Stereo Vision Based 3D Map Construction Using Structure from Motion

Ryuta Harada, Kenichi Asami, and Mochimitu Komori

Department of Applied Science for Integrated System Engineering, Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyusyu, Fukuoka 804-8550, Japan

harada@es.ise.kyutech.ac.jp

Abstract

This paper describes 3D map construction system using stereo vision and the Structure from Motion method. First, parallax image by the triangulation method is generated so that the 3D coordinates corresponding to the feature points are obtained. Next, the Structure from Motion method is applied to the feature points tracking for frame by frame images so that the self-position of the camera can integrate the 3D map. In order to generate more accurate 3D map, the stereo vision system must deal with problems such as aperture problem and noise. The construction of the stereo measurement system and the experiment with mitigating the visual problems are conducted. The practical 3D map for the typical corridor environment was generated from the high-speed stereo image processing within 900 ms per scene.

Keywords: 3D map construction, stereo vision, the Structure from Motion, factorization method, feature tracking.

1. Introduction

These past several years, 3D map reconstruction system with high availability of visual recognition have been greatly expected in variety of academic and industrial fields. The capability of 3D reconstruction in different environments must be equipped with the map construction with enough rich information for navigating the spaces. The stereo vision system^(1,5) enables to measure the 3D coordinates at the feature points in the visual world and to recognize the space relationships between objects. Furthermore, the Structure from Motion method^(2,4,6) is effective to identify the camera position itself in the 3D space. Therefore, we propose the integrated method of stereo vision measurement with edge points and 3D reconstruction with feature points according to movements of the camera. The experiment of the 3D map construction is demonstrated in a practical environment.

The aim of this paper is to construct a stereo vision system integrated with the Structure from Motion. Many stereo vision techniques have been introduced for 3D measurement system. We present a simple method of stereo measurement focusing on edge points to grasp environmental recognition without heavy computation. The stereo vision system is established by using easily available cameras in this research. The results of 3D measurement include the framework for peripheral map with many point groups, but it is difficult to distinguish the movements of objects and the camera itself. Therefore, we combine the localization method by the Structure from Motion to perform visual encoding of the self-position of the camera. The 3D map construction could be completed by integrating with static 3D point groups and dynamic point tracking in the sequences of the camera movement.

2. 3D Position Measurement

2.1 Edge Extraction

Correspondences among the feature points in the left and right images are obtained from the edge image. The edge image shows the location where the brightness of the image is greatly changed. The edge image extracts only the significant points in order to avoid the full searching in the original image. The edge feature points often represents the boundary locations among objects in the real world. The Sobel filter as shown in Fig. 1 to calculate the edge points in the image is used as the conventional derivative operation. Fig. 2 shows a result of edge image for corridor environment.

-1	-2	-1
0	0	0
1	2	1

Vertical direction

-1	0	1
-2	0	2
-1	0	1

Horizontal direction

Fig. 1. Sobel filter for edge extraction

Fig. 2. A result of edge extraction

2.2 Disparity Calculation

Template matching is performed for detecting a specific pattern contained in an image by comparing another image. In this system, the matching by using SAD (Sum of Absolute Difference) is applied, and the disparity of the left and right images is calculated. The procedure of the SAD method is as follows. We cut out the area centered at an edge point of the images, superpose it to all areas which can be compared in another image, and evaluate the absolute value of the difference among the corresponding pixels. By comparing the SAD value to a threshold, the candidate of edge point can be selected. The SAD value says that the similarity of compared areas, where the two areas are not similar if the value is greater, but the two areas are similar if the value is smaller. After comparing all of the areas which can be compared in the searching image, the area with the smallest SAD value is associated with the corresponding area. In this case, the all coordinates of edge points in the reference image are corresponded with the coordinates in the search image. The difference between the horizontal coordinates in the reference and search images is obtained as parallax data.

