3D SCENE RECONSTRUCTION SYSTEM WITH HAND-HELD STEREO CAMERAS

SangUn Yun, Dongbo Min and Kwanghoon Sohn

Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea
khsohn@yonsei.ac.kr

ABSTRACT

3D scene modeling is a challenging problem and has been one of the most important research topic for many years. In this paper, we describe the 3D scene reconstruction system that creates 3D models with multiple stereo image pairs acquired by hand-held device. Our algorithm consists of the following two steps, which is depth reconstruction and model registration. In the first part, we obtain the depth map with stereo matching and camera geometry in each view. The algorithm is based on adaptive window methods in hierarchical frameworks. In the second part, we use SIFT feature to estimate the camera motion. LMedS algorithm reduces the effect of outliers in this process. Experimental results show that the proposed algorithm provides accurate disparity map in various types of images, and the 3D model of real world's scene.

Index Terms— Disparity estimation, SIFT algorithm, robust estimation, scene reconstruction

1. INTRODUCTION

Obtaining 3D models from images is an active research topic in computer vision. A few years ago the main applications were just robot guidance and visual inspection. Nowadays, however, more and more demands are created in visualization and measurements such as planetary rover exploration, forensics, etc. Objects scanning and environment modeling can be divided into two types, outside-looking-in and inside-looking-out structures. The prior one is more suitable for object scanning whereas the other is more suitable for environment modeling. 3D data can be obtained using various range finders and computed from stereo images or monocular sequences. In this paper, we use hand-held stereo camera to obtain depth information for satisfying the requirement, which is the need of low cost and high quality in system. Also, this system has an inside-looking-out structure because the proposed algorithm is for scene reconstruction. 3D reconstruction system consists of several processes. However, we will review essential techniques including depth acquisition and view registration.

"This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA" (IITA-2005-(C1090-0502-0027))

Depth acquisition: The main approaches for depth acquisition include structured light, laser scanning and stereo. The structure light approach uses a projector to illuminate the object with patterns and recovers the 3D shape from monocular images. Auto synchronous laser scanners can be used for both objects and environments due to their long depth field and high accuracy at close range[9]. In this paper, stereo matching algorithm is used for depth acquisition in the low cost system. It is a very difficult problem due to ambiguous region such as the occluded and textureless areas. To solve this problem, a number of algorithms have been proposed [1].

3D view registration: 3D registration is the technique that integrates with multiple 3D date sets by tracking the camera motion with a sensor or matching the data sets manually or automatically. The most common methods for automatic registration of 3D data is Iterative Closest Point (ICP) algorithm [2]. It iteratively minimizes the distances between the overlapping regions of two set of 3D points, lines and surfaces. The alternative method for the estimation of camera motion uses feature matching.However, artificial markers are sensitive to a change of environment, natural marker are preferred to use in the estimation.

2. OVERVIEW OF SYSTEM

The system is made up of hand-held stereo cameras and a computer. Stereo image pairs are captured in the each view and dense disparity is computed. The proposed algorithm estimates disparity and boundary map simultaneously. We assume that captured images are rectified in advance, so that epipolar line becomes horizontal. The estimated disparity maps are converted into depth information with camera parameters. The system does not require any sensors to estimate the camera motion. It automatically computes the rotation (R) and translation (T) matrix using corresponding 3D feature sets. For feature matching, Scale Invariant Feature Transform (SIFT) feature [3] is used and we reduce the effect of outlier using Least Median of Square (LMedS) methods [4]. The 3D data sets are registered by the recovered camera motion and the final scene is reconstructed. The reconstructed model is then converted into surface triangular meshes, which are augmented by mapping texture from the color camera images. Fig.1 shows system overall process.

