IEICE TRANS. INF. & SYST., VOL.E89-D, NO.1 JANUARY 2006

PAPER Special Section on Artificial Reality and Telexistence

Registration of Partial 3D Point Clouds Acquired from a Multi-view Camera for Indoor Scene Reconstruction

Sehwan KIM[†], *Student Member and* Woontack WOO[†a)], *Nonmember*

SUMMARY In this paper, a novel projection-based method is presented to register partial 3D point clouds, acquired from a multi-view camera, for 3D reconstruction of an indoor scene. In general, conventional registration methods for partial 3D point clouds require a high computational complexity and much time for registration. Moreover, these methods are not robust for 3D point cloud which has a low precision. To overcome these drawbacks, a projection-based registration method is proposed. Firstly, depth images are refined based on both temporal and spatial properties. The former involves excluding 3D points with large variation, and the latter fills up holes referring to four neighboring 3D points, respectively. Secondly, 3D point clouds acquired from two views are projected onto the same image plane, and two-step integer mapping is applied to search for correspondences through the modified KLT. Then, fine registration is carried out by minimizing distance errors based on adaptive search range. Finally, we calculate a final color referring to the colors of corresponding points and reconstruct an indoor scene by applying the above procedure to consecutive scenes. The proposed method not only reduces computational complexity by searching for correspondences on a 2D image plane, but also enables effective registration even for 3D points which have a low precision. Furthermore, only a few color and depth images are needed to reconstruct an indoor scene. The generated model can be adopted for interaction with as well as navigation in a virtual environment.

key words: *projection-based registration, virtual environment generation, multi-view camera, scene reconstruction*

1 Introduction

Image-based 3D reconstruction of a real environment plays a key role in providing visual realism for navigation in and interaction with a virtual environment (VE). Especially, the 3D reconstruction of a real environment is essential for Mediated Reality (MR) in which users can remove/augment or replace the generated virtual objects [1]. Nevertheless, most conventional approaches only emphasize reconstruction itself without considering the interaction with the reconstructed models. Whereas the reconstruction of a real environment with 3D modeling tools is tedious and time-consuming, the generated models are not realistic. Though the reconstruction using active range sensors, combined with a camera, generates an accurate and realistic model, they require expensive sensing equipments and time-intensive reconstruction processing. Furthermore, alignment of 3D point cloud with a texture map is also required. On the other hand, image-based 3D reconstruction methods not only preserve realism but also provide a rather simple modeling procedure. Especially, off-the-shelf multi-view cameras, which provide color as well as depth images, enable us to generate image-based models more easily. Therefore, a delicate registration is indispensable to register seamlessly partial 3D point clouds, acquired from the multi-view camera in a few directions, for 3D reconstruction of a real environment. Note that a magnified 2D image, generated by 2D registration, restricts user navigation and has a limitation in providing interaction. On the other hand, 3D registration can not only overcome these disadvantages but also generate interactive models.

Various registration methods for 3D reconstruction of a real environment have been proposed. Besl proposed ICP (Iterative Closest Point) algorithm [2], which has been widely used so far, and Johnson proposed Color ICP to reconstruct an indoor scene [3], [4]. Blais et al. exploited a simulated annealing to minimize a cost metric based on the total distance between all matches in all views [5]. On the other hand, Nishino presented an optimization method based on M-estimator to implement a robust registration method for several range images [6]. Especially, Pulli developed a data acquisition device, and adopted a projective registration method employing planar perspective warping [7]. Sharp defined invariant features to improve ICP [8], and Fisher applied projective ICP to Augmented Reality (AR) applications [9]. However, most of these methods rely on expensive, accurate equipment, and require much time in generating 3D models [5], [6], [8]. In addition, if 3D point clouds have large error variations, there is a limitation in improving registration accuracy [2]–[4]. Furthermore, stereo cameras are usually used for object modeling, and not for an indoor scene reconstruction [7].

On the other hand, many research activities reconstruct a 3D scene from multiple 2D images captured by a hand-held camera [10]–[12]. However, if enough salient features are not found in the images (e.g. indoor environment), it is difficult to reliably estimate the camera poses. Even after estimating camera poses through optimization, 3D point should also be calculated, e.g., by the multi-baseline stereo technique.

To lessen these problems, a novel projection-based registration method for partial 3D point clouds is proposed. Firstly, a depth image is refined based on the spatio-temporal property of 3D point clouds by using adaptive uncertainty regions. Secondly, 3D point clouds, acquired at two camera views, are projected onto the same destination image plane, and correspondences are searched for within the overlapping

Manuscript received March 11, 2005.
Manuscript revised June 21, 2005.
[†]The authors are with U-VR Lab., GIST, Gwangju, 500–712, Korea.
a) E-mail: wwoo@gist.ac.kr
DOI: 10.1093/ietisy/e89-d.1.62

area through the modified KLT (Kanade-Lucas-Tomasi) feature tracker. Then, two sets of 3D point clouds are fine-registered by minimizing Euclidean distance errors with the help of Levenberg-Marquardt algorithm. Thus, an optimized camera pose is estimated. Finally, each linearly-triangulated point is evaluated referring to corresponding points, and a new color is assigned to each point in the overlapping area. Consequently, we reconstruct an indoor environment by applying the above procedure to consecutive views.

The proposed 3D reconstruction method has the following characteristics. First of all, the proposed registration method, adopting a multi-view camera, paves the way for development of a convenient way to reconstruct a real environment. Until now, there have been extensive researches on indoor scene reconstruction using accurate active range sensors. Thus, the registration results are also accurate since precise 3D point clouds are measured. However, the equipment is very large and much time is needed for reconstruction. Even though the precision of 3D point cloud acquired with pre-calibrated multiple lenses is relatively low, its pervasiveness can simplify the modeling procedure [13], [14]. For this purpose, the proposed method estimates a camera's pose using the projected 2D correspondences instead of 3D coordinates. Thus, registration is effectively carried out even if the available 3D data are less accurate than those of the optical sensor. Besides, the computational complexity is reduced since our approach finds correspondences in a 2D image plane. In addition, the proposed method simplifies the 3D reconstruction of a real environment since it requires placing a multi-view camera only at a few positions in an indoor environment. In summary, the proposed method requires a few 2D images and depth images to reconstruct an indoor environment in a more convenient and fast way. This is made possible by employing a multi-view camera which provides 3D information.