Fig. 3 shows a parallax image where red represents close point and blue represents far point to the stereo camera system. The most edge points in the left (reference) image are corresponded to the edge points in the right (search) image. Some edge points are not corresponded because the similarity by SAD is evaluated to be low below the threshold. The stereo camera is manually adjusted in advance and the result of the stereo image processing is generated within 900 ms per a pair of images. This result of a disparity calculation is mapped to the 3D position in the space.

Fig. 3. A result of disparity calculation

2.3 Triangulation

The important technique in performing the 3D position measurement is triangulation. The triangulation is a method for measuring the distance to the target in a remote location from the two observation points, and it uses the triangle determined if the length of one side and the angle of the ends are known. For 3D position measurement, the triangulation is used to convert the points on the obtained parallax images to real 3D coordinates.

As shown in Fig. 4, the distance of the left and right cameras denotes T , and the centers of the projection of each camera denote O_l and O_r . The focal length f of the left and right cameras are the same value, and both images place at the same planes. Principal points c_l and c_r that exist at intersection of the optical axis and the image plane are adjusted to be parallel, and row of each camera is aligned. The point P in the real world can be found in the left and right image plane, and it has the coordinates x_l and x_r . $\triangle Px_lx_r$ and $\triangle PO_lO_r$ have similar relationship, so that the following equation is formed.

$$Z = \frac{fT}{x_l - x_r} \quad (1)$$

Where $x_l - x_r$ denotes the parallax. The focal length f and the distance T between two cameras are predefined values. By using this equation, we can determine the depth Z from the camera to the point P if the parallax of the left and right images is obtained. The values of X and Y coordinate can be calculated by using the similar triangles in another viewpoint. In our system, too small parallax is eliminated because the measured accuracy is not good to make the environmental map. For the all edge points in the parallax image, 3D position is measured by the triangulation method so as to construct the 3D map as point cloud in a location.

Fig. 4. Relationship between disparity and Z coordinate

3. Structure from Motion

The Structure from Motion⁽⁶⁾ is a method for obtaining a 3D structure of the object from plural images captured from the single camera with changing view and the camera position. The procedure of the Structure from Motion conforms to the following steps: First, feature points for plural images for the skeleton of the object is extracted, and the sequential correspondence of the feature points between the images is acquired. Then, a matrix describing the correspondence is generated by covariance, and the 3D coordinates of each feature point and the camera position is estimated by factorization.

A series of images have F piece of time sequence taken by a single camera with relative motion to the space. The prominent feature points P such as corners of the object from the first frame is extracted in order to track the sequential movement. The position of a feature point p on the image of the f -th frame denotes (x_{fp}, y_{fp}) , and a matrix W is defined as follows:

$$W = \begin{bmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & \vdots & \vdots \\ x_{F1} & \cdots & x_{FP} \\ y_{11} & \cdots & y_{1P} \\ \vdots & \vdots & \vdots \\ y_{F1} & \cdots & y_{FP} \end{bmatrix} \quad (2)$$

Where W means a matrix of $2F \times P$, and it has X coordinate in the upper half and Y coordinate in the lower half. Each column corresponds to the tracking result for the one feature point and each row corresponds to the X and Y coordinate of the all feature points in one image. Matrix W' subtracted the average value of the row from each element of W is made.

$$W' = W - \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_F \\ \bar{y}_1 \\ \vdots \\ \bar{y}_F \end{bmatrix} [1 \quad \cdots \quad 1] \quad (3)$$

This represents the position of the feature points in the relative position to the center of gravity. W' is called the measurement matrix in the factorization method in order to restore the structure and motion.

When the measurement matrix W' is factorized as shown in equation (4), the liner algebraic technique of singular value decomposition is performed. According to the singular value decomposition theorem, any matrix W' of $2F \times P$ can be decomposed into a product of the three matrices of $2F \times 3$, 3×3 , and $3 \times P$.

$$W' = MS = \begin{bmatrix} i_1^T \\ \vdots \\ i_F^T \\ j_1^T \\ \vdots \\ j_F^T \end{bmatrix} [s_1 \quad \cdots \quad s_P] \quad (4)$$

4. Experiments

The stereo vision system using easily available USB cameras is constructed for the experiment of making 3D map as shown in Fig. 5. The distance between the cameras is set to 150 mm. Image size is 640×480. We adopt the perspective model for this camera used in the experiment.

