3D Reconstruction of Outdoor Scenes Using Structure from Motion and Depth Data

Keisuke Fujimoto
Hitachi Ltd.
1-280, Higashi-Koigakubo,
Kokubunji-shi
185-8601, Tokyo
keisuke.fujimoto.qb@hitachi.com

Takashi Watanabe
Hitachi Ltd.
1-280, Higashi-Koigakubo,
Kokubunji-shi
185-8601, Tokyo
takashi.watanabe.dh@hitachi.com

ABSTRACT

Recently, low-cost and small RGB-D sensors appear massively at the entertainment market. These sensors can acquire colored 3D models using color images and depth data. However, a limitation of the RGB-D sensor is that sunlight interferes with the pattern projecting LED. The sensor is most suitable only for indoor scenes. Some RGB-D sensors are available in outdoor scenes. However, the measurement range is limited because the light of LED spreads in all directions. In this research, we developed a novel measurement method for RGB-D sensors, which can measure shapes in outdoor scenes. This method uses several measurement data from multiple viewpoints, and estimates the shape and the sensor poses using Structure from Motion (SfM). However, a conventional image-based SfM cannot determine a correct scale. To determine the correct scale, our method uses the depth information that is obtained from partially acquired area which is near to the viewpoints. Then, our method optimizes the shape and the poses by a modified bundle adjustment with the depth information. It minimizes the reprojection error of the features in the acquired images and the depth error between the estimated model and the measurement depth. At last, our method generates dense point cloud using a multi-view stereo algorithm. Using both the acquired images and depth data, our method reconstructs the shape which locates out of measurement range in outdoor environment. In our experiment, we show that our method can measure the range up to 20 meters away by measuring from several viewpoints in the range of 5 meters using a RGB-D sensor in outdoor scenes.

Keywords

RGB-D sensor, Structure from Motion, Bundle Adjustment, Point Cloud, Multi View Stereo

1 INTRODUCTION

Recently, low-cost and small RGB-D sensors appeared. These sensors can acquire colored 3D models using color image and depth data. The 3D models of target objects can be reconstructed using depth data, so these RGB-D sensors has caused a surge in 3D perception research in the past few years. However, a limitation of the RGB-D sensors is that sunlight interferes with the pattern projecting LED. Therefore, these sensors are not available in outdoor scenes. Fig.1 shows measurement result in outdoor scenes. The black color represents the area where cannot be measured. As shown the bottom of Fig.1, almost all the area cannot be measured in outdoor scenes.

In this research, we developed a novel measurement method for RGB-D sensors in outdoor scene. This method uses several measurement data from multiple viewpoints, and estimates the shape and the sensor poses using Structure from Motion (SfM) and a scale adjustment method. SfM algorithms have a scale ambiguity problem. Then, our method uses scale information obtained from partially acquired area which is near to the viewpoints, and our method optimizes the shapes and poses by the modified bundle adjustment with the scale information. The bundle adjustment minimizes the reprojection error of the features in the acquired images and the depth error between measurement data and estimated data. At last, our method generates dense point cloud using a multi-view stereo algorithm. Our method obtains the correct scale and reconstructs the shape which locates out of measurement range in outdoor environment. In our experiment, we show that our method can measure the range up to 20m away with 700mm accuracy by measuring from several viewpoints in the range of 5m using a RGB-D sensor in outdoor scenes.

Figure 1: Measurement data from a RGB-D sensor in outdoor scenes. The top shows acquired color image. The bottom shows color mapped depth image. The black colored area represents out of range. The area which is near to the viewpoint is acquired.

2 RELATED WORK

Recently RGB-D sensors have become very popular in the area of Simultaneous Localization and Mapping (SLAM) [Henry10][Endres12]. The major advantage of these sensors is that they provide a rich source of 3D information at relatively low cost. Unfortunately, in outdoor scenes, sunlight affects the measurement result of these sensors, so these sensors are limited to use in indoor scenes.