The paper is organized as follows. In Sect. 2, we explain the depth image refinement. Fine registration using color and depth images is discussed in Sect. 3. Then, color selection for the overlapping area and extension to the consecutive views are described in Sect. 4. After experimental results are analyzed in Sect. 5, conclusions and future work are presented in Sect. 6.

2. Depth Image Refinement

Erroneous 3D Points

Unlike optical sensor-based methods which use an active range sensing technique, passive techniques use images generated by the light reflected by objects. However, disparity estimation results in inherent stereo mismatching errors, especially at depth discontinuity areas and on homogeneous areas. These errors cause poor registration results. Thus, unreliable regions should be eliminated before registration. In this regard, a depth image is refined by employing its 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. Figure 1 is a flow diagram for 3D reconstruction of two views in a real environment.

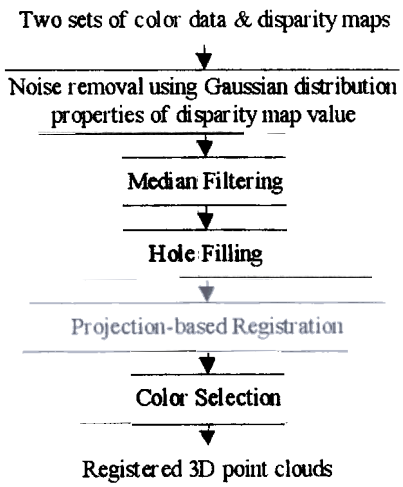


Fig. 1 Flow diagram for 3D reconstructi

To analyze the characteristics of 3D point cloud, mean and standard deviation are calculated for each pixel after acquiring N_f depth images for the same view. However, depth values of some parts have large variations even in a static scene. The reason is that even though the materials of objects are assumed to obey the properties of Lambertian surface, some parts do not satisfy the postulation. In general, depth values drift with time, camera position, illumination conditions of a scene, etc. These factors induce large variations in depth values, especially at depth discontinuity areas as well as on homogeneous areas, when disparity is estimated. Thus, these unstable parts should be removed.

In the depth image of a static scene, depth variation of each pixel is modeled as a Gaussian distribution. After investigating depth value of each pixel, we get rid of the pixels whose depth variation is larger than the threshold value for the ith pixel as follows.

$$\sigma_i \quad \alpha Th_i$$

where σ_i represents a standard deviation for depth variation of the ith pixel. α and Th_i denote a scale factor and a threshold value for the ith pixel, respectively. Then, Median filter is applied to remove spot noises.

However, note that the threshold values depend on 3D coordinates of a scene with respect to the optical center of the multi-view camera. This is because the disparity estimation error increases as an object moves away from the camera. Therefore, Th_i must be expressed as a function of 3D coordinates with respect to the optical center of the camera.

2.2 Adaptive Uncertainty Region

To decide Th_i in terms of the 3D coordinates for a scene,

area through the modified KLT (Kanade-Lucas-Tomasi) feature tracker. Then, two sets of 3D point clouds are fine-registered by minimizing Euclidean distance errors with the help of Levenberg-Marquardt algorithm. Thus, an optimized camera pose is estimated. Finally, each linearly-triangulated point is evaluated referring to corresponding points, and a new color is assigned to each point in the overlapping area. Consequently, we reconstruct an indoor environment by applying the above procedure to consecutive views.

The proposed 3D reconstruction method has the following characteristics. First of all, the proposed registration method, adopting a multi-view camera, paves the way for development of a convenient way to reconstruct a real environment. Until now, there have been extensive researches on indoor scene reconstruction using accurate active range sensors. Thus, the registration results are also accurate since precise 3D point clouds are measured. However, the equipment is very large and much time is needed for reconstruction. Even though the precision of 3D point cloud acquired with pre-calibrated multiple lenses is relatively low, its pervasiveness can simplify the modeling procedure [13], [14]. For this purpose, the proposed method estimates a camera's pose using the projected 2D correspondences instead of 3D coordinates. Thus, registration is effectively carried out even if the available 3D data are less accurate than those of the optical sensor. Besides, the computational complexity is reduced since our approach finds correspondences in a 2D image plane. In addition, the proposed method simplifies the 3D reconstruction of a real environment since it requires placing a multi-view camera only at a few positions in an indoor environment. In summary, the proposed method requires a few 2D images and depth images to reconstruct an indoor environment in a more convenient and fast way. This is made possible by employing a multi-view camera which provides 3D information.

The paper is organized as follows. In Sect. 2, we explain the depth image refinement. Fine registration using color and depth images is discussed in Sect. 3. Then, color selection for the overlapping area and extension to the consecutive views are described in Sect. 4. After experimental results are analyzed in Sect. 5, conclusions and future work are presented in Sect. 6.

2. Depth Image Refinement

2. Erroneous 3D Points

Unlike optical sensor-based methods which use an active range sensing technique, passive techniques use images generated by the light reflected by objects. However, disparity estimation results in inherent stereo mismatching errors, especially at depth discontinuity areas and on homogeneous areas. These errors cause poor registration results. Thus, unreliable regions should be eliminated before registration. In this regard, a depth image is refined by employing its 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. Figure 1 is a flow diagram for 3D reconstruction of two views in a real environment.

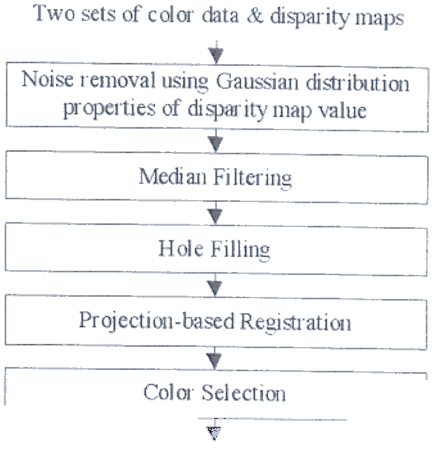


Fig. 1 Flow diagram for 3D reconstruction.

