Indoor Scene Reconstruction Using a Projection-Based Registration Technique of Multi-view Depth Images*

Sehwan Kim and Woontack Woo

GIST U-VR Lab, Gwangju 500-712, S. Korea
{skim, wwoo}@gist.ac.kr

Abstract. A novel registration method is presented for 3D point clouds, acquired from a multi-view camera, for scene reconstruction. In general, conventional registration methods require a high computational complexity, and are not robust for 3D point clouds with a low precision. To remedy these drawbacks, a projection-based registration is proposed. Firstly, depth images are refined based on temporal property by excluding 3D points with large variations, and spatial property by filling holes referring to neighboring 3D points. Secondly, 3D point clouds are projected to find correspondences and fine registration is conducted through minimizing errors. Finally, final colors are evaluated using colors of correspondences, and a 3D virtual environment is reconstructed by applying the above procedure to several views. The proposed method not only reduces computational complexity by searching for correspondences on an image plane, but also enables an effective registration even for 3D points with a low precision. The generated model can be adopted for interaction with a virtual environment as well as navigation in it.

1 Introduction

Image-based reconstruction of a real environment plays a key role in providing visual realism while allowing a user to navigate in and interact with a Virtual Environment (VE). The visual realism of reconstructed models encourages a user to interact with the VE proactively. Furthermore, the generated VE allows the user to manipulate augmented objects while walking around the VE by removing/augmenting virtual objects from the viewpoint of Mediated Reality (MR) [1]. Unlike the methods using modeling tools or ones based on active range sensors, image-based modeling methods not only preserve realism but also provide a simple modeling process. Especially, off-the-shelf multi-view cameras enable to generate models more easily. Thus, a delicate registration is required to register 3D point clouds to reconstruct models. Note that unlike 2D registration, 3D registration may generate interactive image-based models.

Until now, various registration methods have been proposed. ICP (Iterative Closest Point) has been widely used, and Color ICP was proposed by Johnson [2][3]. Another approach is to project data sets onto an image plane and to pair points [4][5]. Color or intensity image was used to improve matches by using the intensity gradients on the projection plane, and back-projecting to get a 3D point pair [6]. Especially, Pulli

* This work was supported in part by MIC through RBRC at GIST, and in part by CTRC at GIST.

Y.-S. Ho and H.J. Kim (Eds.): PCM 2005, Part II, LNCS 3768, pp. 759 – 771, 2005.

adopted a projective registration method employing planar perspective warping, and Bernardini et al. searched the neighborhood and paired locations that maximize the cross-correlations [7][8]. Sharp et al. defined invariant features to improve ICP, and Fisher applied projective ICP to Augmented Reality [9][10]. On the other hand, Nishino et al. presented an optimization method based on M-estimator [11]. However, most methods rely on accurate equipments and require much time for modeling. If 3D points have large error variations, results are not reliable. Stereo cameras are usually exploited for an object modeling instead of an indoor scene reconstruction.

To address these weaknesses, a projection-based registration is proposed. Firstly, a depth image is refined based on the spatio-temporal property of 3D point clouds using adaptive uncertainty region. Secondly, correspondences are searched for through the modified KLT feature tracker for projected 3D point clouds. Then, 3D point clouds are fine-registered by minimizing errors. Finally, each 3D point is evaluated referring to correspondences, and a new color is assigned. Thus, we reconstruct an indoor environment by applying the above procedure to several views. The proposed method is carried out effectively even if the precision of 3D point cloud is relatively low by using the correlation of features with the neighborhood. Thanks to 2D-based registration, computational complexity is also low. Also, the proposed method makes 3D reconstruction easy just by placing a multi-view camera at several positions.

The paper is organized as follows. In Chapter 2, 3D reconstruction of an indoor scene is explained. After experimental results are analyzed in Chapter 3, conclusions and future work are presented in Chapter 4.

2 3D Reconstruction of an Indoor Scene

2.1 Depth Image Refinement

In general, passive techniques use images generated by the light reflected by objects. However, disparity estimation results in inherent stereo mismatching errors, usually at depth discontinuities and on homogeneous areas. These errors cause poor registration results. Thus, unreliable areas should be eliminated before registration. In this regard, a depth image is refined by spatio-temporal property. In the first step, erroneous 3D points are removed using the temporal property that the erroneous 3D points change dramatically in 3D space with time. In the second step, holes are filled by means of the spatial property that there is a spatial correlation among neighboring pixels. Fig. 1 shows a flow diagram for 3D reconstruction. Overall procedure and its projection-based registration part are described in Fig. 1(a) and Fig. 1(b), respectively.