SfM [Davison07][Snavely07] computes camera poses and 3D shapes of scenes as 3D point cloud using only corresponding feature points in each 2D image. These image-based approaches can be applied to outdoor scenes. To find the correspondences between images, features such as corner points are tracked from one image to another image. However, the image-based method has a scale ambiguity problem [Hartley00]. It is impossible to recover the absolute scale of the scene.

To avoid the scale ambiguity problem, sensor fusion approaches are proposed. Pollefeys et al. [Pollefeys08] integrated captured image sequences with GPS data to correct the scale. Nutzi et al. [Nutzi10] proposed to merge the output of a SfM algorithm with IMU (Inertial Measurement Unit) measurements in an Extended Kalman Filter holding the scale as an additional variable in the state. However, these sensor fusion approaches need additional sensors.

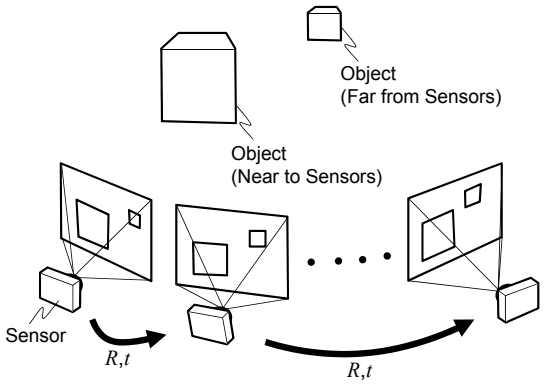

Figure 2: Measurement from several viewpoints. In this case, the near object and the far object exist.

In contrast to these previous works, our method obtains the correct scale and 3D models in outdoor scenes using one RGB-D sensor only without another sensor.

3 PROPOSAL METHOD

Our method estimates the 3D shape of outdoor scenes by a RGB-D sensor using feature points and scale of which distance is acquired. At first, we measure color images and depth images from several viewpoints moving the RGB-D sensor as shown in Fig.2. Next, using the image-based SfM which uses the feature in the acquired images, the 3D shape can be measured robustly in outdoor scenes. However, the correct scale is unknown. Then, we adjust the scale using depth data obtained from acquired area which is near to viewpoints. And our method minimizes the reprojection error and the depth error. The reprojection error is the 2D distance between the features and projected points as shown in Fig.3. The depth error is the distance between the estimated depth and the measured depth as shown in Fig.4. At last, we generate dense point cloud using multi-view stereo.

3.1 Initialization

At first, sensor poses and 3D shapes are initialized using image-based SfM. The SfM computes camera poses and 3D shapes as 3D point cloud using only corresponding feature points in each view. To find the correspondences between images, features such as corner points are tracked from one image to another image. One of the most widely used feature detectors is the SIFT (Scale-invariant feature transform) [Lowe04]. Given a set of corresponding points in two or more images, camera matrices and 3D coordinate of the features are estimated by minimizing reprojection error. In this way, the relative 3D structure of the target scene can be estimated. And we will use the provisional scale which is obtained this initialization step. In our implementation, we used the VisualSFM Software [Wu11].

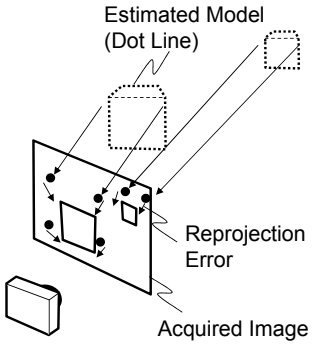

Figure 3: Reprojection Error. It is geometric error corresponding to the 2D image distance between the projected point and the measured point. The error is shown as length of the arrows from projected points.

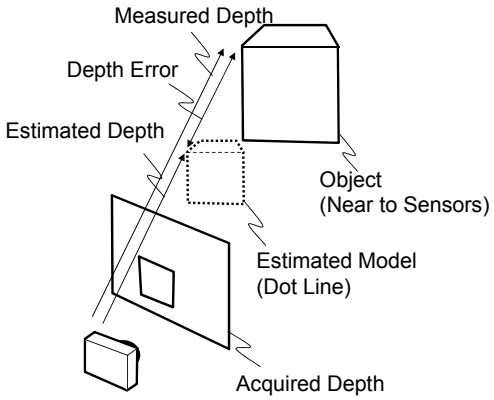

Figure 4: Depth Error. It is a distance between a estimated depth and a measured depth. The estimated depth is obtained by reconstructed shape.