To analyze the characteristics of 3D point cloud, mean and standard deviation are calculated for each pixel after acquiring N_f depth images for the same view. However, depth values of some parts have large variations even in a static scene. The reason is that even though the materials of objects are assumed to obey the properties of Lambertian surface, some parts do not satisfy the postulation. In general, depth values drift with time, camera position, illumination conditions of a scene, etc. These factors induce large variations in depth values, especially at depth discontinuity areas as well as on homogeneous areas, when disparity is estimated. Thus, these unstable parts should be removed.

In the depth image of a static scene, depth variation of each pixel is modeled as a Gaussian distribution. After investigating depth value of each pixel, we get rid of the pixels whose depth variation is larger than the threshold value for the ith pixel as follows.

$$\sigma_i > \alpha Th_i \tag{1}$$

where σ_i represents a standard deviation for depth variation of the ith pixel. α and Th_i denote a scale factor and a threshold value for the ith pixel, respectively. Then, Median filter is applied to remove spot noises.

However, note that the threshold values depend on 3D coordinates of a scene with respect to the optical center of the multi-view camera. This is because the disparity estimation error increases as an object moves away from the camera. Therefore, Th_i must be expressed as a function of 3D coordinates with respect to the optical center of the camera.

2.2 Adaptive Uncertainty Region

To decide Th_i in terms of the 3D coordinates for a scene,

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

where (x', y', z') is a final uncertainty region in terms of 3D coordinates of a scene with respect to the optical center. Therefore, Eq. (1) is reexpressed reflecting 3D coordinates, (x_c, y_c, z_c), with respect to optical center of the camera as well as direction of ray as follows.

$$\sigma_i > \alpha Th_i(x_c, y_c, z_c) \tag{8}$$

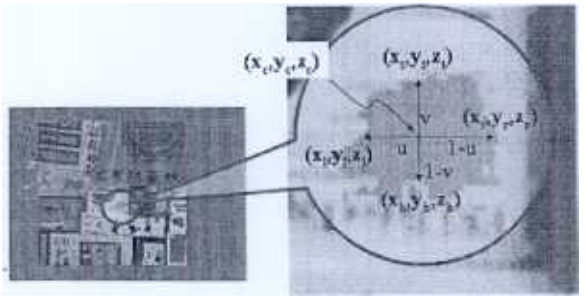
2.3 Hole Filling

Through the above step, some parts of a scene, which do not satisfy the assumption of Lambertian surface, are removed. The depth discontinuity and homogeneous areas are also excluded. However, hole filling is required on holes, generated during the preceding step, and homogeneous areas where disparity may not be estimated. Thus, the spatial property for a current point, i.e. spatial correlation among 3D points of four neighboring pixels, is exploited. The 3D point of the current pixel is estimated by calculating the ratios in horizontal and vertical directions as follows.

$$\begin{aligned} x_c &= ((1-u) \times x_l + u \times x_r + (1-v) \times x_t + v \times x_b)/2 \\ y_c &= ((1-u) \times y_l + u \times y_r + (1-v) \times y_t + v \times y_b)/2 \\ z_c &= ((1-u) \times z_l + u \times z_r + (1-v) \times z_t + v \times z_b)/2 \end{aligned} \tag{9}$$

where (x_c, y_c, z_c) is 3D coordinates of the current pixel (within a hole) in an image as shown in Fig. 4. We can reach four valid points in horizontal and vertical directions starting from the current position. Then, corresponding 3D coordinates are (x_t, y_t, z_t), (x_b, y_b, z_b), (x_l, y_l, z_l) and (x_r, y_r, z_r) for top, bottom, left and right directions, respectively. u and v denote the ratios in horizontal and vertical directions, respectively. That is, 3D coordinates of the current pixel, (x_c, y_c, z_c) is estimated.

However, errors may be generated if depth difference between neighboring points is large. Therefore, this step is not carried out if depth difference is larger than a threshold, Th_{dd}, e.g., at object boundaries. After investigating valid 3D coordinates in each of four directions, we apply the step only to the holes whose size is so small that we can consider each of them as a plane.



Hole filling.

3. Registration of Partial 3D Point Clouds

The depth image refinement not only gets rid of inherent stereo mismatching errors but also 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 since the error bound of 3D point cloud is relatively large. Thus, a projection-based registration method is proposed to carry out a pairing process that searches for correspondences between 3D point clouds of destination and source views. Figure 5 shows the projection-based registration of Fig. 1 in detail.

3 Initial Registration

We let a multi-view camera located around an indoor environment and its pose is estimated by tracking features of a scene. Firstly, a coplanar calibration pattern and structural constraints of the multi-view camera are used to calibrate the camera [15], [16]. Then, intrinsic parameters and camera pose are estimated with respect to a reference position in a real environment. Finally, a rigid-body transformation is applied to 3D points of features at each camera view to estimate the poses of the camera at each view. Therefore, partial 3D point clouds, acquired from a multi-view camera at each position, are registered initially.

Figure 6 shows projected images onto the destination view after the initial registration. Figure 6 (a) and Fig. 6 (b) are projection results of 3D point clouds, which are acquired from the destination and source views, onto the destination view. Mid-luminance value is assigned to unprojected

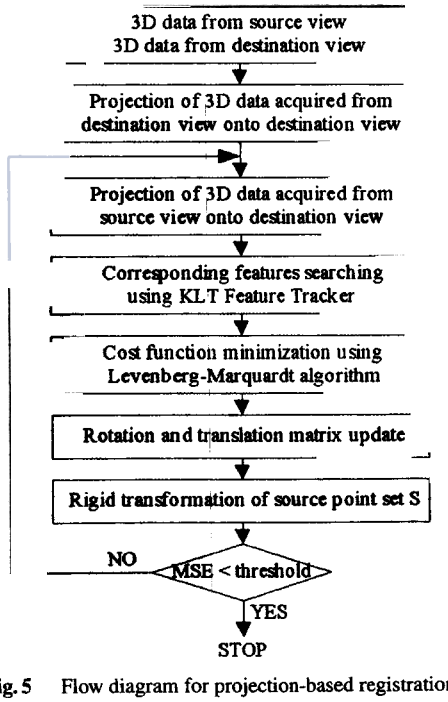


Fig. 5 Flow diagram for projection-based registration.

