AN AUTOMATIC 3D RECONSTRUCTION METHOD BASED ON MULTI-VIEW STEREO VISION FOR THE MOGAO GROTTOES

Jie Xiong [a, *], Sidong Zhong [a], Lin Zheng [a]

[a] School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China – xiongjiewhu1989@163.com, sdzhong@whu.edu.cn, zl@whu.edu.cn

KEY WORDS: 3D Reconstruction, Multi-view vision, Epipolar Constraint, Correlation, Texture mapping, Mogao Grottoes

ABSTRACT:

This paper presents an automatic three-dimensional reconstruction method based on multi-view stereo vision for the Mogao Grottoes. 3D digitization technique has been used in cultural heritage conservation and replication over the past decade, especially the methods based on binocular stereo vision. However, mismatched points are inevitable in traditional binocular stereo matching due to repeatable or similar features of binocular images. In order to reduce the probability of mismatching greatly and improve the measure precision, a portable four-camera photographic measurement system is used for 3D modelling of a scene. Four cameras of the measurement system form six binocular systems with baselines of different lengths to add extra matching constraints and offer multiple measurements. Matching error based on epipolar constraint is introduced to remove the mismatched points. Finally, an accurate point cloud can be generated by multi-images matching and sub-pixel interpolation. Delaunay triangulation and texture mapping are performed to obtain the 3D model of a scene. The method has been tested on 3D reconstruction several scenes of the Mogao Grottoes and good results verify the effectiveness of the method.

1. INTRODUCTION

With the rapid development of computer technology and sensing technology, three-dimensional (3D) digitization of objects has attracted more and more attention over the past decades. 3D modelling technology has been widely applied in various digitization fields, especially cultural heritage conservation. 3D digitization of cultural heritage is mainly used for digital recording and replication of cultural heritage. Considering the precious value of cultural heritage objects, non-contact and non-destructive measure approaches are generally taken to acquire 3D models. For realistic application, automatic, fast and low-cost 3D reconstruction methods with high precision are required.

A number of active and passive technologies (Pavlidis et al., 2007) are developed for 3D digitization of cultural heritage. Laser scanning methods (Huang et al., 2013) and structured light methods (Zhang et al., 2011) are typical active methods. The most significant advantage of laser scanning is high accuracy in geometry measurements. Nevertheless, the models reconstructed by laser scanning usually lack good texture and such devices have high cost. As passive methods, vision-based methods have the ability to capture both geometry information and texture information, requiring less expensive devices. According to the amount of cameras used, vision-based methods are divided into monocular vision, binocular vision and multi-view vision. Monocular vision methods can obtain depth information from two-dimensional characteristics of a single image or multiple images from a single view (Massot and Hérault, 2008; Haro and Pardàs, 2010). Such methods are usually not very robust to the environment. Moreover, monocular vision methods can gain 3D information from a sequence of images from different views (shape from motion, SFM) (Chen et al., 2012). The SFM method has a high time cost and space cost. Binocular vision method can acquire 3D geometry information from a pair of images captured from two known position and angles. This method has high automation and stability in reconstruction. But this method easily leads to mismatched points due to repeatable or similar features of binocular images (Scharstein and Szeliski, 2002). In order to reduce the possibility of mismatching, 3D measurement systems based on multi-view vision have been developed (Setti et al., 2012). Generally, the systems have complex structure.

This paper presents an automatic 3D reconstruction method based on multi-view stereo vision. This method has reconstructed 3D models of several scenes of No.172 cave in the Mogao Grottoes using a portable four-camera photographic measurement system (PFPMS) (Zhong and Liu, 2012). The PFPMS is composed of four cameras to add extra matching constraints and offer redundant measurement, resulting in reducing the possibility of mismatching and improving measure accuracy relative to traditional binocular systems.

2. 3D RECONSTRUCTION METHODOLOGY

The authors take reconstruction of a scene of a wall for example to illustrate the whole process of 3D reconstruction, including multi-view images acquisition, multi-view image processing, triangulation and texture mapping.

