCLUSION

An algorithm for the auto-alignment of stereo images followed by the self-extraction of objects of interest using an unsupervised search is presented. The method has shown to be capable of shortlisting regions of high matches between the left and right images and to select regions of high correspondence as output.

Experiment Set I

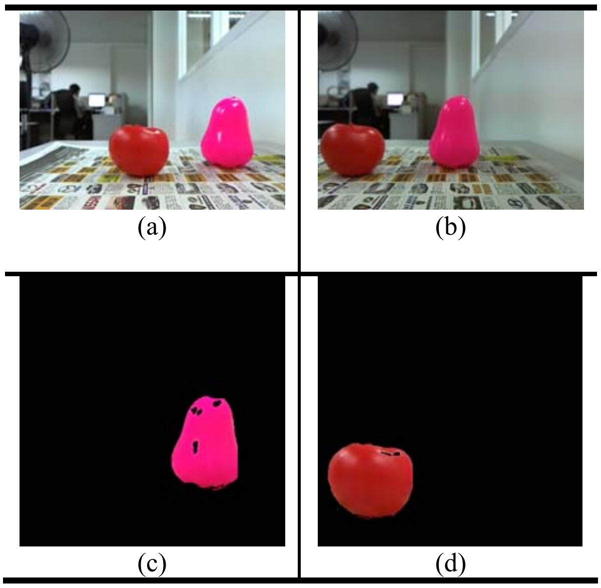


Figure 9. (a) Left image (b) Right image. (c) Extracted object 1. (d) Extracted object 2. Both objects have been reconstructed based on original images.

Experiment Set II

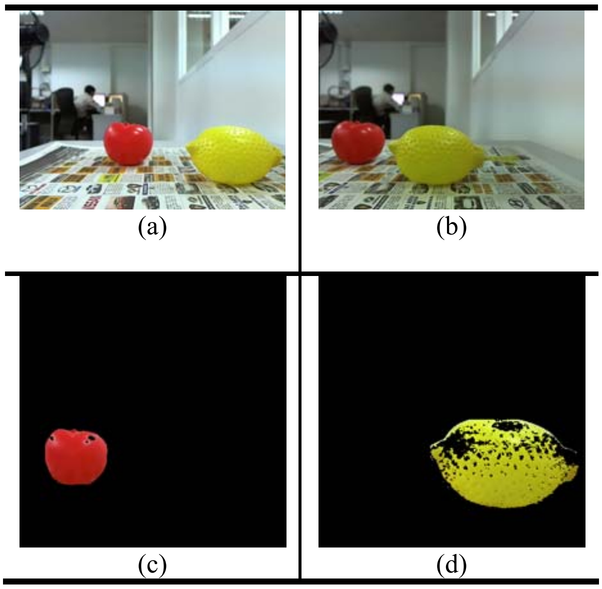


Figure 10. (a) Left image (b) Right image. (c) Extracted object 1. (d) Extracted object 2. Both objects have been reconstructed based on original images.

Experiment Set III

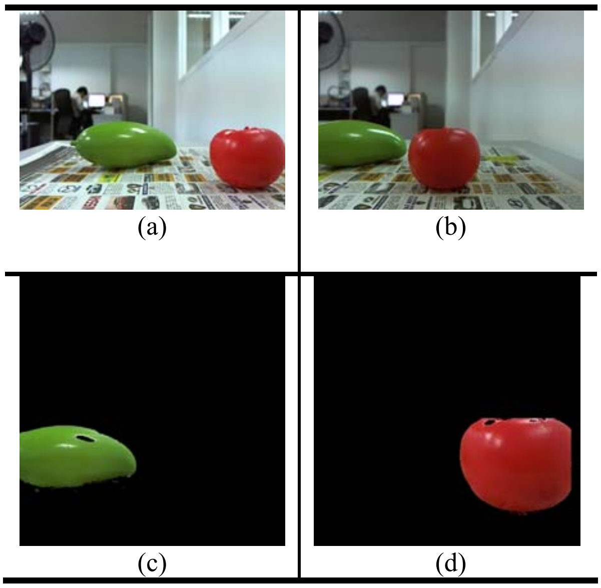


Figure 11. (a) Left image (b) Right image. (c) Extracted object 1. (d) Extracted object 2. Both objects have been reconstructed based on original images.