3.2 Scale Adjustment

In this section, we explain the way to adjust scale using depth data obtained from acquired area which is near to viewpoints. Let d_{ij} be the depth of features j in images i, which is obtained by the estimated 3D structure of the target scene with the provisional scale in the above section. Let d'_{ij} be the depth which is obtained from measurement data of feature points. Then the following relation holds

$$d'_{ij} = sd_{ij} \tag{1}$$

where the scale s is the ratio between the provisional estimated scale and the measured scale. However, typically, the above condition does not necessarily satisfy because of the estimation error of SfM and the measurement error of RGB-D sensor. In our method, we compute the optimal scale s^* by minimizing follow equation:

$$s^* = \arg\min_s \sum_i \sum_j (d'_{ij} - sd_{ij})^2. \tag{2}$$

The scale s^* is given by

$$s^* = \frac{\sum_i \sum_j d'_{ij} d_{ij}}{\sum_i \sum_j d_{ij}^2}. \tag{3}$$

Using the scale s^*, the 3D coordinate q of features is updated by

$$q \leftarrow s^* q \tag{4}$$

and the camera pose $\boldsymbol{T}$ is

$$\boldsymbol{T} \leftarrow s^* \boldsymbol{T} \tag{5}$$

where $\boldsymbol{T}$ is 3D vector which represents camera's position.

3.3 Optimization

In this section, we show our optimization approach using a modified bundle adjustment. Conventional bundle adjustments minimize the reprojection error to estimate camera poses and 3D shapes [Triggs00]. In contrast to these bundle adjustment methods, our modified bundle adjustment uses the 3D positions of features and the depth data acquired by RGB-D sensor. In our research, we assume a pinhole camera model. In this model, the mapping from 3D coordinates of points in space to 2D image coordinates can be represented in homogeneous coordinates. Let $\boldsymbol{q}$ be representation of the 3D point in homogeneous coordinates, and let (u', v') be representation of the projected point in the pinhole camera. Then the following relation holds

$$\begin{pmatrix} u'_{ij} \\ v'_{ij} \\ 1 \end{pmatrix} \propto \boldsymbol{P}_i \begin{pmatrix} \boldsymbol{q}_j \\ 1 \end{pmatrix} \tag{6}$$

where $\boldsymbol{P}$ is a 3×4 camera matrix which is given by combining a camera calibration matrix $\boldsymbol{K}$, rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{T}$. The camera matrix is given by

$$\boldsymbol{P}_i = \boldsymbol{K} \begin{bmatrix} \boldsymbol{R}_i & \boldsymbol{T}_i \end{bmatrix} \tag{7}$$

where the camera calibration matrix $\boldsymbol{K}$ is an upper triangular matrix that is consist of the focal length and the principal point, the rotation matrix $\boldsymbol{R}$ is a 3×3 matrix that represents camera orientation, and the translation vector $\boldsymbol{T}$ is a three vector that represents camera's position. The image points (u', v') given by

$$u'_{ij} = \frac{\boldsymbol{P}_i^{(1)} \begin{bmatrix} \boldsymbol{q}_j^T & 1 \end{bmatrix}^T}{\boldsymbol{P}_i^{(3)} \begin{bmatrix} \boldsymbol{q}_j^T & 1 \end{bmatrix}^T} \tag{8}$$

$$v'_{ij} = \frac{\boldsymbol{P}_i^{(2)} \begin{bmatrix} \boldsymbol{q}_j^T & 1 \end{bmatrix}^T}{\boldsymbol{P}_i^{(3)} \begin{bmatrix} \boldsymbol{q}_j^T & 1 \end{bmatrix}^T} \tag{9}$$

where $\boldsymbol{P}^{(k)}$ is k-th row of the camera matrix $\boldsymbol{P}$. The reprojection error between project points and observed points E_1 is given by

$$\begin{aligned} E_1 &= \frac{1}{2}||\boldsymbol{e}_1||^2 \\ &= \frac{1}{2} \sum_i \sum_j ((u_{ij} - u'_{ij})^2 + (v_{ij} - v'_{ij})^2). \end{aligned} \tag{10}$$