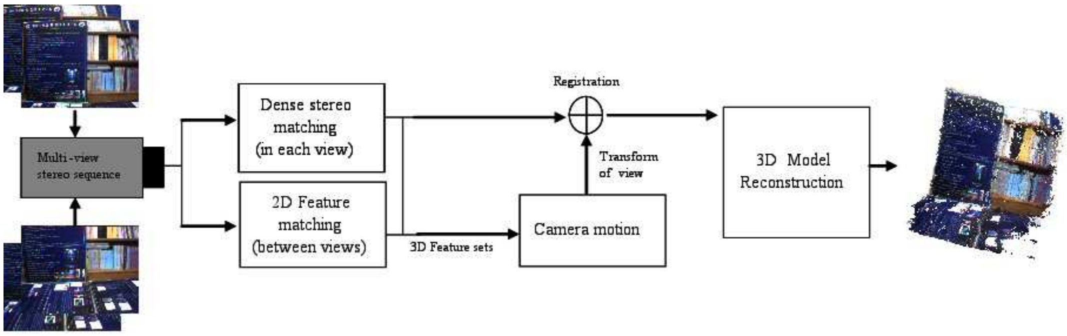

Fig. 1. Scene modeling system architecture

3. IMPLEMENTATION

3.1. Dense stereo matching

In this paper,we improve our previous work about part of optimal window size decision [10]. We improve a technique of window size decision. Generally, the performance of window methods depends on how well the selection of window adaptively at each pixel [5]. As window size is increased from small to large, the results range from accurate disparity boundaries but noisy in low textured areas, to more reliable in low textured areas but blurred disparity boundaries. It is very difficult to obtain the optimal window for each pixel, which are reliable in both low textured areas and object boundaries. We classify an image into the depth-discontinuous and continuous region, and different approaches are used in each region as shown in Fig.2. To solve the matching problem in homogeneous regions, a window size increases until the window includes sufficient texture information or encounters the boundary map to overcome the problems in the textureless regions. In boundary regions, to reduce 'overfitting' problem, we use the 9 different shape window models, as shown in Fig.3. Each window is decided according to the direction of the corresponding edge. Therefore, the key technique in the algorithm is the estimation of dege direction and boundary map to decide optimal windows.

Edge direction detection: To estimate the direction of edges, we use the texture information for each level block. We define four types of edge directions (horizontal, vertical, diagonal and anti-diagonal). To determine the type of block, we utilize four 'Sobel' masks. For each region, four types of 'Sobel' masks are applied and the directional type with the largest summation of absolute 'Sobel' value is selected and we sum the absolute 'Sobel' values of all the directional types in Eq. (1). We use this information to decide the location of the homogeneous region [6].

$$Sum = Sum_h + Sum_v + Sum_d + Sum_{ad} \quad (1)$$

,where Sum_h,Sum_v,Sum_d and Sum_{ad} represent the sums of absolute 'Sobel' values in each direction, respectively. The region of block is considered as a homogeneous region when the Sum is less than the threshold (E_{th}). In order to consider the characteristics of an image, the value of E_{th} is computed as follows:

$$E_{th} = \frac{1}{N_x N_y} \sum_{(i,j)\in I_r} \sum_{k\in(h,v,d,ad)} (I_k \times c) \text{ , } c \in [0,1] \quad (2)$$

,where $I_h(i,j)$, $I_v(i,j)$, $I_d(i,j)$ and $I_{ad}(i,j)$ are the output values for the each directional 'Sobel' masks. N_x, N_y indicate the image size, respectively. We can control the proportion of the homogeneous region by adjusting the constant c.

Boundary map estimation: It is necessary for the boundary map to select an optimal window. According to the boundary information, an image can be divided into the homogeneous and object boundary region and direction of each pixel is calculated at boundary regions. Generally, the boundary regions are determined by the disparity information, which should be estimated. Therefore, we use a hierarchical framework to estimate the boundary and disparity information simultaneously. The pixel is boundary if

$$\frac{\max(D_1, D_2)}{\min(D_1, D_2)} > B_{th} \quad (3)$$
$$D_1 = |d' - d|, \; D_2 = |d - d'| \begin{bmatrix} horizon.\ d'(i,j)=d(i,j\pm k) \\ vert.\ d'(i,j)=d(i\pm k,j) \end{bmatrix}$$

,where D_1 and D_2 are the difference between disparity of neighboring pixels and k is a window size in each level. The region of block is decided as object boundary when the rate of difference is more than the threshold (B_{th}).

Disparity estimation: According to the information of edge direction and boundary, adaptive window size/ shape and the

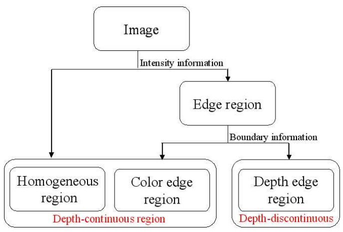

Fig. 2. Classification of Textureless and boundary region

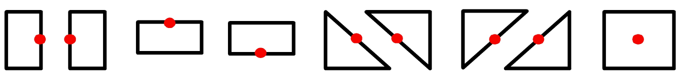

Fig. 3. Definition of 9 window models

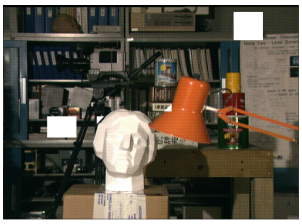

Fig. 4. Adaptive windows at each pixel in image

reference region are computed to execute accurate and fast disparity estimation. Fig. 4. shows the shape and size of window in $Tsukuba$ image. We can confirm that an optimal window is estimated at each pixel adequately. The sum of absolute difference (SAD) is used as the matching cost and the disparity is finally selected by Winner-Takes-All (WTA) method in Eq. (4).

$$E(x,y,d) = \frac{1}{||N||} \sum_{(x',y')\in N} |I_l(x',y') - I_r(x'+d,y')|$$
$$\hat{d}(x,y) = \arg\min_d E(x,y,d) \quad (4)$$

,where I_l and I_r are input images and N represent an optimal window for each pixel. Finally, we obtain final disparity map through the post processing schemes that filtered the outlier.

