Multi-Image Matching for a General Motion Stereo Camera Model

Ja Seong Ku*, Kyoung Mu Lee**, and Sang Uk Lee*
*School of Electrical Engineering, Seoul National University
**Department of Electronics & Electrical Engineering, Hong-ik University
E-mail: jsku@sting.snu.ac.kr, kmlee@wow.hongik.ac.kr, sanguk@sting.snu.ac.kr

Abstract

Motion stereo is to extract the 3-D information of an object from images of a moving camera, using geometric relationships between corresponding points. This paper presents an accurate and robust motion stereo algorithm employing multiple images, taken under a general motion. The object functions for individual stereo pairs are represented, with respect to the distance, then these object functions are integrated, considering the position of cameras and the shape of the object functions. By integrating the general motion stereo images, we not only reduce the ambiguities in correspondence, but also improve the precision of reconstruction. Also by introducing an adaptive window technique, we can alleviate the effect of projective distortion in matching features and improve accuracy greatly. Experimental results on the synthetic and real data set are presented to demonstrate the performance of the proposed algorithm.

1. Introduction

The problem of extracting the 3-D information from images has been one of the most challenging and important problems in computer vision. Among the existing 3-D reconstruction techniques, shape from stereo, which extracts the 3-D information from multiple images using geometric constraints between corresponding features, has been investigated quite intensively by many researchers. Especially, the techniques using stereo images taken by a single moving camera, known as the motion stereo, have received a lot of attention, due to its practicability and accuracy. These techniques can be readily applied to target tracking, navigation, 3-D recognition and so on.

Two typical motion stereo methods are the so called lateral and axial motion stereo. The lateral motion stereo acquires images by moving a camera in the direction perpendicular to its optical axis, while the axial motion stereo moves in the parallel direction. It is known that the axial motion stereo has several advantages over the lateral motion stereo, such as the small search range for finding the corresponding features and the low probability of missing them [1]. However, it provides relatively poor estimate, compared with the lateral stereo, due to the large error over the region near the center of an image. Although the lateral and axial motion stereo methods are relatively easy to analyze, those methods are quite restrictive in practical applications, since it is not easy to maintain the camera motion perpendicular or parallel to the optical axis in many real situations.

There also have been many stereo techniques that use multiple images [2]-[6]. Most of these techniques find the corresponding points in each stereo image pair and then select the correct combinations for the corresponding points. These techniques are simply the extension of the conventional binocular stereo technique. In contrast, there have been techniques that use multiple images simultaneously to find the corresponding points. Tsai [2] proposed the algorithm that uses multiple images to find easily the sharper extrema of an object function for the corresponding points. Tsai positioned eight cameras at the specific points, called the potential conjugate points on multiple perspective views. Okutomi *et al.* [3] and Kanade [4] presented the method that uses multiple stereo pairs with various baselines, obtained by a lateral displacement of a camera. In this method, they represented the SSD(Sum of Squared Difference) for each stereo pair, with respect to the inverse distance, rather than the disparity. Then, the sum of the resulting SSD functions is used for an object function to find the corresponding points, demonstrating that the algorithm could remove ambiguity and improve accuracy.

But, notice that these two methods also assume the restrictive conditions on the camera position. Therefore, this paper presents an accurate and robust motion stereo algorithm based on multiple images,

0-8186-8821-1/98 $10.00

taken under a general motion which includes both the rotational and translational motion. By using multiple stereo images taken under a general motion in batch mode, we not only reduce ambiguities in correspondence, but also improve accuracy relating to the reconstruction significantly.

The object functions for individual stereo pairs are represented, with respect to the distance, and then these object functions are combined considering the position of cameras and the shape of the object functions. Also by introducing an adaptive window technique, we can alleviate the effect of projective distortion in matching features and improve accuracy greatly. In this paper, it is assumed that the internal and external parameters of cameras are already available.

2. Multi-image matching for a general motion stereo

Similar to the method proposed by Okutomi *et al.* [3], we first represent the SSD, the object function, with respect to the distance, so that the results of each stereo image pair can be integrated in a single framework. Then, by defining a global cost function as the weighted sum of all the SSD's, and minimizing it, we can determine the initial depth estimate. Moreover, to overcome the occlusion problem, we analyze the SSD functions using the initial estimate and employ highly confident SSD's selectively.