 ISBN 978-80-86943-66-4

where (u_{ij}, v_{ij}) are the measured feature points, and $\boldsymbol{e}_1$ is residual error. Then, we explain depth error between estimated depth and measured depth. In the pinhole camera model, the depth of features are given by

$$d'_{ij} = \boldsymbol{P}_i^{(3)} \begin{bmatrix} \boldsymbol{q}_j^T & 1 \end{bmatrix}^T. \quad (11)$$

Then, the depth error is given by

$$\begin{aligned} E_2 &= \frac{1}{2}||\boldsymbol{e}_2||^2 \\ &= \frac{1}{2}\sum_i \sum_j (d_{ij} - d'_{ij})^2 \end{aligned} \quad (12)$$

The total error from reprojection error and depth error is given by

$$E = rE_1 + (1-r)E_2 \quad (13)$$

where r is the weight parameter between the 2D distance on the images and the 3D distance in the reconstructed structure. Next, we explain the way to minimize the error. To solve non-linear least squares problems, the Levenberg-Marquardt (LM) algorithm is most widely used. The method interpolates between the Gauss-Newton algorithm and the gradient descent method. This method updates parameters $\boldsymbol{x}$ with

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1}\boldsymbol{g} \quad (14)$$

where $\boldsymbol{I}$ is identity matrix, $\boldsymbol{H}$ is hessian matrix, and $\boldsymbol{g}$ is gradient vector. λ is damping factor which adjusts the step size at each iteration. Using residual error $\boldsymbol{e}_1$ in (10) and $\boldsymbol{e}_2$ in (12), the total error (13) is

$$\begin{aligned} E &= \frac{r}{2}||\boldsymbol{e}_1||^2 + \frac{1-r}{2}||\boldsymbol{e}_2||^2 \\ &= \frac{1}{2}\begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix}^T \begin{pmatrix} r\boldsymbol{I} & 0 \\ 0 & (1-r)\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix} \end{aligned} \quad (15)$$

The gradient vector and the approximated hessian matrix is given by

$$\boldsymbol{g} = \begin{pmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{pmatrix}^T \begin{pmatrix} r\boldsymbol{I} & 0 \\ 0 & (1-r)\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{pmatrix} \quad (16)$$

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{pmatrix}^T \begin{pmatrix} r\boldsymbol{I} & 0 \\ 0 & (1-r)\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{pmatrix} \quad (17)$$

where the matrix $\boldsymbol{J}$ is jacobian. Using residual error, the jacobian is given by

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}_1 \\ \boldsymbol{J}_2 \end{pmatrix} = \begin{pmatrix} \frac{d\boldsymbol{e}_1}{d\boldsymbol{x}} \\ \frac{d\boldsymbol{e}_2}{d\boldsymbol{x}} \end{pmatrix} \quad (18)$$

Note that we gave the approximated hessian matrix. The hessian matrix is a symmetric positive definite matrix and the solution to (14) can be obtained.

Unfortunately, due to the large number of unknowns contributing to the minimized reprojection error and depth error, the computational cost of LM algorithm becomes high. The cost depends on computing cost of inverse of the hessian matrix. To solve this problem, the Sparse Bundle Adjustment (SBA) [Triggs00] takes advantage of the sparse structure of the hessian matrix. The SBA solves huge minimization problems over many thousands of variables within seconds on a standard PC for the image-based SfM. Fortunately, in our method, it is easy to apply sparse technique to deal with both reprojection error and depth error. In the SBA formula, replacing our jacobian matrix and our residual vector to original SBA's jacobian matrix and residual vector, this problem can be solved.