(a) (b)

Fig. 6 Projection of 3D point cloud onto 2D image plane, (a) projection of 3D point cloud of destination view onto its own view, (b) projection of 3D point cloud of source view onto the destination view.

points to differentiate them from projected ones. It should be noted that the projection of 3D point cloud acquired from the source view induces self-occlusion. This is eliminated based on the rays that originate at the camera center and pass though each pixel. Theoretically, Fig. 6 (b) should exactly overlap with Fig. 6 (a). However, discrepancies occur due to the errors in disparity estimation, camera calibration, etc. Therefore, accurate geometric relationship between two views is found by minimizing errors between correspondences within the overlapping area in terms of projection matrix, P_S, of a source view. Accordingly, fine registration should be accomplished to compensate the errors induced by disparity estimation, camera calibration, and so forth.

3.2 Projected Image Refinement

After initial registration, two sets of 3D point clouds, acquired from the source and destination views, are projected onto the destination view using evaluated calibration parameters. Then, the modified KLT feature tracker is applied to find the correspondences between two image planes. However, the projection of 3D point cloud acquired from the source view onto the destination view produces floating-point numbers. Thus, some pixels do not have any value as shown in Fig. 8 (a). These unprojected pixels can generate false alarms when corresponding features are searched for through the modified KLT feature tracker. Therefore, to preserve an original image as well as to remove unprojected pixels, a special care should be taken. In this case, linear interpolation is useless since object boundaries are smoothed. On the other hand, bi-linear interpolation cannot be used since the exact relationship of two images is unknown.

A two-step integer mapping is presented to meet these requirements. In Fig. 7, grid points are on the lines. White circles represent grid points, and black circles denote projected pixels of 3D point cloud of the source view onto the destination view.

At the first step, a search range is set to $-0.5 \sim +0.5$ along x and y axes, respectively, for each grid point. The color of each grid point is evaluated by considering weights, which are decided by relative distances with all pixels within the search range. However, there exist grid points which do not include any projected point at the first step. At the second step, the search range is expanded to $-1.0 \sim +1.0$

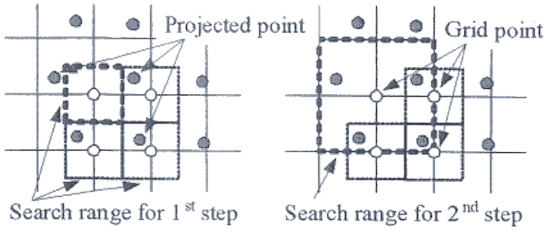
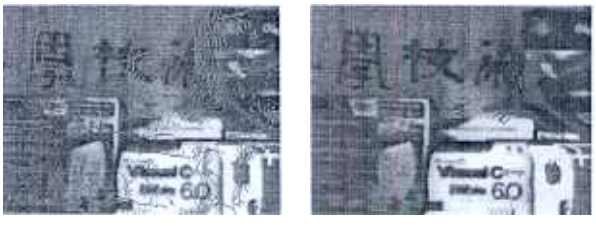


Fig. 7 Two-step integer mapping.



(a) (b)

Fig. 8 Two-step integer mapping, (a) before, (b) after.

and a similar procedure is accomplished. Figure 8 shows the results. Figure 8 (a) is an enlarged part of Fig. 6 (b), and Fig. 8 (b) is the result after the two-step integer mapping is applied. The presented method improves the modified KLT feature tracking performance by removing unmapped grid points and preserving an original image as it is, at the same time.

3.3 Fine Registration Using Correspondences

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 close 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 pro-

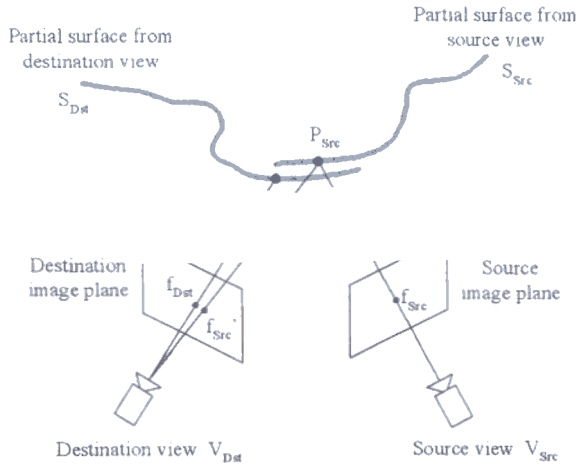


Fig. 9 Selection of corresponding features.

jected onto the destination image plane and corresponding features, f_{Src}', are searched for. This is illustrated in Fig. 9.

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. 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 [17]. By exploiting RANSAC, we can eliminate outliners and obtain only correct pairs between source and destination views.

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

$$L = \sum_{i=0}^{N_{feat}-1} \kappa \left\{ \left(1 - \frac{\left\| f_{Dst,i} - f_{Src,i} \right\|}{Dist_{max}} \right) \left\| f_{Dst,i} - f_{Src,i}' \right\| + \kappa_2 \left\| c_{Ds} \quad c_{Src,i}' \right\|^2 \right\} \tag{10}$$

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.

$$\kappa_1 = \begin{cases} 1 & \text{if } \|P_{Dst} - P_{Src}\| < Th \\ 0 & \text{o/w} \end{cases} \tag{11}$$

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

In summary, we search for correspondences and use some features 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.

Given two sets of corresponding points,

$$\text{Find } \{R_{Src}, T_{Src}\} \text{ w.r.t } \{R_{Dst}, T_{Dst}\} \quad \text{such that } \arg\min_{\{R_{Src}, T_{Src}\}} L \tag{12}$$

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 [18].