2.1. Object functions for individual stereo pairs

In this paper, we employ an area-based matching technique for the correspondence problem, and the SSD is used for the correlation measure between two windows. When the SSD is calculated under a general motion stereo, the size and shape of corresponding windows in the reference and the other image can be different, due to dilation, rotation and perspective distortion, caused by the general camera motion. Thus, it is inappropriate to use the window of fixed size and shape for correspondence. In this paper, we employ an adaptive window method, where the size and shape of the corresponding window varies, according to the relative position of the camera. Moreover, a sub-pixel registration technique is used to achieve higher precision.

Fig. 1 shows a camera projection model under a general motion, including both the translational and rotational motion simultaneously. Suppose that we have taken N+1 images at O, C_1, ..., C_i, ..., C_N,

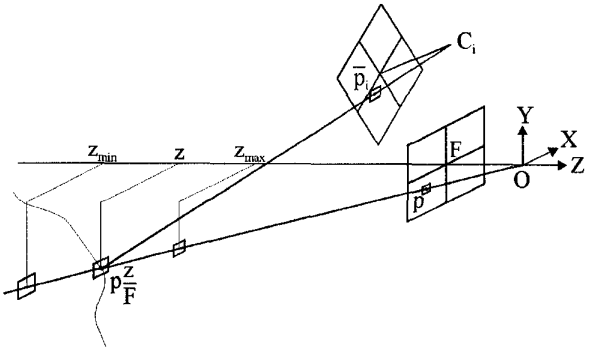


Fig. 1. General motion stereo camera model.

where O and C_i denote the focal point of the reference image and the i-th image, respectively. Let $f_0(x)$ and $f_i(x)$ be the reference and the i-th image intensity function, respectively. The SSD function of the i-th image over point p in the reference image is given by

$$SSD_i^p(z) = \sum_{q \in W} \{ f_0(q) - f_i(\overline{q_i}(z)) \}^2, \quad (1)$$

where W denotes a rectangular window surrounding the point p, q denotes the points in the window W, and $\overline{q_i}(z)$ denotes the corresponding points to q in the i-th image. If an ideal pin-hole camera model is assumed and the coordinate system is arranged as shown in Fig. 1, then $\overline{q_i}(z)$ can be calculated by

$$\overline{q_i}(z) = M_i q \frac{z}{F}, \quad (2)$$

where M_i is the projection matrix of the i-th camera and F is the focal length of the reference camera, respectively.

Finally, the object function for one stereo image pair can be obtained, by changing the z value within the search range $[z_{min}, z_{max}]$ with the search step size Δz. If the approximate range of the object position is known, then it can be used as the search range.

2.2. Integration of the object functions

The calculated SSD's are multiplied by the weights of each image, and the object function is finally derived by the summation of these weighted SSD's. The object function is given by

$$G_{initial}^p(z) = \sum_{i=1}^{N} w_i^p \cdot SSD_i^p(z). \quad (3)$$

In eq. (3), the weight w_i^p is the measure implying how confident the i-th image is for the point p of

the reference image. In this paper, we use the weight function w_i^p, given by

$$w_i^p = \sin\theta_i \cdot |C_i|, \quad (4)$$

where

$$\cos\theta_i = \frac{p \cdot C_i}{|p| \cdot |C_i|}, \quad 0 \le \theta_i \le 90^\circ . \quad (5)$$

In eq. (5), θ_i is the angle between p and C_i. Note that in case of the lateral motion stereo, as shown in Fig. 2(a), θ_i is constant. Thus, the weight w_i^p is directly proportional to the length of the baseline, implying that longer baseline generally yields smaller error in the lateral motion stereo. Fig. 2(b) shows the case where C_i is constant, that is, all C_i's are on the circle of which center is O. In this case, w_i^p is the distance from the point C_i to the line $\overrightarrow{Op}$. If C_i is on the line $\overrightarrow{Op}$, w_i^p becomes zero, implying that this image is no longer reliable. For example, in case of the axial motion stereo, we cannot acquire any information from the feature point in the right center of image. Whereas, if C_i is perpendicular to $\overrightarrow{Op}$, then the image is considered as the most reliable. Thus, the weight w_i^p can be interpreted as a generalized length of baseline.