In this experiment, the corridor environment shown in Fig. 6 with approximately 4,000 mm of total length was measured at 9 observation points. Along with the corridor of a straight position and T-branch, the stereo vision system obtained 3D coordinates every 6,060 mm progress.

The 3D coordinates are integrated according to the camera movement, and visualized as the 3D map shown in Fig. 7.

Fig. 5. Stereo vision system for the experiment

Fig. 6. Experimental environment for typical corridor

Fig. 7. A result of 3D map generation for typical corridor

As shown in Fig. 7, a rough outline of the 3D map was constructed. However, there are some noise on the floor of the corridor. This reason is considered that the cameras detects a light from fluorescent lamp to reflect the floor of the corridor. Also there are some mismatch correspondences by SAD similarity evaluation and the accuracy of the 3D measurement at distant point decreases by the limitation of image resolution. From these reasons, it is necessary to perform processing to erase the reflected light and to correct the measured data at the near position.

Next, the self-position estimation by the Structure from Motion was performed in the same corridor. In the sequential frames while moving by 100 mm to the right, the feature points between each frame was tracked from the first to the second frames. Fig. 7 shows the first frame where the feature points are detected and the tracking of the feature points to the second frame. The green circles in the image is feature point of the current frame and the green line is the trajectory of the feature point from the previous frame.

Fig. 8. A result of 3D map generation for typical corridor

Table. 1. The displacement of the moving distance

	X coordinate		Y coordinate		Z coordinate	
	value	error	value	error	value	error
1 st -2 nd	88	12	0	0	0	0
2 nd -3 rd	108	8	0	0	0	0

Table 1 shows the displacement of the moving distance of the X, Y, and Z coordinates of the feature points from the first to the third frame. From this table, when the camera moves 100 mm to the right, distances of Y and Z coordinates have not changed and distance of X coordinate has nearly exact value 100mm.

5. Conclusions

This paper presents the 3D map construction using the stereo vision and the Structure from Motion. The 3D measurement was achieved by the triangulation and the 3D map was generated. The self-position estimation of the camera was performed by tracking the feature points. As the future work, the improvement of feature matching accuracy is needed. The integration of the stereo vision system and the Structure from Motion would be required for the automatic map extension. Moreover, the point cloud library would be useful to visualize the 3D map and to handle the generated data for the many applications.

References

- (1) H. Hagiwara, Y. Touma, K. Asami, and M. Komori : "FPGA-Based Stereo Vision System Using Gradient Feature Correspondence", Journal of Robotics and Mechatronics, Vol. 27, No. 6, pp. 681-690, 2015.
- (2) G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla : "Segmentation and Recognition Using Structure from Motion Point Clouds", Proc. of the 10th European Conference on Computer Vision, Part I, pp. 44-57, 2008.
- (3) P. Torr and A. Zisserman : "Feature Based Methods for Structure and Motion Estimation", Proc. of the International Workshop on Vision Algorithms: Theory and Practice, pp. 278-294, 1999.
- (4) P. Li, D. Farin, R. Gunnewiek, and P. With : "On Creating Depth Maps from Monoscopic Video Using Structure from Motion", Proc. of the 27th Symposium on Information Theory, pp. 508-515, 2006.
- (5) M. Chandraker, J. Limy, and D. Kriegman : "Moving in Stereo: Efficient Structure and Motion Using Lines", Proc of the 12th International Conference on Computer Vision, pp. 1741-1748, 2009.
- (6) C. Tomasi and T. Kanade : "Shape and Motion from Image Streams: A Factorization Method", International Journal of Computer Vision, Vol. 9 (2), pp. 137-154, 1992.