2.1 Multi-view images acquisition

As the main hardware system, the PFPMS consists of four cameras with the same configuration parameters, which observes the target object at a distance of 2.0-5.0 m. Four

* Corresponding author

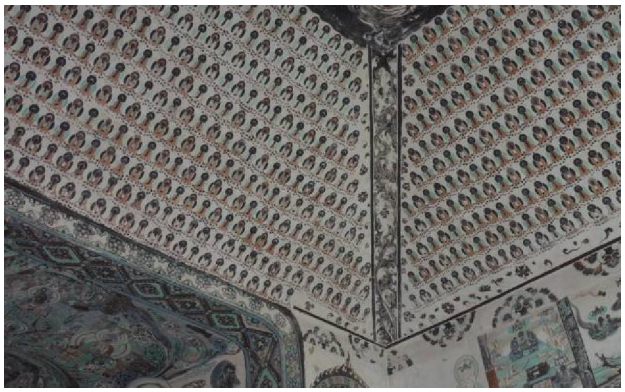
images with a high image resolution of 6016×4000 can be captured synchronously by a button controller connected to the switch of shutters of the four cameras. The overall structure of the PFPMS is similar to a common binocular vision system and the difference is that two cameras with upper-lower distribution are substitute for each camera of a binocular system respectively, as shown in Figure 1. The four cameras have rectangular distribution and their optical axes are parallel to each other to minimize the impact of perspective distortion on feature matching. The distance between the left or right cameras is about 15 cm and the distance between the upper or lower cameras is about 75 cm. On the one hand, the baseline between the left two cameras or the right two cameras is short. As a result, the very small difference between the two images captured by them can help improve accuracy of matching. Furthermore, the left cameras and the right cameras can form four binocular systems with long baseline to calculate space position of feature points. Thus every point can be measured four times to improve precision. The corresponding parameters of the four cameras and the parameters of relative position of any two cameras need to be obtained before the measurement. The cameras can be calibrated with a tradition pinhole model (Tsai, 1987), and then 3D space coordinates of any point can be calculated with its coordinates in the four images.

Figure 1. The PFPMS

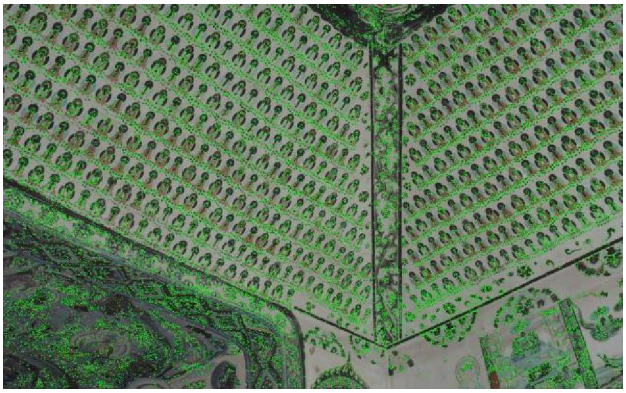
2.2 Multi-view images processing

The multi-view images processing is divided into extraction of feature points and matching of feature points. For convenience, let UL, LL, UR, LR image represent the upper-left, lower-left, upper-right, lower-right image respectively.

2.2.1 Extraction of feature points: With high detecting speed and high position accuracy, Harris corners (Harris and Stephens, 1988) are chosen as feature points for matching. We adopt corner extraction of image partition to ensure corner points' uniform distribution. Figure 2(a)(b) show the original UL image and its distribution of Harris corners respectively.



(a) The Original UL image



(b) Distribution of Harris Corners in UL image

Figure 2. Extraction of Harris Corners

In order to reduce the search range, the Harris corners detected are stored in a sub-regional way. The feature points in the four images can be extracted in this way.