2.3. Refinement of an initial estimate

In general, a longer baseline is not always desired. As the separation of the camera increases, the two images become less similar. Thus, some objects obtained by one camera may not even be visible to the other [7]. This paper, however, presents the method to decrease this kind of error.

First, using the weighted sum of SSD's, we find the initial depth estimate $z_{initial}$. Then, the final depth z_{final} can be refined by analyzing the SSD's. If the shape of the SDD yields no local minimum around $z_{initial}$, it could be due to the occlusion of objects. Thus, the SDD of this image pair is excluded. Moreover, if the SSD of one image pair yields relatively high value, compared with the SSD's of the other image pairs, it is also excluded. Finally, the overall object function is given by

$$G_{final}^p(z) = \sum_{i=1}^{N} d_i^p \cdot w_i^p \cdot SSD_i^p(z), \quad (6)$$

where

$$d_i^p = \begin{cases} 0, & \text{if } SSD_i^p(z) \text{ yields no local minimum around } z_{initial} \\ & \text{or prominently higher value.} \\ 1, & \text{otherwise.} \end{cases} \quad (7)$$

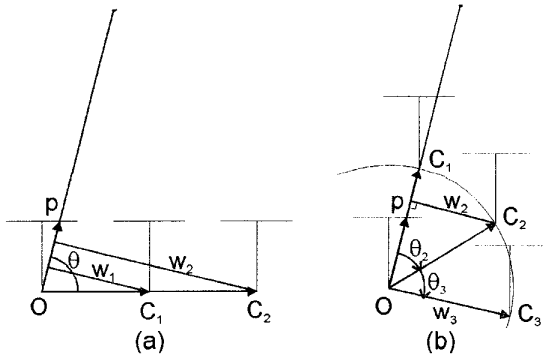


Fig. 2. Weight function: (a) Constant θ; (b) constant $|C_i|$.

3. Experimental results

We have tested the proposed algorithm on both synthetic and real images.

Fig. 3 shows the data set we have synthesized. We use five images, for example, two of which are shown in Fig. 3(a) and (b). Fig. 3(a) is the image that we use as the reference image, and Fig. 3(b) is the fourth image. We select 63 feature points in the reference image, which are shown in Fig. 3(c). Fig. 3(d) presents 76 feature lines, which are only used to visualize the extracted 3-D information for the feature points. In Fig. 4, the average error of 63 feature

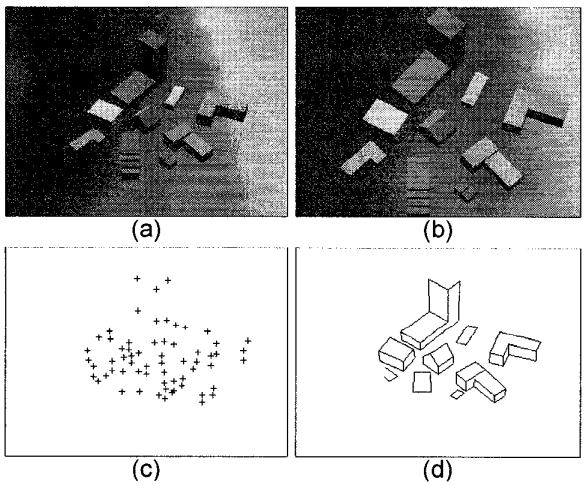


Fig. 3. Synthetic data set: (a) Reference image; (b) fourth image; (c) feature points; (d) feature lines.

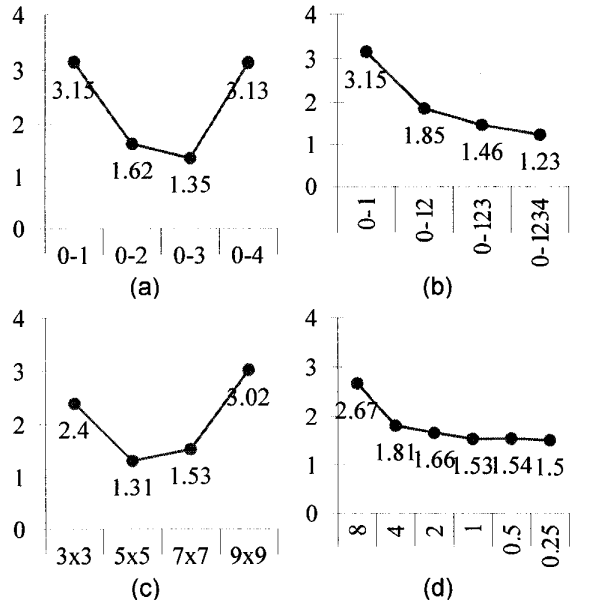


Fig. 4. Average error of 63 feature points: (a) Using one pair of images; (b) using multiple images; (c) window size; (d) search step size.