Experiment Set IV

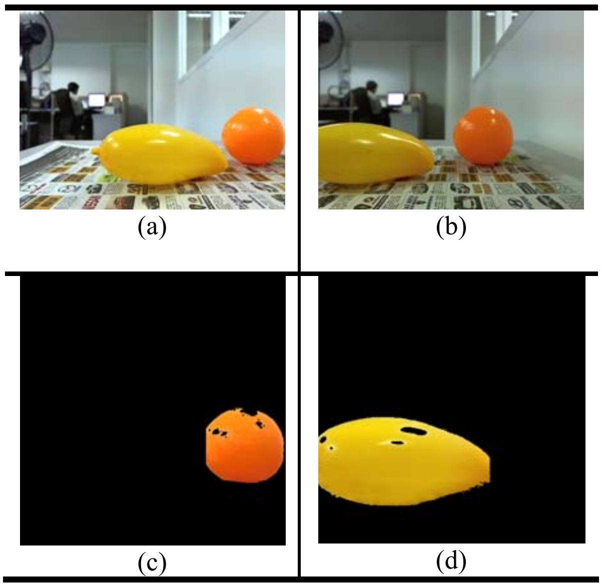


Figure 12. (a) Left image (b) Right image. (c) Extracted object 1. (d) Extracted object 2. Both objects have been reconstructed based on original images.

Table 1. Comparison of actual depth and estimated depth for extracted objects

Experiment	**Object**	**X point at left image (px)**	**X point at right image (px)**	**Disparity (px)**	**Actual Depth (cm)**	**Estimated Depth (cm)**
1	Tomato	286	87	199	18.40	22.64
	Water Apple	496	327	169	25.30	26.66
2	Tomato	261	102	159	26.50	28.34
	Lemon	514	285	229	18.50	19.68
3	Tomato	542	303	239	16.80	18.85
	Mango	272	103	169	26.05	26.66
4	Mango	348	139	209	18.65	21.56
	Orange	567	398	169	25.50	26.66

The extracted objects are free from background interference and retain characteristics from the original images using a reconstruction method. Analysis of the estimated depth proves that the algorithm has successfully aligned the stereo images with the correct depth of the objects. This study will be further extended to applications in real-life drones and robots for the simulation of human-like vision and behaviour.

REFERENCES

Alwan, R.H.; Naji, M.A. 1996. "Automatic Stereo Image Matching using Edge Detection Technique." *International Archives of Photogrammetry and Remote Sensing* XXXI, No. B3, pp. 29-35.

Bleyer, M.; Rother, C.; Kohli, P.; Scharstein, D.; Sinha, S. 2011. "Object Stereo – Joint Stereo Matching and Object Segmentation." In *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition* (Jun. 21-23) IEEE, Colorado Springs, USA, pp. 3081-3088.

Bleyer, M.; Rhemann, C.; Rother, C. 2012. "Extracting 3D Scene-consistent Object Proposals and Depth from Stereo Images." *Lecture Notes in Computer Science* 7576, pp. 467-481.

Chen, S.Y.; Tong, H; Cattani, C. 2012. "Markov Models for Image Labeling" *Mathematical Problems in Engineering* Vol. 2012, AID 814356, 18 pages., doi: 10.1155/2012/814356.

Chen, S.Y.; Wang, Z.J. 2012. "Acceleration Strategies in Generalized Belief Propogation." *IEEE Transactions on Industrial Informatics* 8, pp. 41-48.

Fan, J.; Yau, D.K.Y.; Elmagarmid, A.K.; Aref, W.G. 2001. "Automatic Image Segmentation by Integrating Color-Edge Extraction and Seeded Region Growing." *IEEE Transactions on Image Processing* 10, No. 10, pp. 1454-1466.