3.4 Generating Dense Point Cloud

Our algorithm uses only feature points, so reconstructed 3D structure is sparse. Then, at last, we apply multi-view stereo algorithm [Seitz06] which generates dense point cloud. The multi-view stereo algorithm uses the camera poses estimated in above section. Using the poses, our method generates dense point cloud whose scale is correct. In our research, we apply The Patch-based Multi-view stereo (PMVS) [Furukawa10] to generate dense point cloud.

4 EXPERIMENTAL RESULT

In this section, we show our experimental result. We test our method using Kinect v2 in outdoor scenes. We compared our result with ground truth data which is acquired by high accuracy Laser Range Finder (LRF). As a LRF, we choose a RIEGL VZ-400. The accuracy of VZ-400 is about 5mm at 100m range.

We measured from 70 viewpoints in the range of 5 meters using RGB-D sensor in outdoor scenes. The top of Fig.1 shows the acquired RGB image, and the bottom of Fig.1 shows the acquired depth map. In the bottom of Fig.1 black color represents the area where cannot be measured. The top of Fig.5 shows the result of the estimated depth map which is estimated by our method, and the bottom of Fig.5 shows the ground truth. As the Fig.5, the resolution of the depth map in our method is lower than the ground truth. The reason is that the number of points generated PMVS is less than LRF data. For visibility, we determined the resolution according to the number of points. As shown in Fig.5, our result can estimate the depth out of RGB-D sensor's range. Fig.6 shows the estimated accuracy. The line depicts the root mean square (RMS) error. RMS is computed using the difference between our result and ground truth around each place. The graph shows our method can estimate the range up to about 20 meters away with 700 mm accuracy. And accuracy is higher in the location close to the sensor. The running time of our scale adjustment and bundle adjustment is 16.5s and 60MB memory is used in this case.

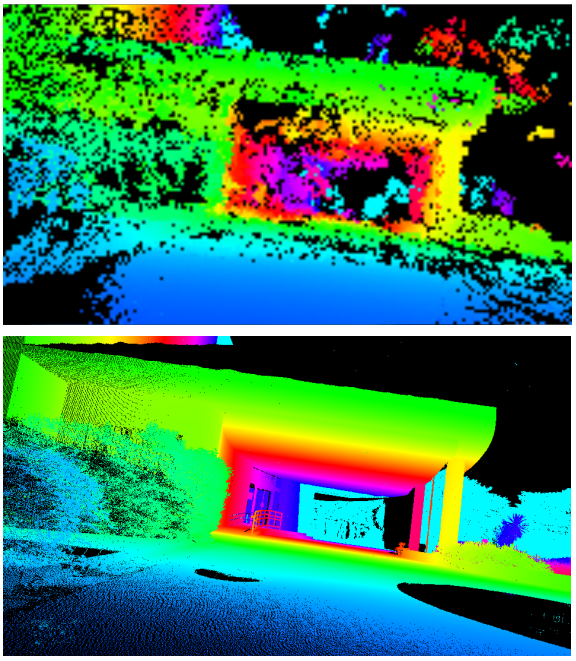

Figure 5: Color mapped depth image. The top shows estimated depth image using our method. The bottom shows ground truth.

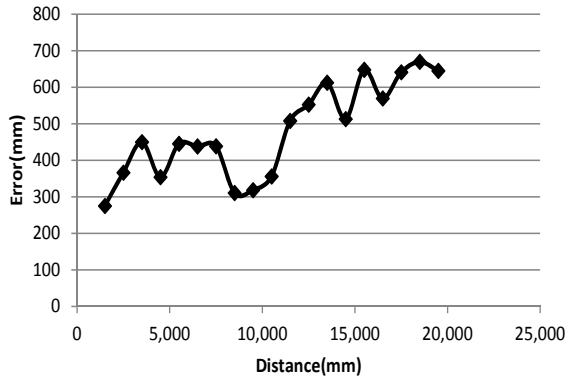

Figure 6: Accuracy of our method. The line shows the root mean square error which is computed using the difference between our estimated depth and ground truth.