3.2. 3D model registration

Feature Matching: Feature detection methods, such as the Harris detector, are sensitive to the affine distortion of image. Therefore, they are not suitable to build feature sets in image that acquired by camera under various environments. SIFT feature[3] is widely used because it is invariant to affine transforms. These characteristics are suitable as the problem of image obtained by hand-held camera, such as different angles. SIFT feature algorithm is based upon finding locations within the scale space of an image which can be reliably extracted. Features are identified by detecting maxima and minima in the difference of gaussian (DOG) pyramid. A subpixel location, scale and orientation are associated with each SIFT feature. In order to achieve high specificity, a local feature is formed by measuring the local image gradients at many orientations in coordinates relative to the location, scale and orientation of the feature.

Camera pose estimation: According to assigning the estimated disparity at each feature, we can obtain 3D coordinate

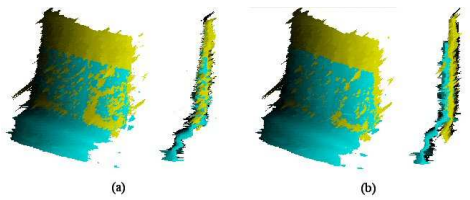

Fig. 5. Registered model without texture : (a) LMedS method (MER:0.253925) (b) Least square method (MER:4.394239)

in the features. Subsequently we reconstructed the 3D location of the pixels (X,Y,Z) in a real scene using the following equations:

$$X = \frac{x}{d}B \qquad Y = \frac{y}{d}B \qquad Z = \frac{f}{d}B \quad (5)$$

,where (x,y,d) are the SIFT feature location in image and disparity. B is the baseline distance and f is the focal length. To integrate the 3D data sets, we recover the R,T matrix. We need to 4 corresponding features to estimate the R,T matrix. Although the SIFT feature matching algorithm has low bad matching error rate, if the outliers are used in estimation of R,T matrix, the recovered camera motion is incorrect and it is impossible to register the model correctly. Moreover, if feature sets used in camera pose estimation is gathered in some region, it is also caused by error of R,T matrix. Therefore, we employ an LMedS approach. The number of sample are decided as following Eq. (6)

$$P = 1 - (1 - (1-\varepsilon)^p)^m \quad (6)$$

,where P is the probability of a good sample for LMedS. ε , p, m denote the ratio of false matched, the sample size and the number of sample required. To obtain 99.9% probability of a good samples, we choose the m = 108 with ε =0.5, p=4. MER (Matching Error Rate) is calculated on the sum of distance with transformed feature and reference feature for objective evalution, as shown in Fig. 5. The yellow and ciya models mean the 3D data set in each view. When we use just least square methods, MER is 4.394239 and each 3D data set does not registered accuately. However, when we employ the LMedS method, MER is 0.253925 and the quality of 3D model is improved.

4. EXPERIMENTAL RESULTS

We performed the experiments using personal computer with a Pentium IV 3.0GHz processor and 1GB RAM. A Bumblbee^{TM} stereo camera from Point Grey Research (PGR) [7] is used. This camera has 0.12m base-line and 6.0mm focal length at 640×480 image resolution.

4.1. Evaluation of proposed stereo algorithm

To evaluate the performance of our approach, we used four stereo image pairs with different contents and photographing environments. A $Tsukuba$ and $Sawtooth$ are provided on "www.middleburry.edu/stereo" with ground truth disparity maps[8]. The Lab and the $Studio$ image pairs are natural stereo sequence captured by Bumblbee^{TM}. The maximum disparity of the image pair is 15 pixels in a size of 640×480. The $Tsukuba$ and $Sawtooth$ image pairs are performed on the test bed of Scharstein's homepage [8] and the bad pixel ratios are 2.55 and 0.77 %. However, only a subjective evaluation is performed for the Lab and the $Studio$ image pairs because the ground truth disparity map is not provided. Fig.6