4. Surface Reconstruction of Registered 3D Point Clouds

4 Integration

Even after the fine registration phase, there exist points, which do not have the same 3D coordinates even though they are the same point in the real world, due to disparity estimation error. Thus, corresponding points should be manipulated so that they may occupy the same 3D coordinates in the reconstructed space. After the estimation of the camera parameters of source view with respect to destination view, trimming and color selection, for duplicate 3D point clouds within the overlapping area, are carried out. After registration of two views, each grid point of the overlapping area in the destination view has its own correspondence in the source view. This means that we can obtain the corresponding features in 3D point cloud of the original source view. Thus, final 3D coordinates are calculated by a linear triangulation method for the overlapping area [19]. Note that the following clues are exploited; 3D point, for each pixel obtained from a multi-view camera, should lie within an adaptive uncertainty region, and the 3D point within the uncertainty region conforms a Gaussian distribution.

Then, color adjustment is required to consider changes in lighting conditions depending on camera position. We suppose that all materials within a captured scene satisfy the property of Lambertian surface.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \left(v \times \begin{bmatrix} R_{Dst} \\ G_{Dst} \\ B_{Dst} \end{bmatrix} + u \times \begin{bmatrix} R_{Src} \\ G_{Src} \\ B_{Src} \end{bmatrix} \right) (u+v) \tag{13}$$

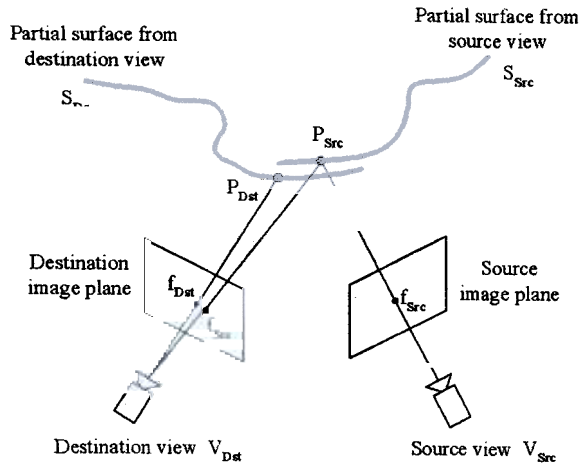


Fig. 9 Selection of corresponding features.

jected onto the destination image plane and corresponding features, f_{Src}', are searched for. This is illustrated in Fig. 9.

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. 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 [17]. By exploiting RANSAC, we can eliminate outliners and obtain only correct pairs between source and destination views.

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

$$L = \sum_{i=0}^{N_{feat}-1} \kappa_1 \left\{ \left(\frac{\left\| f_{Dst,i} - f_{Src,i}' \right\|}{Dist_{max}} \right) \left\| f_{Dst,i} - f_{Src,i}' \right\|^2 + \kappa_2 \left\| c_{Dst,i} - c_{Src,i}' \right\|^2 \right\} \quad (10)$$

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.

$$\kappa_1 = \begin{cases} 1 & \text{if } \|P_{Dst} - P_{Src}\| < Th \\ 0 & \text{o/w} \end{cases} \quad (11)$$

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

In summary, we search for correspondences and use some features 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.

Given two sets of corresponding points,
Find $\{R_{Src}, T_{Src}\}$ w.r.t $\{R_{Dst}, T_{Dst}\}$ (12)
such that $\arg\min_{\{R_{Src}, T_{Src}\}} L$

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 [18].

4. Surface Reconstruction of Registered 3D Point Clouds

4. Integration

Even after the fine registration phase, there exist points, which do not have the same 3D coordinates even though they are the same point in the real world, due to disparity estimation error. Thus, corresponding points should be manipulated so that they may occupy the same 3D coordinates in the reconstructed space. After the estimation of the camera parameters of source view with respect to destination view, trimming and color selection, for duplicate 3D point clouds within the overlapping area, are carried out. After registration of two views, each grid point of the overlapping area in the destination view has its own correspondence in the source view. This means that we can obtain the corresponding features in 3D point cloud of the original source view. Thus, final 3D coordinates are calculated by a linear triangulation method for the overlapping area [19]. Note that the following clues are exploited; 3D point, for each pixel obtained from a multi-view camera, should lie within an adaptive uncertainty region, and the 3D point within the uncertainty region conforms a Gaussian distribution.

Then, color adjustment is required to consider changes in lighting conditions depending on camera position. We suppose that all materials within a captured scene satisfy the property of Lambertian surface.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \left(v \times \begin{bmatrix} R_{Dst} \\ G_{Dst} \\ B_{Dst} \end{bmatrix} + u \times \begin{bmatrix} R_{Src} \\ G_{Src} \\ B_{Src} \end{bmatrix} \right) / (u+v) \quad (13)$$

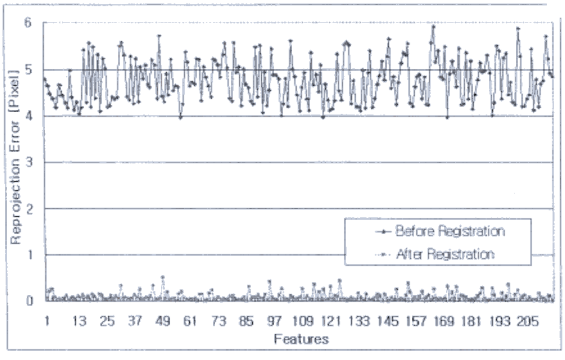


Fig. 13 Reprojection error before and registration.

Table 1 Average rotation and translation errors after registration.

Rotation (degrees)	Pitch	Yaw	Roll
	-0.057912	-0.036512	-0.004023
	x	y	z
	-0.000195	-0.0024234	+0.0008412

where x_i and x_i' are the measured correspondences, and $\hat{x}_i$ and $\hat{x}_i'$ are the estimated correspondences. N_{feat} denotes the number of corresponding features and $d(x, y)$ represents the Euclidean distance between x and y, respectively. In this case, $N_{feat} = 216$. We can also observe that the reprojection error is much reduced after registration.

Table 1 shows the average rotation and translation errors after registration when the standard deviations of the rotation and translation are 0.3° and 0.02 m, respectively. Note that even though errors are large enough to represent the error bound of the multi-view camera, the registration results are feasible.