2.2.2 Matching of feature points: Template matching methods are used to search the homologous image points in most stereo matching algorithms. Traditional template matching methods mainly include sum of squared differences (SSD), sum of absolute differences (SAD), normalized cross correlation (NCC) and zero mean normalized cross correlation (ZNCC) (Lazaros et al., 2008). These methods weigh the degree of similarity between two points by calculating the difference between the pixels inside the rectangle window around one point and the pixels inside the rectangle window around the other point. ZNCC is chosen for matching due to its stronger anti-noise ability. Let $I_1(x,y)$, $I_2(x,y)$ represent the intensity of the pixel at (x,y) in image 1 and image 2 respectively and ZNCC can be given by the following expression.

$$ZNCC(x_1,y_1,x_2,y_2)= \frac{\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left[I_1(x_1+i,y_1+j)-\bar{I}_1\right]\left[I_2(x_2+i,y_2+j)-\bar{I}_2\right]}{\sqrt{\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left[I_1(x_1+i,y_1+j)-\bar{I}_1\right]^2}\sqrt{\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left[I_2(x_2+i,y_2+j)-\bar{I}_2\right]^2}} \quad (1)$$

$$\bar{I}_1=\frac{1}{2N+1}\sum_{i=-N}^{N}\sum_{j=-N}^{N}I_1(x_1+i,y_1+j) \quad (2)$$

$$\bar{I}_2=\frac{1}{2N+1}\sum_{i=-N}^{N}\sum_{j=-N}^{N}I_2(x_2+i,y_2+j) \quad (3)$$

where N = the half of the size of template window. (N is set to 10 pixels in actually matching)

x_1, y_1 = coordinates of the matched point in image 1.

x_2, y_2 = coordinates of the matched point in image 2.

$\bar{I}_1$ = average intensity of the pixels inside the window around (x_1, y_1).

$\bar{I}_2$ = average intensity of the pixels inside the window around (x_2, y_2).

Figure 3 shows the main matching scheme based on epipolar constraint. Let l_{UL-LL}, l_{UL-UR}, l_{UL-LR}, l_{LL-UR}, l_{LL-LR}, l_{UR-LR}

represent the epipolar lines which can be obtained from the known parameters of the four cameras (Xu et al., 2012). The matching process is described as the following steps:

a) For any point P_{UL} in UL image, search its corresponding point along the epipolar line l_{UL-LL} in LL image and find some possible points which have a ZNCC value above 0.9. Rank these points by ZNCC value from high to low and the top five points are chosen as the candidate matched points.

b) Let P_{LL} represent the first candidate point. l_{UL-UR}, l_{UL-LR}, l_{LL-UR}, l_{LL-LR} can be obtained from the position of P_{LL} and P_{UL} respectively.

c) Find the matched point P_{UR} in the rectangle region (40 pixel $\times$ 40 pixel) around the intersection of l_{LL-UR} and l_{UL-UR} based on the maximum ZNCC value with P_{UL}. If the maximum value is less than 0.7, remove P_{LL} from the candidate points and return to step (b).

d) Find the matched point P_{LR} in the rectangle region (40 pixel $\times$ 40 pixel) around the intersection of l_{UL-LR} and l_{LL-LR} based on the maximum ZNCC value with P_{UL}. If the maximum value is less than 0.7, remove P_{LL} from the candidate points and return to step (b).

e) Calculate the ZNCC value between P_{UR} and P_{LR}. If the value is less than 0.9, remove P_{LL} from the candidate points and return to step (b).

f) Obtain l_{UR-LR} from the position of P_{UR} and calculate the distance from P_{LR} to l_{UR-LR}. If the distance is less than 5 pixels, remove P_{LL} from the candidate points and return to step (b).

g) Calculate the matching error defined as the sum of the distance between each matched point and the intersection of the epipolar lines of two other matched points with long baseline relative to it. If the matching error is less than 20 pixels, P_{UL}, P_{LL}, P_{UR}, P_{LR} can be regard as the homologous points and return step (a) for the matching of the next point. Otherwise, remove P_{LL} from the candidate points and return to step (b).