Above described scene is measured in the shade, so the acquired image is not bright as shown Fig1. Then, we measured from 90 viewpoints in the range of 5 meters using RGB-D sensor in the direct sunlight (not against sun). The top of Fig.7 shows the acquired RGB image, the middle of Fig.7 shows the generated depth map, and the bottom of Fig.7 shows the ground truth. Fig.6 shows the estimated accuracy. The graph shows our method can estimate the range up to about 20 meters away with 500 mm accuracy. The result shows that our method can work in the bright environment. In this case, the running time of our scale adjustment and bundle adjustment is 69.2s and 135MB memory is used.

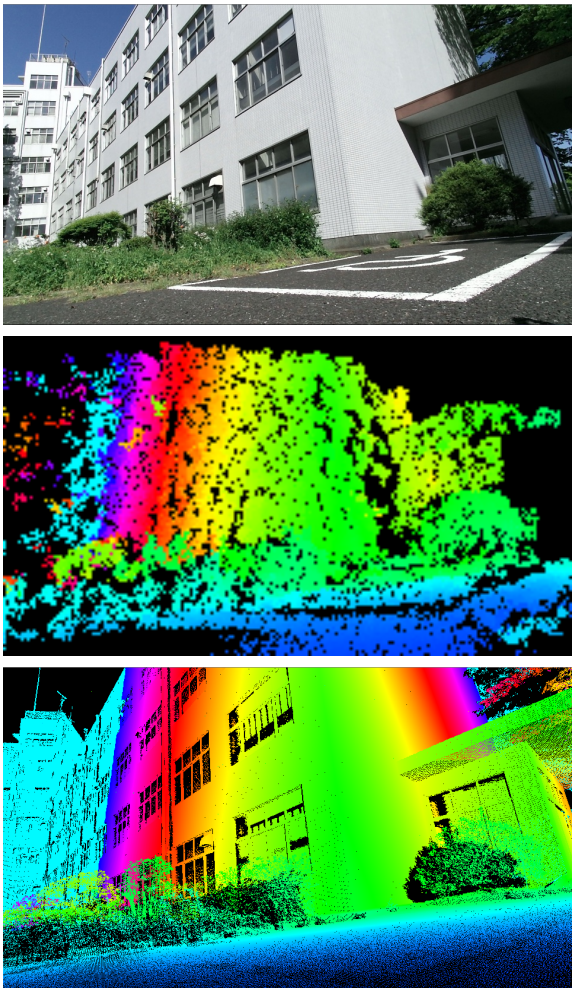

Figure 7: Color mapped depth image in the bright scene. The top shows color image, the middle shows estimated depth image. The bottom shows ground truth.

5 DISCUSSION

As Fig.5 and Fig.7, the result shows the our depth maps became noisy. The reason is correspondence error between color images in the process of MVS described in Sec.3.4. For example, error models are generated in the sky (Fig.5). And our method cannot reconstruct the ground surface model although the ground exists in the captured area. In our method, MVS algorithm generates dense 3D points using correspondence of image patches, so the method cannot make correspondence on texture-less area. Therefore, it is difficult to reconstruct the model of these texture-less areas correctly.

As shown the graph, the accuracy from 4m to 6m became low. In this area, only ground surface exists, so the error became large. The reason of this low accuracy is correspondence error as described above.