In the depth map of a static scene, depth variation of each pixel is modeled as a Gaussian distribution. After investigating the depth value of each pixel, we get rid of the pixels whose depth variation is larger than the threshold value, Th_i, for the i^{th} pixel.

$$\sigma_i > \alpha Th_i(x_c, y_c, z_c) \tag{1}$$

where σ_i represents a standard deviation for the i^{th} pixel and α denotes a scale factor. $(x_c\ y_c\ z_c)^T$ is a translation vector from the optical center of a camera to the center of uncertainty region.

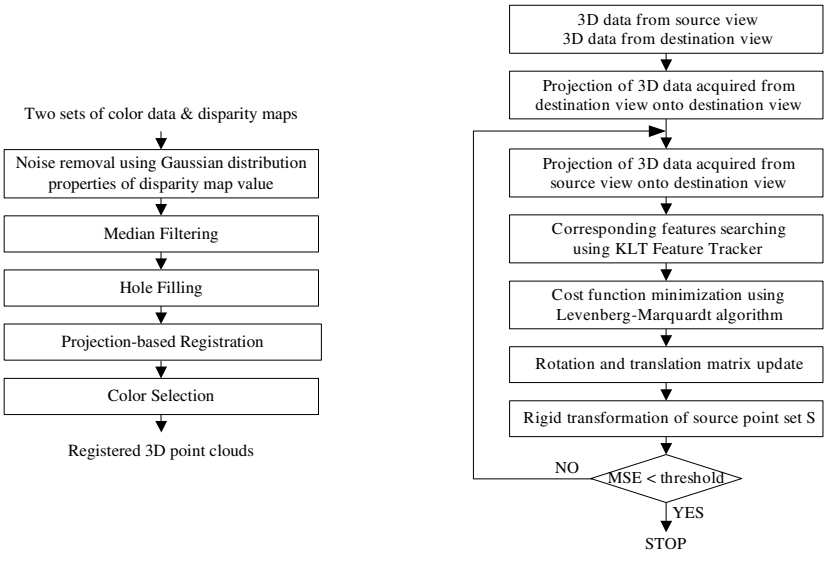


Fig. 1. Flow diagram for 3D reconstruction (a) overall procedure (b) projection-based registration part of (a)

The adaptively changing error bound appears to be ellipsoidal in a 3D space, and we call it as *Adaptive Uncertainty Region (AUR)*. The AURs are determined based on the error tolerance of each axis [12]. Its uncertainty distance, for each axis, is modeled as a Gaussian distribution. The Gaussian distributions increase linearly with the distance along x or y axis, and increase monotonically with the distance along z axis, respectively. The ellipsoid is also rotated with respect to the optical center reflecting the direction of ray that originates at the camera center and passes though each pixel.

$$\frac{x^2}{(\Delta x)^2}+\frac{y^2}{(\Delta y)^2}+\frac{z^2}{(\Delta z)^2}=1 \tag{2}$$

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = R_1 R_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} \tag{3}$$

$$R_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & z_c/d & y_c/d \\ 0 & -y_c/d & z_c/d \end{pmatrix} \quad R_2 = \begin{pmatrix} d & 0 & x_c \\ 0 & 1 & 0 \\ -x_c & 0 & d \end{pmatrix} \quad d = \sqrt{{y_c}^2 + {z_c}^2}$$

where $(x' y' z')^T$ is a final uncertainty region in terms of 3D coordinates of a scene with respect to the optical center. Δx, Δy and Δz represent uncertainty distances along each axis.

Then, Median filter is applied to remove spot noises. Finally, hole filling is required for the holes, generated during the above step, and homogeneous areas. That is, spatial property for a current 3D point, i.e. spatial correlation among 3D points of neighboring pixels, is exploited.