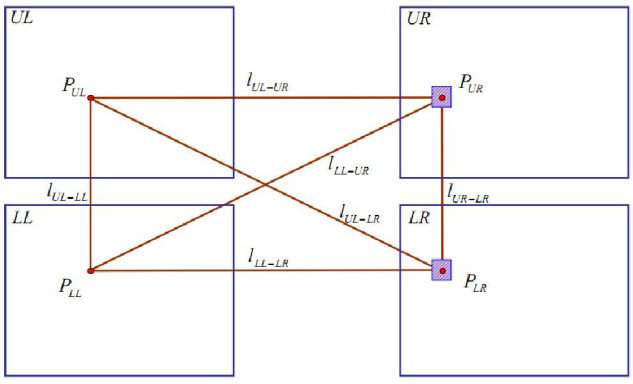


Figure 3. Matching scheme

Figure 4 shows the whole processing flow of the above-mentioned matching method.

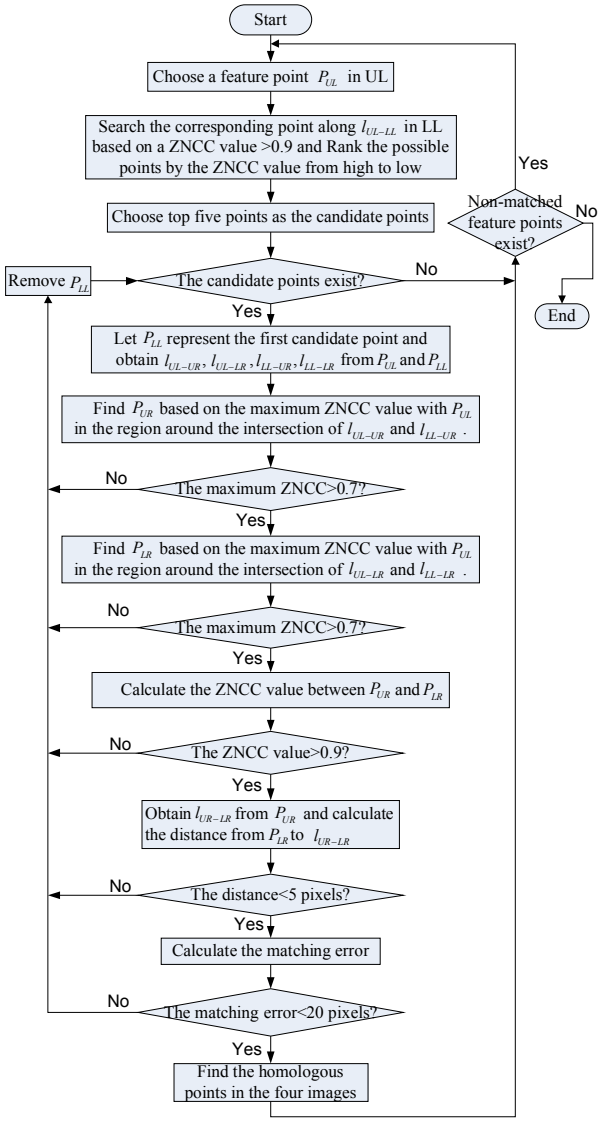


Figure 4. The processing flow of the matching method

After matching, sub-pixel interpolation operation can be performed to improve measurement precision. Bicubic interpolation is chosen to gain sub-pixel position of the homologous points because of its smooth interpolation effect. As the four cameras of the PFPMS form four binocular systems with a long baseline (UL-UR, UL-LR, LL-UR, LL-LR), for every matched point, take the average of four space coordinates respectively calculated from it and its homologous points as the final coordinates of its corresponding space point. Figure 5 shows the 3D point cloud.

2.3 Triangulation and texture mapping

Generally, the surface of the object can be expressed with a triangulated irregular net. Delaunay Triangulation (Tsai, 1993) is performed to process the point cloud obtained from stereo matching. In order to avoid appearance of some long and narrow triangles with long sides during triangulation, the

length of every triangle's sides should be limited. Figure 6 shows Delaunay triangulation of the point cloud.

In order to reconstruct a model with texture, every point's colour information extracted from one of the captured images can be used for texture mapping. The UL image is selected as the texture image. To every triangle, each vertex's texture coordinates can be obtained from the image coordinates of its matched point in the texture image, and the texture coordinates of internal points can be calculated by linear interpolation of the vertexes' texture coordinates. The texture image is mapped automatically to a model in this way. Finally, the 3d model of the scene is generated, as shown in Figure 7.