5.2 Real Data

The proposed method is also applied to the real data as follows. Figure 14 shows the results of depth image refinement. Figure 14 (a) and Fig. 14 (b) show an original image and a corresponding depth image, respectively. Corresponding 3D point cloud and the results of depth image refinement are shown in Fig. 14 (c) and Fig. 14 (d), respectively. We set N_f to 30 and Th_{dd} to 0.15. We can observe that invalid areas, such as object boundaries, homogeneous areas and non-Lambertian surfaces, are effectively removed. Holes, whose depth differences are small, are also filled.

Figure 15 demonstrates the results of minimizing the distance errors between corresponding features. That is, we applied the modified KLT feature tracker to Fig. 6 (a) and Fig. 6 (b), and minimized the distance between them. Figure 15 (a) and Fig. 15 (b) are projected images of 3D point clouds of the destination and source views onto the destination view. Corresponding features are also marked. Enlarged areas are also shown in Fig. 15 (c) and Fig. 15 (d) at the initial and final steps. Red and white markers represent corresponding features of source and destination views, respectively. We can see that the distances between correspondences are effectively minimized.

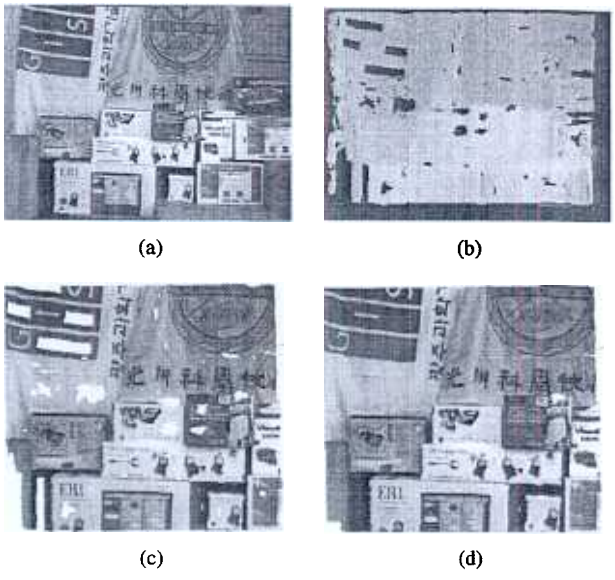


Fig. 14 Depth image refinement, (a) original image, (b) depth image, (c) 3D point cloud before depth image refinement, (d) 3D point cloud after depth image refinement.

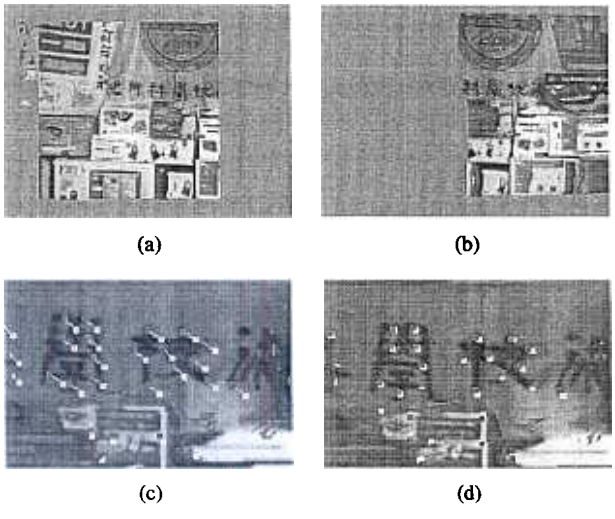


Fig. 15 Corresponding features, (a) projection of 3D point cloud of destination view onto its own view, (b) projection of 3D point cloud of source view to destination view, (c) enlarged area of (b) before error minimization, (d) enlarged area of (b) after error minimization.

Figure 16 illustrates registration results. Figure 16 (a) and Fig. 16 (b) show a combined 3D point cloud acquired from both views and a registered 3D point cloud after applying the proposed method, respectively. On the other hand, Fig. 16 (c) and Fig. 16 (d), Fig. 16 (e) and Fig. 16 (f) are enlarged areas for the corresponding scenes. By observing the boundary shape of a circle, Chinese or English characters, we can see that the registration works well. Furthermore, the navigation, from left to right view within the generated VE as shown from Fig. 16 (g) to Fig. 16 (i), proves the validity of the depth information of the model and some motion

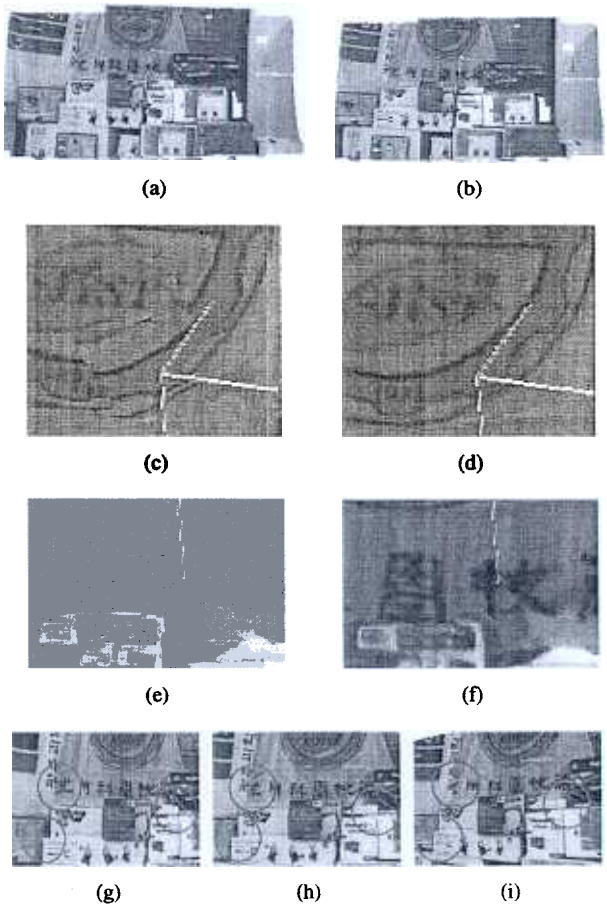


Fig. 16 Registration results, (a) combined 3D point cloud before registration, (b) registered 3D point cloud, (c) enlarged area I of (a), (d) enlarged area I of (b), (e) enlarged area II of (a), (f) enlarged area II of (b), (g) left view, (h) front view, (i) right view.

parallax.