2.2 Initial Registration

The depth image refinement removes inherent stereo mismatching errors, and reduces the error bound of 3D point cloud. However, the precision of 3D point cloud is still low for registration. That is, the registration method exploiting the conventional ICP, which employs the shortest distance, is inappropriate. Thus, a projection-based registration method is proposed by effectively carrying out a pairing process that searches for correspondences between 3D point clouds acquired from two views, destination and source views. As shown in Fig. 6, we let a multi-view camera be located around a wall while acquiring partial surfaces successively. Destination and source views mean the views of a camera at the previous and current positions, respectively.

In the *initial registration* phase, a rigid-body transformation is applied to 3D points of corresponding features to estimate the poses of a multi-view camera [13]. Actually, any method, such as semi-automatic one [7], can be used for the initial registration. Thus, 3D point clouds are initially registered. Fig. 2 shows the projection of each 3D point cloud acquired from destination and source views onto the destination view after the initial registration. Fig. 2(a) and Fig. 2(b) are projection results of 3D point clouds, which are acquired from the destination and source views, onto the destination view. A constant value is assigned to unprojected pixels to differentiate them from projected ones. It should be noted that the projection of 3D point cloud acquired from the source view causes self-occlusion. This is eliminated based on the rays that originate at the camera center and pass though each pixel. Theoretically, Fig. 2(b) should be a subset of Fig. 2(a). However, discrepancies exist due to the errors in disparity estimation, camera calibration, etc. Therefore, an accurate geometric relationship between two views is found by minimizing distance errors between correspondences. Thus, *fine registration* should be employed to compensate the errors.

(a) (b)

Fig. 2. Projected images (a) projection of 3D point cloud of destination view onto its own view (b) projection of 3D point cloud of source view onto destination view

In general, projection of 3D point cloud acquired from the source view onto the destination view results in floating-point numbers. Thus, there occur unprojected

pixels. These generate false alarms when corresponding features are searched for through a modified KLT [14]. In this case, linear interpolation is useless since it makes the object boundaries smoothed. On the other hand, bi-linear interpolation cannot be used since the relation of two images is unknown.

Not only to preserve an original image but also to remove unprojected pixels, a two-step integer mapping is presented as shown in Fig. 3. Firstly, a search range is set to −0.5~0.5 along x and y axes for a grid point. The color of grid point is decided by relative distances with neighboring pixels. Secondly, the search range is expanded to −1.0~1.0 and a similar procedure is conducted for grid points which do not include any projected point at the first step. Fig. 4 shows the results. Fig. 4(a) is an enlarged part of Fig. 2(b), and Fig. 4(b) is the result after applying the mapping.

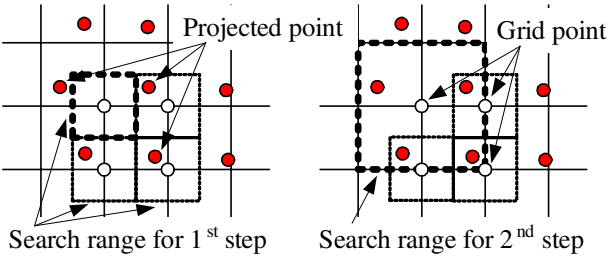


Fig. 3. Two-step integer mapping

(a) (b)

Fig. 4. Two-step integer mapping results (a) before (b) after

2.3 Projection-Based Registration for Partial 3D Point Clouds

Correct pairing plays a key role in accurate registration to compensate the errors induced by the disparity estimation, camera calibration, etc. In *fine registration phase*, corresponding features, on 2D image plane instead of 3D space, are employed. That is, a feature-based approach is proposed by exploiting corresponding features within the overlapping area.

Let us consider two textured surfaces that are already in the initial alignment. If you render the acquired 3D surfaces, as they would be seen from an arbitrary viewpoint, the resulting 2D color images are also in alignment. Each point on the source surface projects to the same pixel as its corresponding point on the destination surface. If we could move the partial surface of source view such that its projected image aligns well with the image of the other surface, we could be confident that visible

surface points projecting to the same pixel correspond to the same point on the object surface. We can then find good point pairs by pairing points that project to the same pixel.

We apply our registration method to align two partial surfaces by iteratively adjusting extrinsic calibration parameters of source view with respect to destination view. In other words, we apply a Euclidean transformation T: $\Re^3 \rightarrow \Re^3$ to the source surface. The destination surface, S_{Dst}, is projected onto its own image plane and features, f_{Dst}, are extracted in the projected image plane. On the other hand, at each iteration, the source surface, S_{Src}, is projected onto the destination image plane and corresponding features, f_{Src}', are searched for. This is illustrated in Fig. 5.