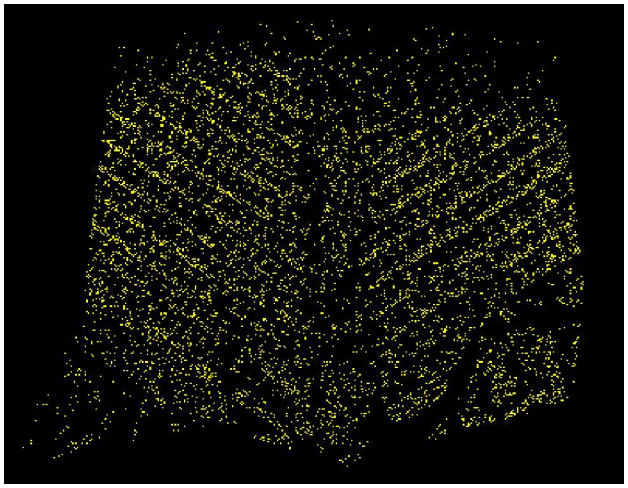


Figure 5. The 3D point cloud (7484 points)

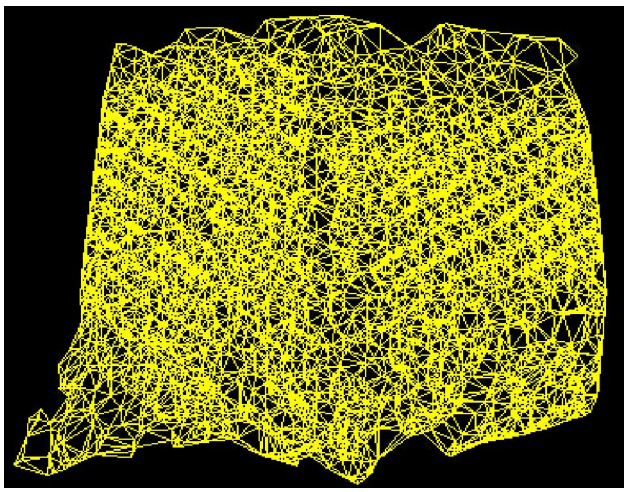
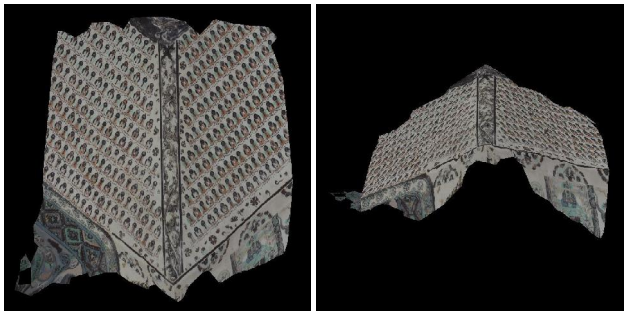


Figure 6. Delaunay Triangulation of the point cloud



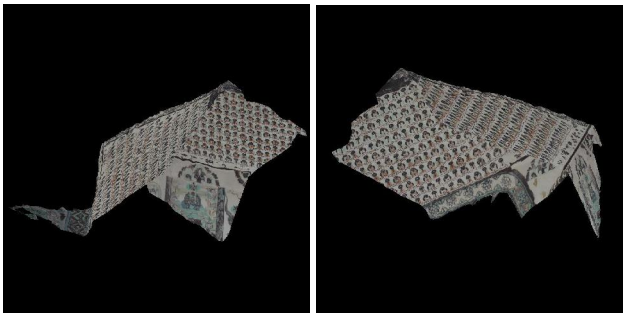


Figure 7. The 3D model of the scene from different views

3. EXPERIMENT RESULTS

The 3D model of the scene of a wall has been obtained with the matching method above and a good result is given. In order to test the stability and adaptability of the method, the 3D models of Scene A and Scene B are reconstructed respectively. Figure 8 and Figure 9 show the results. The results show these models describe geometric structure and texture of the scenes realistically as a whole. Not dense enough points are extracted from the region with poor texture. It leads to the loss of local model. For instance, incomplete models of the Buddhas indicate this point.