points are compared over a different condition. If we use only one pair of image, the error can be dependent on the used image, as shown in Fig. 4(a). But, if we combine multiple images, the more images yield the smaller error, as shown in Fig. 4(b). Fig. 4(c) and (d) show the average error, according to the window size and the search step size, respectively. From those results, it is concluded that the window size of 5×5 and the search step size of 1 mm provide the best performance. The extracted 3-D objects in different viewing directions are depicted in Fig. 5. The average depth error of the 63 feature points is presented in Table 1(a). Considering that the camera is about 1 m away from the objects, the resulting error is good enough. To compare the results with other methods, we perform the binocular stereo matching between the reference image and the other four images, and the average of these four

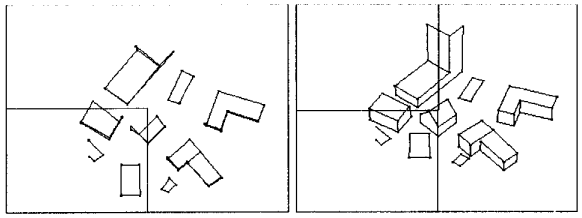


Fig. 5. Results on the synthetic data set in different viewing directions.

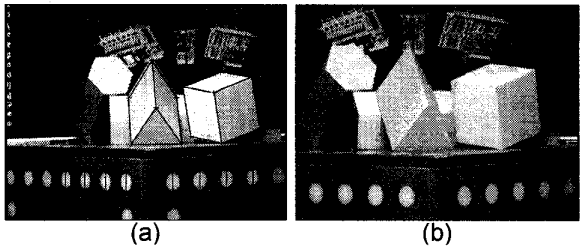


Fig. 6. Block data set: (a) Reference image and overlaid feature points and lines; (b) fifth image.

results are presented in Table 1(b). We also perform the binocular stereo matching, and then select an image that yields the minimum SSD, of which the result is shown in Table 1(c). Comparing these results, it can be observed that the proposed algorithm could reduce the error by 40% and 60% respectively.

Table 1. Average error in synthetic data set

(a)	Proposed Algorithm	1.23 mm
(b)	Stereo Matching and Averaging	2.06 mm
(c)	Stereo Matching and Selecting Min. SSD	3.12 mm

Fig. 6 shows the block data set we have tested. We use six images. Fig. 6(a) is the reference image, and Fig. 6(b) is the fifth image. We select 29 feature points and 37 feature lines in the reference image, which are shown in Fig. 6(a). Since this data set is real image, the true depth information for the feature points are not available. Instead, we know the distance between these feature points, that is, length of the blocks. The extracted 3-D information for the feature points and lines are rendered in different viewing directions in Fig. 7. The average error of 37 feature lines is presented in Table 2(a). Considering that the camera is about 1 m away from the blocks, of which the size is 3~5 cm, the result is promising. Compared with the other two methods, the proposed algorithm could reduce the error by 51% and 47%, respectively.

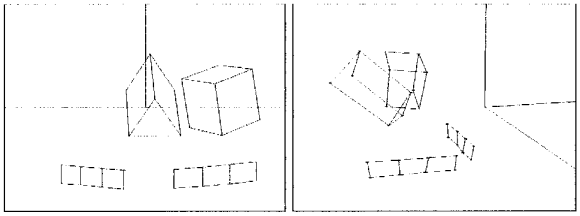
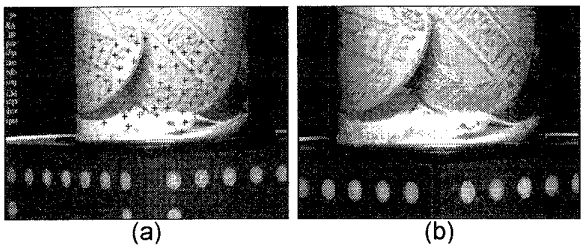


Fig. 7. Results on the block data set in different viewing directions.