The registration results for another scene are shown in Fig. 17, which explain that the visual quality of the proposed method is better than that of ICP. Figure 17 (a) and Fig. 17 (b) show left and right images, respectively. After initial registration, we can obtain the results as shown in Fig. 17 (c). Note that heart shape, face part of bear and some letters are smeared. In Fig. 17 (d) and Fig. 17 (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.

The registration and modeling results for two walls are shown in Fig. 18. We let the multi-view camera located around a wall, and acquired a color image and 3D point cloud. In Fig. 18 (a), two walls are shown at the same time. On the other hand, Fig. 18 (b) and Fig. 18 (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

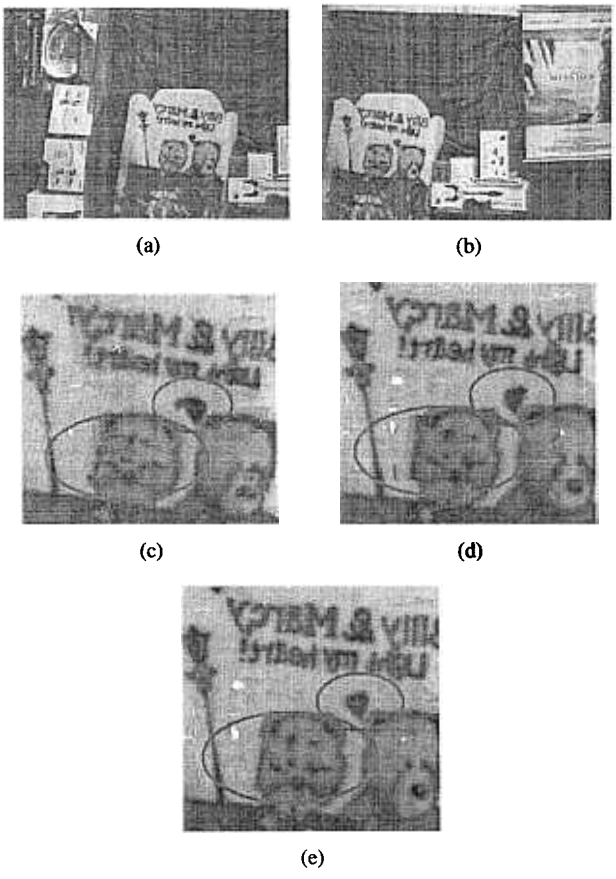


Fig. 17 The comparison of visual quality, (a) left image, (b) right image, (c) initial registration, (d) ICP, (e) proposed method.

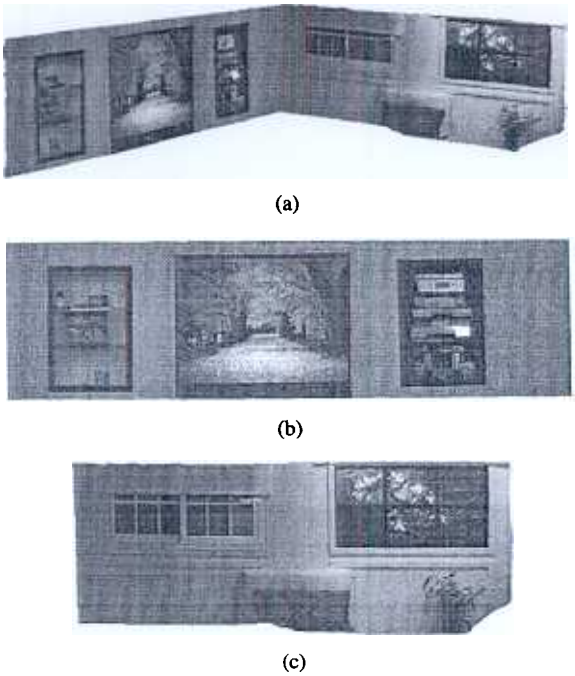


Fig. 18 Indoor scene reconstruction, (a) two walls, (b) left wall, (c) right wall.

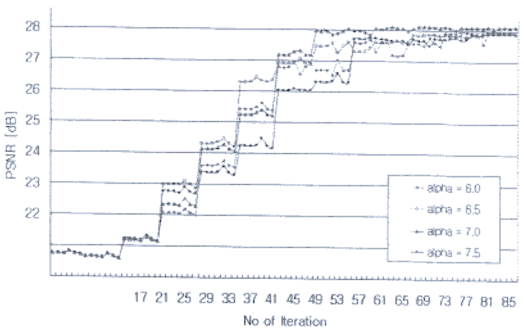


Fig. 19 PSNR according to alpha (α).

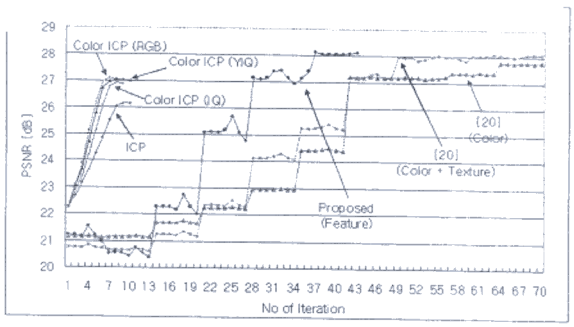


Fig. 20 Comparison of convergence rate.

an indoor environment.

We compared the proposed method with the conventional ICP and color ICP on the basis of PSNR (Peak Signal to Noise Ratio), which represents visual quality of registered results [2], [3]. In general, accurate registration is difficult for 3D point cloud, which has large disparity estimation errors, through ICP and color ICP. This is because those methods are just based on the closest distance and do not consider neighboring pixels. Projection-based ICP just projects 3D point cloud of source view to destination view and searches for paired 3D point. However, it is difficult to guarantee correct corresponding features in case of the large error bound. On the other hand, the proposed method tracks the corresponding features by exploiting surrounding information for each block in a 2D image plane. Then, it minimizes the distance between the corresponding features. Thus, we can see that the proposed method provides results that are more reliable. We took advantage of the following measure for comparing performance in the overlapping area.

$$PSNR = 20\log_{10}\frac{255}{\sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(Y_{Src,i} - Y_{Dst,i})^2}} \text{ (dB)} \quad (15)$$

where N is the number of pixels, which are valid for both images; and $Y_{Src,i}$ and $Y_{Dst,i}$ denote the luminance value of the ith point on the projected source and destination data, respectively.