(a) The original UL image of Scene A

(b) The 3D model of Scene A

Figure 8. 3D Reconstruction of Scene A

(a) The original UL image of Scene B

(b) The 3D model of Scene B

Figure 9. 3D Reconstruction of Scene B

4. CONCLUTION

This paper proposes an automatic 3D reconstruction method based on multi-view stereo vision. 3D models of several scenes of No.172 cave in the Mogao Grottoes have been reconstructed using a portable four-camera photographic measurement system. The cameras of the measurement form two binocular systems with a short baseline and four binocular systems with a long baseline. The binocular system with a short baseline is used for rapidly matching with the small difference between the two images. The binocular systems with a long baseline are used for multiple measurements. Compared with a traditional binocular system, the PFPMS have the advantage of reducing the possibility of mismatching and improving measurement accuracy. The experiment results show the effective of this matching method.

The limitation of the method is that the point cloud is not enough dense in a region with poor texture. Besides, only several models of local scenes can be reconstructed but are not complete. The future work will be focused on obtaining a dense point cloud by introducing structured light and stitching of the models reconstructed from different perspectives.

ACKNOWLEDGEMENTS

This work is supported by the 973 Program (2012CB725301) and National surveying and mapping geographic information public welfare industry special funding scientific research projects (210600001).

REFERENCES

Chen, S., Wang, Y. and Cattani, C., 2012. Key issues in modeling of complex 3D structures from video sequences. http://www.hindawi.com/journals/mpe/2012/856523/.

Haro, G. and Pardàs, M., 2010. Shape from incomplete silhouettes based on the reprojection error. Image and Vision Computing, 28(9), pp. 1354-1368.

Harris, C. and Stephens, M., 1988. A combined corner and edge detector. In Proceedings of Alvey Vision Conference, The Plessey Company, ed., pp. 189–192.

Huang, H., Brenner, C. and Sester, M., 2013. A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data. In: *ISPRS Journal of Photogrammetry & Remote Sensing*, 79, pp. 29-43.

Lazaros, N., Sirakoulis, G. C. and Gasteratos, A., 2008. Review of stereo vision algorithms: from software to hardware. International Journal of Optomechatronics, 2(4), pp. 435-462.

Massot, C. and Hérault, J., 2008. Model of frequency analysis in the visual cortex and the shape from texture problem. International Journal of Computer Vision, 76(2), pp. 165-182.

Pavlidis, G., Koutsoudis, A., Arnaoutoglou, F., Tsioukas, V. and Chamzas, C., 2007. Methods for 3D digitization of cultural heritage. Journal of cultural heritage, 8(1), pp. 93-98.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision, 47(1-3), pp. 7-42.

Setti, F., Bini, R., Lunardelli, M., Bosetti, P., Bruschi, S. and De Cecco, M., 2012. Shape measurement system for single point incremental forming (SPIF) manufacts by using trinocular vision and random pattern. Measurement science and technology, 23(11), 115402.

Tsai, R. Y., 1987. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. Robotics and Automation, IEEE Journal of, 3(4), pp. 323-344.

Tsai, V. J., 1993. Delaunay triangulations in TIN creation: an overview and a linear-time algorithm. International Journal of Geographical Information Science, 7(6), pp. 501-524.

Xu, S. B., Xu, D. S. and Fang, H., 2012. Stereo Matching Algorithm Based on Detecting Feature Points. In Advanced Materials Research, Vol. 433, pp. 6190-6194.

Zhang, K., Hu, Q. and Wang, S., 2011. A fast 3D construction of heritage based on rotating structured light. In International Symposium on Lidar and Radar Mapping Technologies.

International Society for Optics and Photonics, pp. 82861Z-82861Z.

Zhong, S. D. and Liu, Y., 2012. Portable four-camera three-dimensional photographic measurement system and method. http://worldwide.espacenet.com/publicationDetails/biblio?CC=CN&NR=102679961B&KC=B&FT=D.