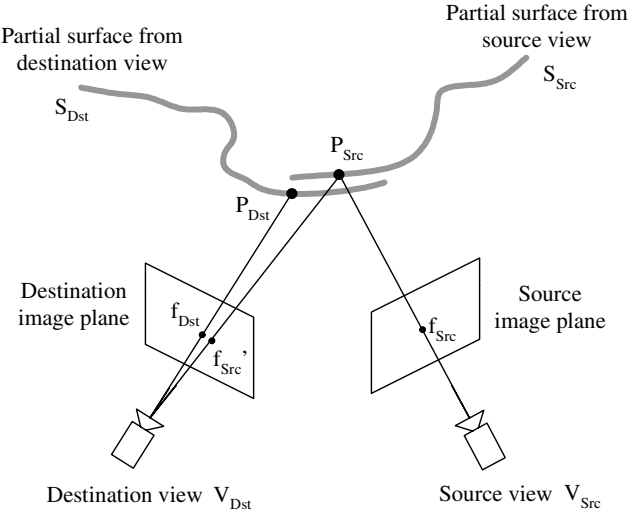


Fig. 5. Selection of corresponding features

For each feature f_{Dst} of the destination image, the corresponding feature f_{Src}' of the source image is found in the neighborhood of the same position as f_{Dst} using the modified KLT feature tracker. P_{Dst} and P_{Src} are 3D points of f_{Dst} and f_{Src}, respectively. And c_{Dst} and c_{Src} are RGB color components of f_{Dst} and f_{Src}, respectively.

Firstly, features are extracted over the overlapping area, Ω, in the destination image. The modified KLT feature tracker is adopted to extract feature corners that are robust to noise and can be tracked well. For this, local autocorrelation and eigenvalues are computed. After features are extracted, S_{Src} is projected onto the destination image plane using the same calibration parameters as the projection of S_{Dst}. Then, correspondences are searched for in the projected source image using cross-correlation in sub-pixel unit. However, there may occur some mismatches that should be filtered out. In order to guarantee correct paring, RANSAC is applied at each iteration [15]. By exploiting RANSAC, we can eliminate outliners and obtain only correct pairs between source and destination views.

Projecting S_{Src} onto the destination image plane produces an image I_{Src}'. Then, we can define a cost function measuring the mismatch between I_{Src}' and destination image I_{Dst}.

$$L = \sum_{i=0}^{N-1} \kappa_1 \left\{ \left(1 - \frac{\left\| f_{Dst,i} - f_{Src,i}' \right\|}{Dist_{max}} \right) \left\| f_{Dst,i} - f_{Src,i}' \right\|^2 + \kappa_2 \left\| c_{Dst,i} - c_{Src,i}' \right\|^2 \right\} \tag{4}$$

where κ_1 is described as follows to exclude the pair whose distance in 3D space exceeds a preset threshold *Th*. In other words, the pair, whose depth difference is large, is not included.

$$\kappa_1 = \begin{cases} 1 & if \ \left\| P_{Dst} - P_{Src} \right\| < Th \\ 0 & o/w \end{cases} \tag{5}$$

κ_2 is a weighting factor for color information and *N* denotes the number of features. And $\| \ . \ \|$ and $Dist_{max}$ represent the norm and a maximum distance between f_{Dst} and f_{Src}', respectively.

In summary, we search for correspondences and use them to define a total cost function within the overlapping area. By minimizing the cost function, a final pose of the source view is estimated. That is, we can estimate the pose of source view $\{R_{Src}, T_{Src}\}$, with respect to the pose of destination view $\{R_{Dst}, T_{Dst}\}$ through minimizing the errors on *N* feature points as follows.

$$\begin{aligned} &Given\ two\ sets\ of\ corresponding\ points, \\ &Find\ \{R_{Src}, T_{Src}\}\ w.r.t\ \{R_{Dst}, T_{Dst}\} \\ &such\ that\ \arg\min_{\{R_{Src}, T_{Src}\}} L \end{aligned} \tag{6}$$

The total error is minimized through Levenberg-Marquardt non-linear optimization algorithm.