Figure 19 depicts PSNR according to α when block size is 40×40 and the number of block $N_B = 192$ [20]. As shown in the results, convergence rate as well as PSNR may be different depending on α. When α is 7.0, not only convergence is the fastest but also PSNR is the highest. However, special care should be taken while choosing the optimal value for α. Also, much time is required to extract texture information from using a Gabor filter.

As shown in Fig. 20, the visual quality of the proposed method is superior to those of conventional ICP and color ICP. The visual quality of the method using color and texture simultaneously ($\alpha = 7.0$, $N_B = 192$) is better than that of the method using only color. Furthermore, the convergence of the feature-based method is faster than that of the color and texture-based method because it employs the corresponding features. Besides, visual quality of the proposed method is also similar to the color and texture-based method. Although it seems that the conventional ICP and color ICP converge faster, they take longer to search for correspondences than does the proposed method. The reason is that our method is based on a 2D image plane while the conventional ICP and color ICP search for correspondences in 3D space.

6. Conclusions and Future Work

We proposed a novel registration method that exploits partial 3D point clouds acquired from a multi-view camera to carry out 3D reconstruction of an indoor environment. We proved that even though the error of depth information is relatively large compared to that of laser-scanned data, 3D point clouds are effectively registered between two views. Furthermore, the time required for registration is reduced. We also showed that an effective reconstruction is possible using a few views of the real environment instead of many 2D images. There are still several remaining challenges. First, we have to reduce convergence rate for registration. Global registration should also be optimized for 3D reconstruction of the entire indoor environment. Natural augmentation of virtual objects into the reconstructed room environment requires light source estimation and analysis to match illumination conditions of the VE. Finally, dense disparity estimation is required to obtain better results.

Acknowledgments

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA, and in part by CTRC at GIST.

References

[1] S. Mann, "Mediated reality," TR 260, MIT Media Lab Perceptual Computing Section, Cambridge, MA, 1994.
[2] P.J. Besl and N.D. McKay, "A method for registration of 3-D shapes," IEEE Trans. Pattern Anal. Mach. Intell., 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 Corporation,

Cambridge Research Lab, 1996.
[4] S. Kang and R. Szeliski, "3-D scene data recovery using omnidirectional multibaseline stereo," Tech. Report CRL-95-6, Oct. 1995.
[5] G. Blais and M.D. Levine, "Registering multiview range data to create 3-D computer objects," IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.8, pp.820–824, 1995.
[6] K. Nishino and K. Ikeuchi, "Robust simultaneous registration of multiple range images comprising a large number of points," ACCV2002, pp.454–461, 2002.
[7] K. Pulli, Surface Reconstruction and Display from Range and Color Data, Ph.D. Dissertation, University of Washington, 1997.
[8] G.C. Sharp, S.W. Lee, and D.K. Wehe, "Invariant features and the registration of rigid bodies," IEEE Int'l Conf., Robotics and Automation, pp.932–937, 1999.
[9] R. Fisher, "Projective ICP and stabilizing architectural augmented reality overlays," Int. Symp. on Virtual and Augmented Architecture (VAA01), pp 69–80, 2001.
[10] T. Rodriguez, P. Sturm, M. Wilczkowiak, A. Bartoli, M. Personnaz, N. Guilbert, F. Kahl, M. Johansson, A. Heyden, J.M. Menendez, I. Ronda, and F. Jaureguiza, "VISIRE. Photorealistic 3D reconstruction from video sequences," ICIP03, vol.III, pp.705–708, 2003.
[11] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," Int. J. Comput. Vis., vol.59, no.3, pp. 207–232, 2004.
[12] T. Sato, M. Kanbara, and N. Yokoya, "Outdoor scene reconstruction from multiple image sequences captured by a hand-held video camera," Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent System (MFI2003), pp.113–118, 2003.
[13] Point Grey Research Inc., http://www.ptgrey.com, 2002.
[14] VIDERE DESIGN, http://www.videredesign.com, 2004.
[15] R. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-shelf TV Cameras and Lenses," IEEE J. Robot. Autom., vol.3, no.4, pp.323–344, 1987.
[16] K. Kim and W. Woo, "Multi-view camera tracking for modeling of indoor environment," Lecture Notes in Computer Science 3331, pp.288–297, 2004.
[17] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol.24, pp.381–395, 1981.
[18] S. Kim and W. Woo, "Projection-based registration using multi-view camera for indoor scene reconstruction," 3-D Digital Imaging and Modeling (3DIM), pp.484–491, 2005.
[19] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2004.
[20] S. Kim, K. Kim, and W. Woo, "Projection-based registration using color and texture information for virtual environment generation," Lecture Notes in Computer Science 3331, pp.434–443, 2004.

Sehwan Kim received his B.S. degree in Electronics Engineering from University of Seoul (UOS), Seoul, Korea, in 1998 and M.S. degree in Dept. of Info. & Comm. from Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 2000, respectively. Now, he is a Ph.D. candidate in Dept. of Info. & Comm. at GIST since 2000. Research Interest: Virtual/Mixed Reality, 3D computer vision and its applications including attentive AR and mediated reality, HCI.

Woontack Woo received his B.S. degree in EE from Kyungpook National University, Daegu, Korea, in 1989 and M.S. degree in EE from POSTECH, Pohang, Korea, in 1991. He received his Ph.D. degree in EE-Systems from University of Southern California, Los Angeles, USA. During 1999–2001, as an invited researcher, he worked for ATR, Kyoto, Japan. In 2001, as an Assistant Professor, he joined Gwangju Institute of Science and Technology (GIST), Gwangju, Korea and now at GIST he is leading U-VR Lab. Research Interest: 3D computer vision and its applications including attentive AR and mediated reality, HCI, affective sensing and context-aware for ubiquitos computing, etc.