Next, we will discuss the accuracy of our result. In our method, we use PMVS algorithm for generating dense point cloud. However, the accuracy of our re-

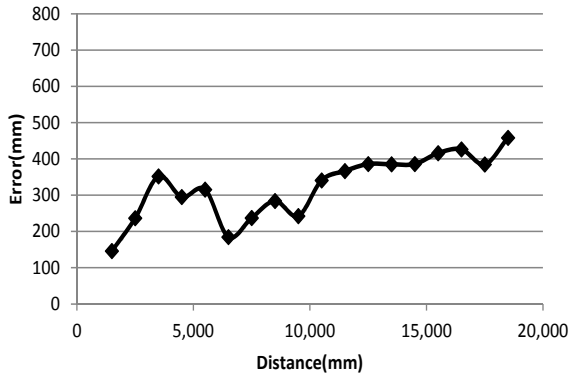

Figure 8: Accuracy of our method in the bright scene.

sults was less quality than the original results of PMVS [Furukawa10]. The accuracy of reconstructed models depends on the accuracy of obtained camera poses in PMVS. In our method, the camera poses are estimated by our bundle adjustment, so the accuracy is chiefly affected by the bundle adjustment. One of the reasons for the low accuracy is that we did not use robust algorithm in our current implementation, in particularly the estimated scale influences the accuracy because the error becomes larger as the object leaves more the sensor position. The scale is determined by the ratio of estimated distance by the features to acquired distance directly.

6 CONCLUSION

We developed a measurement method which reconstructs 3D shapes in outdoor scenes. This method uses measurement data acquired from multiple viewpoints, and estimates the 3D shape and the sensor poses using reprojection error and depth error. Our method obtains the correct scale in the area that could not be measured directly from RGB-D sensor using both the acquired color images and depth images. In our experiment, we show that our method can measure the range up to 20 meters away with 700 mm accuracy by measuring from several viewpoints in the range of 5 meters using RGB-D sensor in outdoor scenes.

As a future work, we plan to apply robust approaches and to compare the accuracy of our method with another approach that can determine the scale. And we will extend to on-line algorithm using local bundle adjustment and video-based real-time MVS.

7 REFERENCES

[Davison07] A. J. Davison, N. D. Molton, I. Reid, and O. Stasse, “MonoSLAM: Real-time Single Camera SLAM” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 29, No. 6, pp. 1052–1067, 2007.

[Endres12] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, W. Burgard, “An Evaluation of the RGB-D SLAM System,” In Proc of the IEEE Int’l Conf. on Robotics and Automation, 2012.

[Furukawa10] Y. Furukawa and J. Ponce, “Accurate, Dense, and Robust Multi-View Stereopsis,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 32, No. 8, pp. 1362–1376, 2010.

[Hartley00] R. Hartley and A. Zisserman, “Multiple View Geometry in Computer Vision,” Cambridge University Press, 2000.

[Henry10] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, “RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments,” in Proc. Int’l Symp. Experimental Robot, 2010.

[Lowe04] D. G. Lowe, “Distinctive Image Features from Scale-invariant Keypoints,” Int’l J. of Computer Vision, Vol. 60, No. 2, pp.91–110, 2004.

[Nutzi10] G. Nutzi, S. Weiss, D. Scaramuzza, and R. Siegwart, “Fusion of IMU and Vision for Absolute Scale Estimation in Monocular SLAM,” Journal of Intelligent Robotic Systems, vol. 61, pp. 287–299, 2010.

[Pollefeys08] M. Pollefeys, D. Nister, J. M. Frahm, A. Akbarzadeh, et al, “Detailed Real-time Urban 3D Reconstruction from Video,” Int’l J. of Computer Vision, Vol. 78, No. 2–3, pp.143–167, 2008.

[Seitz06] S. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, “A Comparison and Evaluation of Multi-view Stereo Reconstruction Algorithms,” In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition , Vol. 1, pp. 519–526, 2006.

[Snavely07] N. Snavely, S. Seitz, R. Szeliski, “Modeling the World from Internet Photo Collections,” Int’l J. of Computer Vision, Vol. 80, No. 2, pp. 189–210, 2008.

[Triggs00] B. Triggs, P. McLauchlan, R. Hartley, A. Fitzgibbon, “Bundle Adjustment - a Modern Synthesis,” LNCS, Springer Verlag, Vol. 1883, pp. 298–375, 2000.

[Wu11] C. Wu, “VisualSFM: A Visual Structure from Motion System,” `http://homes.cs.washington.edu/~ccwu/vsfm`.