By employing the proposed method, the correspondences between destination and projected source images can be found. Therefore, correspondences between destination and original source images can be established after applying RANSAC. Usually, it is hard to find correspondences between two views with a wide baseline. However, if depth image and initial camera pose are available, corresponding features can be extracted effectively.

2.4 Surface Reconstruction of Registered 3D Point Clouds

After pose estimation, trimming and color selection are required. We obtain the correspondences through a registration process. Then, final 3D coordinates are calculated by a linear triangulation method [16].

Color adjustment is also required to consider changes in lighting conditions depending on the camera position.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \left(v \times \begin{bmatrix} R_{Dst} \\ G_{Dst} \\ B_{Dst} \end{bmatrix} + u \times \begin{bmatrix} R_{Src} \\ G_{Src} \\ B_{Src} \end{bmatrix} \right) / (u + v) \tag{7}$$

where u and v are distances from left and right borders of the overlapping area to the current pixel, respectively. R_{Dst} (*or* G_{Dst}, B_{Dst}) and R_{Src} (*or* G_{Src}, B_{Src}) are red (or green, blue) of a current pixel for both images. On the other hand, R'(*or* G', B') means final colors within the overlapping area.

To reconstruct a final surface, after placing a multi-view camera at several positions and acquiring images and 3D point clouds, we apply the above procedure to them. Fig. 6 shows a conceptual diagram.

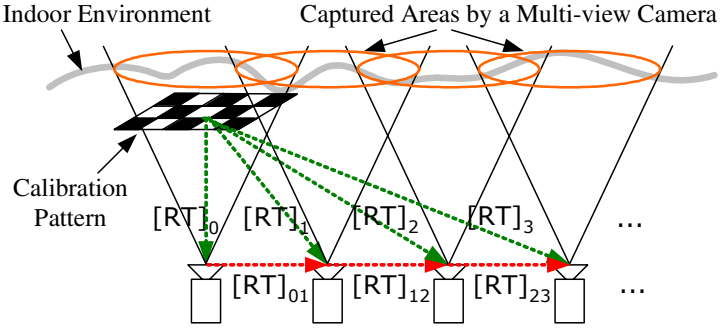


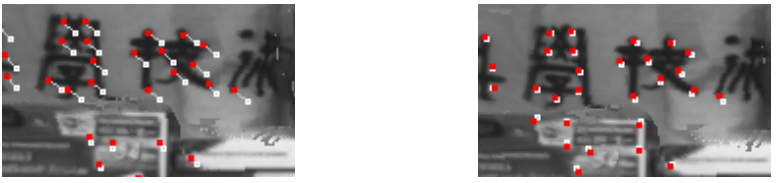
Fig. 6. Indoor scene reconstruction

We first place a pattern whose relative position is already known. Then, we estimate intrinsic/extrinsic parameters with respect to the world coordinate. A camera pose with respect to a specific point of the pattern, $[R\ T]_0$, is evaluated by exploiting Intra-/Inter-calibration and structural information of the multi-view camera [13][17]. After moving the camera to other positions, through the above-mentioned procedure, we can get a relative pose $[R\ T]_i$ at each position, and carry out surface reconstruction.

3 Experimental Results and Analysis

The experiments were carried out under a normal illumination condition of general indoor environment. We used Digiclops which is a multi-view camera for image acquisition [18]. It calculates 3D coordinates through disparity estimation. We employed a planar pattern with 7×5 grid points. Distance between two consecutive points is 10.6 cm. For depth image refinement, 30 frames are used to get mean and standard deviation.