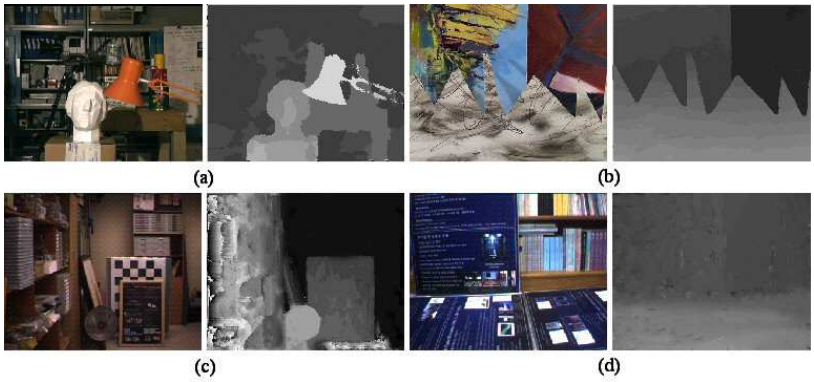

Fig. 6. Results of (a) Tsukuba (b) Sawtooth (c) Studio (d) Lab

is shown a test sequence pairs and results of disparity map. They show quantitative results for the stereo images using reconstruction of depth for 3D reconstruction.

4.2. Evaluation of 3D reconstruction

We reconstruct the model of corner region and bookshelf in our lab. The camera was moved freely at different portions. Fig.7 shows two selected views with some overlapping region from Lab sequences and total scene models are created and visualized in the OpenGL format. Fig.7 (a),(b) shows the two individual models and Fig.7 (c) shows the registered model. The proposed model reconstructing system can integrate two 3D models automatically without any initial estimation or user interaction. Experimental results show that the scenes are reconstructed seamlessly although there are some errors of dense stereo and SIFT feature matching. If the scene reconstruct under the large environments, the stereo camera with wider baseline and higher resolution is necessary.

5. CONCLUSION

In this paper, we presented a 3D scene reconstruction system. The stereo sequences are captured in natural views and the total model is obtained by estimating camera pose and registration method automatically. The proposed stereo matching algorithm finds precise disparity map in a stereo image pair for depth reconstruction. This method decreases the errors of disparity in object boundary or textureless region, and therefore it provide reliable disparities. We also reconstruct 3D model from the estimated disparity map and estimate the camera motion based on SIFT matching. Camera motion accuracy is affected by the accuracy of SIFT feature matching and their distribution. For this reason, we reduce the effect of false match using LMedS. It can be recovered the camera motion correctly and the proposed system create 3D scene model using multiple 3D data set. For future work, it needs to improve the quality of disparity map and multiview registration to obtain dense depth information. It is also planned to develop a complete full 3D modeling system from multi-view images.

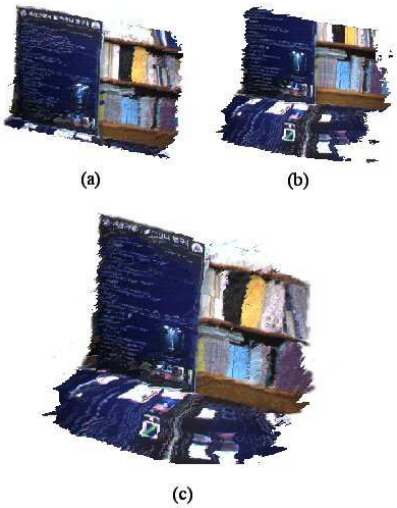

Fig. 7. (a),(b) 1st and 2nd view 3D model (c) Final 3D model

6. REFERENCES

[1] D. Scharstein and R. Szeliski: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, IJCV, Vol. 47 7-42,2002

[2] P. Besl and N. Mckay: A method for registration of 3-d shapes, IEEE PAMI, Vol.14(2) 239-256,1992

[3] D.G. Lowe: Object recognition from local scale-invariant features. IEEE Proc.ICCV, 1150-1157, 1999

[4] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,Artificial Intelligence, Vol. 78 87-119, 1995

[5] T. Kanade and M. Okutomi: A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiments, PAMI, vol. 16, no. 9 920-932, 1994

[6] Y. Kim, J. H. Lee, C. Park and K. Sohn.:MPEG-4 compatible stereoscopic sequence CODEC for stereo broadcasting, IEEE Trans.CE, vol. 51, no. 4, pp. 1227-1236, 1995

[7] Point Grey Research. http://www.ptgrey.com

[8] Scharstein's homepage. http://www.middleburry.edu/stereo

[9] S.Se, P.Jasiobedzki: Intant scene modeler for crime scene reconstruction, IEEE Proc. A3DISS, 2005

[10] S. Yun, D. Min, K. Sohn: Fast dense stereo matching using adaptive window in hierarchical framework, Proc. ISVC, 316-325, 2006