Ikeda, O. 2005. "Shape Reconstruction from Two Colour Images using Photometric Stereo Combined with Segmentation and Stereopsis." In *Proceedings of the 2005 International Conference on Video and Signal Based Surveillance* (Sep. 15-16), IEEE, Como, Italy, pp. 434-438.

Kamencay, P.; Breznan, M.; Jarina, R.; Lukac, P.; Zachariasova, M. 2012. "Improved Depth Map Estimation from Stereo Images Based on Hybrid Method." *Radioengineering* 21, No. 1, April, pp. 70-78.

Kouskouridas, R.; Gasteratos, A. 2011. "Location Assignment of Recognized Objects via a Multi-Camera System." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 4, No. 3 (Sep).

Kytö, M.; Nuutinen, M.; Oittinen, P. 2011. "Method for Measuring Stereo Camera Depth Accuracy based on Stereoscopic Vision." In *Proceedings of SPIE 7864, Three-Dimensional Imaging, Interaction and Measurement Conference*, 786401.

Lam, D.; Hong, R.Z.; DeSouza, G.N. 2009. "3D Human Modeling using Virtual Multi-View Stereopsis and Object-Camera Motion Estimation." In *Proceedings of the 2009 International Conference on Intelligent Robots and Systems* (Oct. 11-15) IEEE/RSJ, St. Louis, USA, pp. 4294-4299.

Lin, C-Y.; Setiawan, E. 2008. "Object Orientation Recognition Based on SIFT and SVM by Using Stereo Camera." In *Proceedings of the 2008 International Conference on Robotics and Biomimetics* (Feb. 21-26) IEEE, Bangkok, Thailand, pp. 1371-1376.

Meng, S.; Jiang, H. 2012. "A Novel Depth Map Generation Method Based on K-means Clustering." In *Proceedings of the Fourth Internation Conference on Digital Home* (Nov. 23-25), Guangzhou, China, pp. 28-32.

Oh, J.; Lee, C. 2007. "Development of a Stereo Vision System for Industrial Robots." In *Proceedings of the 2007 International Conference on Control, Automation and Systems* (Oct. 17-20), Seoul, Korea, pp. 659-663.

Su, L.; Luo, C.; Zhu, F. 2006. "Obtaining Obstacle Information by an Omnidirectional Stereo Vision System." In *Proceedings of the 2006 International Conference on Information Acquisition* (Aug. 20-23) IEEE, Shandong, China, pp. 48-52.

Sutton, D.; Green, R. 2010. "Evaluation of Real Time Stereo Vision System Using Web Cameras. "In *Proceedings of the 25th International Conference of Image and Vision Computing New Zealand* (Nov. 8-9) Queenstown, New Zealand, pp. 1-10.

Tong, H.; Liu, S.; Liu, N.; Barnes, N. 2010. "A Novel Object-Oriented Stereo Matching on Multi-scale Superpixels for Low-Resolution Depth Mapping." In *Proceedings of the 32nd Annual International Conference of the IEEE EMBS* (Aug. 31-Sep. 4) Buenos Aires, Argentina, pp. 5046-5049.

Weng, J.; Ahuja, N.; Huang, T.S. 1992. "Matching Two Perspective Views." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, No. 8 (Aug), pp. 806-825.

AUTHOR BIOGRAPHIES

LING-WEI LEE was born in Kuala Lumpur and studied at the University of Nottingham Malaysia Campus where she obtained her honours degree for Computer Science in 2007. She worked for about a year before returning to her alma mater to pursue her PhD with research focus in the field of systems biology. She is now currently working as a computational research scientist with Neuramatix. She can be reached at lingwei@neuramatix.com.

FAEZNOR DIANA BINTI ZAINORDIN was born in Kuala Lumpur, Malaysia and is currently pursuing her undergraduate studies in the field of Computer Science, majoring in Artificial Intelligence. She is attached to the Faculty of Computer Science and Information Technology, University of Malaya. She can be reached at sasuke_eno91@yahoo.com.