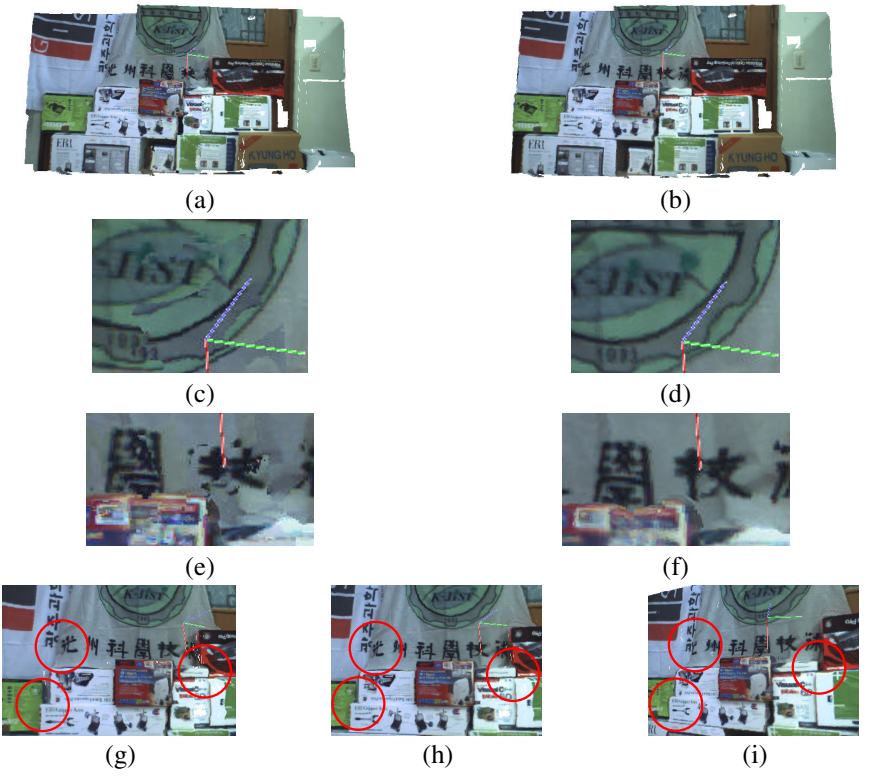
Fig. 7 demonstrates the results of minimizing distance errors between correspondences. Enlarged areas are shown in Fig. 7(a) and Fig. 7(b) at the initial and final steps. Red and white markers represent features of source and destination views, respectively, on the destination view. We can see that the distances between correspondences are effectively minimized.



(a) (b)

Fig. 7. Distance error minimization (a) before (b) after

Fig. 8 illustrates registration results. Fig. 8(a) is a combined 3D point cloud and Fig. 8(b) is the results after registration. We can see that the registration works well by observing the boundary of a circle, Chinese or English characters in Fig. 8(c) to Fig. 8 (f). Furthermore, the navigation, from left to right view within the VE as shown in Fig. 8(g) to Fig. 8(i), proves the depth information of the model and some motion parallax.



(a) (b)

(c) (d)

(e) (f)

(g) (h) (i)

Fig. 8. Registration results (a) combined 3D point cloud (b) registered 3D point cloud (c) enlarged area I in (a) (d) enlarged area I in (b) (e) enlarged area II in (a) (f) enlarged area II in (b) (g) left view (h) front view (i) right view

The registration results for another scene are shown in Fig. 9, which explain that the visual quality of the proposed method is better than that of ICP. Fig. 9(a) and Fig. 9(b) show left and right images, respectively. After initial registration, we can obtain the results as shown in Fig. 9(c). Note that heart shape, face part of bear and some letters are smeared. In Fig. 9(d) and Fig. 9(e), we can see the final registration results of ICP and the proposed method, respectively. Actually, total error is larger than the conventional ICP in terms of the closest distance. However, we observed that the visual quality of the proposed method is much better than that of the conventional ICP. The reason is that the conventional ICP only considers the closest distance instead of data themselves.