(a) (b)

Fig. 8. Cup data set: (a) Reference image and overlaid feature points; (b) fourth image.

Table 2. Average error in block data set

(a)	Proposed Algorithm	1.30 mm
(b)	Stereo Matching and Averaging	2.64 mm
(c)	Stereo Matching and Selecting Min. SSD	2.44 mm

Fig. 8 shows the cup data set. We use five images, and Fig. 8(a) and (b) presents the reference image and the fourth image, respectively. We select 100 feature points on the surface of the cup in the reference image, which are shown in Fig. 8(a). The extracted 3-D feature points are presented in Fig. 9, showing that the results are qualitatively accurate.

Fig. 10 shows the CIL(Calibrated Imaging Lab.) data set. We use eleven images, and Fig. 10(a) and (b) presents the reference image and the tenth image, respectively. In this experiment, we have verified the advantage of the proposed object function over the sum of SSD's. 26 feature points, of which the true value is known, are provided. The average and maximum error of 26 points are presented in Table 3. It is noticed that the maximum error can be reduced, by using the proposed object function, rather than the sum of SSD's.

Table 3. Average and maximum error in CIL data set

		Average	Maximum
(a)	Sum of SSD's	3.47 mm	10.00 mm
(b)	Weighted sum of SSD's	3.27 mm	7.44 mm
(c)	Proposed object function	3.22 mm	6.44 mm

4. Conclusion

In this paper, we have proposed a new motion stereo algorithm to extract the 3-D information from multiple images, taken under a general motion. The object functions for individual stereo image pairs were represented, with respect to the distance, then these object functions were integrated in batch mode. By integrating the general motion stereo images and

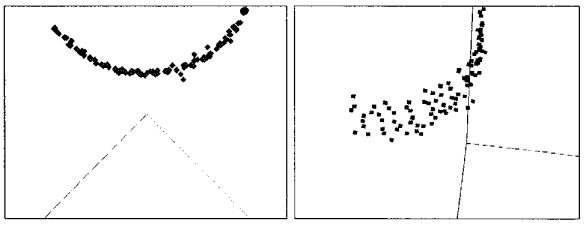


Fig. 9. Results on the cup data set in different viewing directions.

employing the adaptive window technique, we could improve the stability and the accuracy of the reconstruction significantly. Experiments on the synthetic and real stereo images showed promising results. Since we employ a general camera motion, it can be easily applied to the real applications, such as obstacle avoidance, smart missile navigation, and so on.

References

[1] N. Alvertos, D. Brazakovic, and R. C. Gonzalez, "Camera geometries for image matching in 3-D machine vision," *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. 11, No. 9, pp. 897-915, Sep. 1989.

[2] R. Y. Tsai, "Multiframe image point matching and 3-d surface reconstruction," *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. PAMI-5, No. 2, pp. 159-174, Mar. 1983.

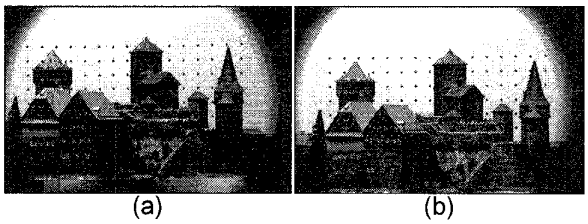
[3] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. 15, No. 4, pp. 353-363, Apr. 1993.

[4] T. Kanade, "A stereo machine for video-rate dense depth mapping and its new applications," *Proc. Computer Vision and Patt. Recogn.*, pp. 196-202, 1996.

[5] R. Collins, "A space-sweep approach to true multi-image matching," *Proc. Image Understanding Workshop*, pp. 1213-1220, Feb. 1996.

[6] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," *Proc. Int. Conf. Computer Vision*, pp 754-760, Jan. 1998.

[7] B. K. P. Horn, *Robot Vision*, Cambridge, MA: MIT Press, 1986.



(a) (b)

Fig. 10. CIL data set: (a) Reference image; (b) tenth image.