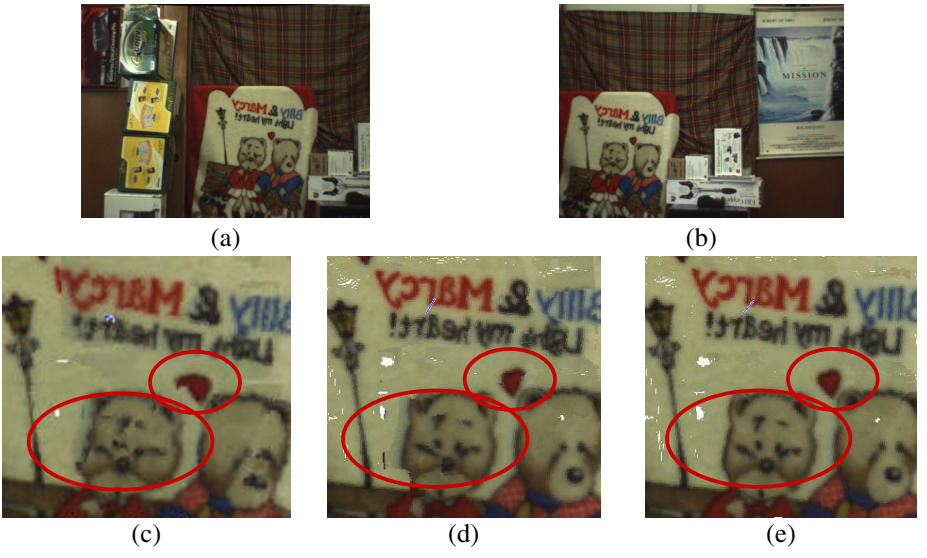


Fig. 9. The comparison of visual quality (a) left image (b) right image (c) initial registration (d) ICP (e) proposed method

The registration and modeling results for two walls are shown in Fig. 10. To get this result, we moved the multi-view camera several times around two walls and registered the acquired 3D point clouds. In Fig. 10(a), two walls are shown. On the other hand, Fig. 10(b) and Fig. 10(c) are scenes for each wall. By applying the proposed method to several sets of 3D point clouds, we can do a dense 3D reconstruction for an indoor environment.

As shown in Fig. 11, the proposed method is superior to ICP or color ICP in the sense of PSNR (Peak Signal to Noise Ratio). This is because the proposed one tracks correspondences by using neighborhood information in an image plane as well as geometric information. The convergence rate of the proposed one is also faster than that of color/texture-based method (α=7.0, N_B=192). The reason is that ICP or color ICP takes longer to search for correspondences than does the proposed method.

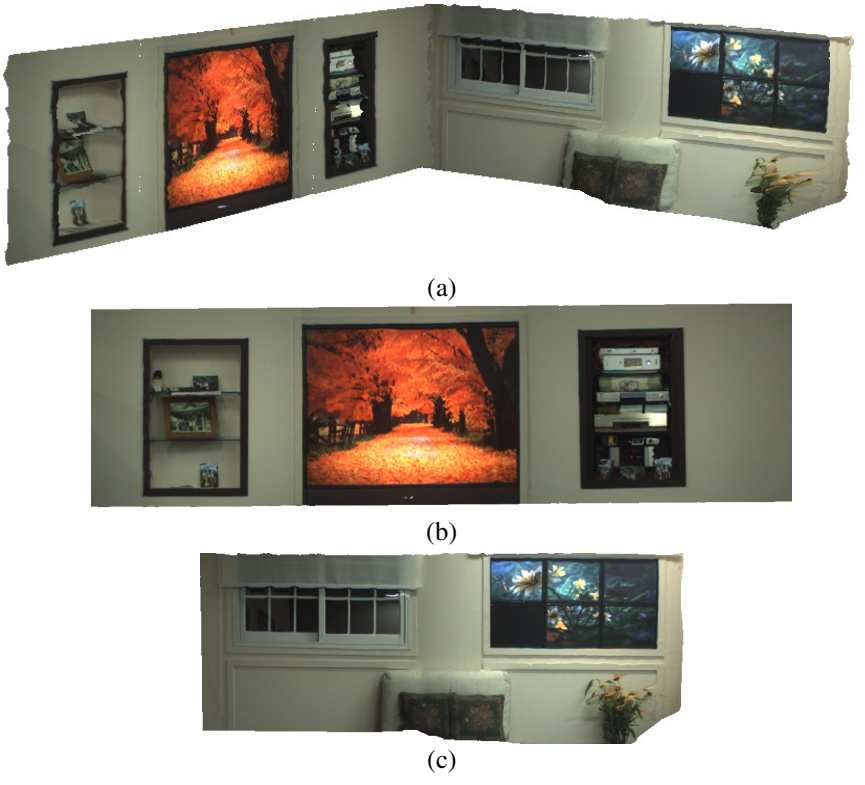


Fig. 10. Indoor Scene reconstruction (a) two walls (b) left wall (c) right wall

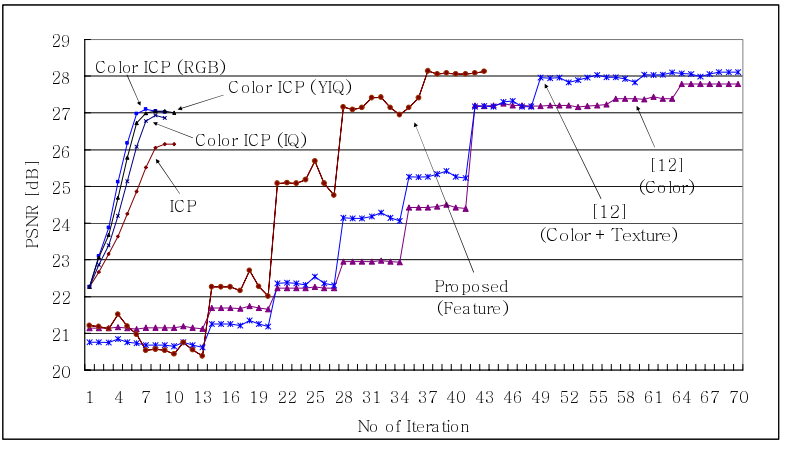


Fig. 11. Performance comparison

4 Conclusions and Future Work

We proposed a novel registration method for 3D point clouds to carry out 3D reconstruction for an indoor environment. We proved that even though the error of depth information is relatively large, effective registration is possible. Furthermore, the required time for registration can be reduced. Only a few views of a real environment are enough for reconstruction instead of numerous 2D images. There are still remaining challenges. Global registration should be optimized to do reconstruction for the entire indoor environment. Natural augmentation of virtual objects into the reconstructed environment requires light source estimation and analysis to match illumination condition of the VE. Finally, dense disparity estimation is required to obtain better results.

References

1. S. Mann, "Mediated Reality," *TR 260*, M.I.T. Media Lab Perceptual Computing Section, Cambridge, Massachusetts, 1994.
2. P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," *IEEE Trans. on PAMI*, vol. 14, no. 2, pp. 239-256, 1992.
3. A. Johnson and S. Kang, "Registration and Integration of Textured 3-D Data," Tech. report CRL96/4, Digital Equipment Corp., Cambridge Research Lab, Oct., 1996.
4. G. Blais and M. D. Levine, "Registering multiview range data to create 3-D computer objects," *IEEE Trans. PAMI*, vol. 17, no. 8, pp. 820-824, 1995.
5. T. Masuda and N. Yokoya. "A robust method for registration and segmentation of multiple range images," *Computer Vision and Image Understanding*, 61(3):295–307, May 1995.
6. S. Weik. "Registration of 3-d partial surface models using luminance and depth information," *Proc. of 3-D Digital Imaging and Modeling*, pp. 93-100, 1997.
7. K. Pulli, *Surface Reconstruction and Display from Range and Color Data*, Ph.D. dissertation, UW, 1997.
8. F. Bernardini, I.M. Martin, and H. Rushmeier, "High-quality texture reconstruction from multiple scans," *IEEE Trans. Visualization and Computer Graphics*, 7(4):318–332, 2001.
9. G. C. Sharp, S. W. Lee and D. K. Wehe, "Invariant Features and the Registration of Rigid Bodies," *IEEE Int'l Conf., on Robotics and Automation*, pp. 932-937, 1999.
10. R. Fisher, "Projective ICP and Stabilizing Architectural Augmented Reality Overlays," *VAA01*, pp 69-80, 2001.
11. K. Nishino and K. Ikeuchi, "Robust Simultaneous Registration of Multiple Range Images Comprising A Large Number of Points," *ACCV2002*, 2002.
12. S. Kim, K. Kim and W. Woo, "Projection-based Registration using Color and Texture Information for Virtual Environment Generation," *LNCS 3331*, pp. 434-443, 2004.
13. K. Kim and W. Woo, "Multi-view Camera Tracking for Modeling of Indoor Environment," *LNCS 3331*, pp.288-297, 2004.
14. KLT: Kanade-Lucas-Tomasi Feature Tracker, http://www.ces.clemson.edu/~stb/klt/, 2005.
15. M. A. Fischler, R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of ACM*, Vol 24, pp 381-395, 1981.

16. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.
17. R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323-344, 1987.
18. Point Grey Research Inc., http://www.ptgrey.com, 2002.