Asia International Symposium on Mechatronics 2008
27-31 August 2008 Hokkaido University, Japan

# Construction of 3D Environment Model from an Omni-Directional Image Sequence

Ryosuke Kawanishi, Atsushi Yamashita and Toru Kaneko

Department of Mechanical Engineering, Shizuoka University, Shizuoka, Japan
(Tel : +81-53-478-1604; E-mail: {f0730034, tayamas, tmtkane}@ipc.shizuoka.ac.jp)

**Abstract:** Map information is important for path planning and self-localization when mobile robots accomplish autonomous tasks. In unknown environments, however, mobile robots should generate an environment map by themselves. We propose a method for 3D environment modeling by a mobile robot. A 3D environment model can be generaterd from the result of 3D measurement using image data. To realize a 3D measurement of objects more efficiently, the robot uses an image sequence acquired by an omni-directional camera which has a wide field of view. The measurement method is based on structure from motion. A triangular mesh is constructed from measurement data. Experimental results showed the effectiveness of the proposed method.

**Keywords:** Omni-Directional Image Sequence, Structure from Motion, 3D Environment Model

## 1. INTRODUCTION

Map information is important for path planning and self-localization when mobile robots accomplish autonomous tasks. In unknown environment, however, mobile robots should construct an environment map by themselves.

3D measurement using image data makes it possible to generate map information [1]. However, an image acquired by a conventional camera has limited field of view [2]. For the problem, a camera which has wide field of view has been proposed, a fisheye cameras [3], an omni-directional camera [4] [5] and so on. Taking account of installation on a mobile robot, an omni-directional camera is suitable because it can get an around view image at once. Gluckman and Nayer show that an omni-directional camera is effective in measurement and recognition of environment [6].

A stereo vision method using two omni-directional cameras is also proposed [7]. The measurement accuracy by stereo vision depends on the baseline length. The longer the baseline length is, the better accuracy is. However, the measurement accuracy of the above method is limited because a baseline length cannot be longer than the robot size. Then, a motion stereo method using stereo pair images which is taken with a single camera at each observation point is equivalent to binocular stereo vision. In this method, baseline length can be longer without restriction of the robot size [8]. Therefore, the method can measure with high accuracy than a measurement by binocular stereo vision.

Measurement accuracy can be expected by using a sensor fusion method using laser [9], GPS [10] or sonar [11] with an omni-directional camera. However, there are some problems that measurement objects and situations are limited and complexity of calibration process for each device increases.

Structure from motion (SFM) is a kind of motion stereo method. SFM estimates the camera motion (the relative relations of camera positions and orientations), and then measures objects in images taken at each observation point. The method extracts and tracks feature points to get corresponding points in an omni-directional image sequence. By position relations of the corresponding points, the method estimates camera motions and measures environment. A triangular mesh is generated from measurement data. By texture-mapping, a 3D environment model is constructed.

Estimation of precise camera motion is important for improvement in accuracy of measurement by SFM. It is necessary to optimize a triangular mesh to match the physical shape of the environment. Then, we propose a method for precise estimation of camera motion and generation of an optimized triangular mesh for construction of the model which adapts to physical shape.

## 2. OUTLINE

A mobile robot executes 3D measurement and modeling by using an omni-directional camera (Fig.1). The mobile robot acquires an omni-directional image sequence during its locomotion.

The process of our method is shown in Fig.2. The method extracts and tracks feature points to get corresponding points in an omni-directional image sequence. By the linear estimation, it estimates the camera motion by using the positions of corresponding points in two images taken at each observation points. With the camera motion and the image coordinates of the feature points, the 3D coordinates of object points are calculated by tri-

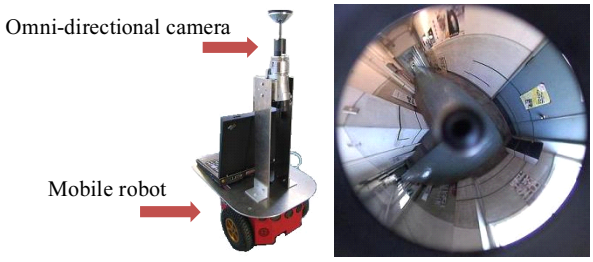



Fig. 1 Mobile Robot and Omni-Directional Image

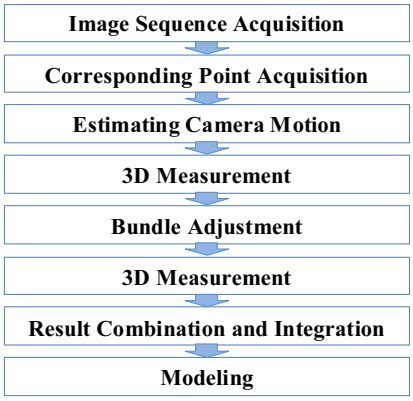



Fig. 2 Process

angulation. By the nonlinear estimation, it can estimate the camera motion more precisely than the linear one. Finally, the individual measurement data is combination and integration.

## 3. ALGORITHM

### 3.1 Corresponding Point Acquisition

For getting correspondent points between images in the omni-directional image sequence, the method extracts feature points in the first image and then tracks them along the sequence. In our method, we use Lucas Kanade tracker algorithm with image pyramid representation [12] (Fig.3).

Tracking feature points is so easy that the points are characteristic visually (high contrast, rich texture, etc.). These points are regarded as corresponding between two images taken at before and after the robot movement. Then, we extract the points which are more characteristic in the images.

### 3.2 Essential Matrix Calculation

We define a unit vector originating from the center of projection to an object point in 3D space as a ray vector $\mathbf{r} = [x, y, z]^T$, where T stands for transposition of vector or matrix. An omni-directional camera we use has a hyperboloid mirror in front of a lens of a conventional camera. Therefore, as shown in Fig. 4, ray vector $\mathbf{r}$ is directed from the focus of the hyperboloid mirror to the refection point of the ray on the mirror surface (Fig.4).

Ray vector $\mathbf{r}$ is calculated from image coordinates $[u, v]^T$ of the feature using Eq.(1) and (2).

$$\mathbf{r} = \begin{bmatrix} su \\ sv \\ sf - 2c \end{bmatrix} \tag{1}$$

$$s = \frac{a^2\left(f\sqrt{a^2+b^2} + b\sqrt{u^2+v^2+f^2}\right)}{a^2f^2 - b^2(u^2+v^2)} \tag{2}$$

In the equations, $a$, $b$ and $c$ are the hyperboloid parameters and $f$ is an image distance which is the distance between a center of a lens and the image plain.

Matrix $\mathbf{E}$ which satisfies Eq.(3) is called an essential matrix,

$$\mathbf{r'}_i^T \mathbf{E} \mathbf{r}_i = 0 \tag{3}$$

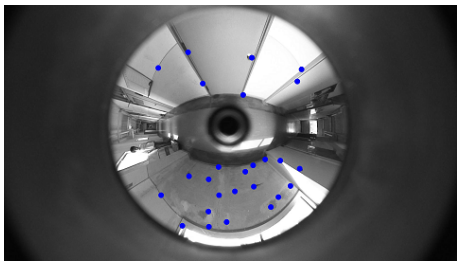


Fig. 3 Feature Extraction and Tracking

where ray vectors $\mathbf{r}_i^T = [x_i, y_i, z_i]^T$, $\mathbf{r'}_i^T = [x'_i, y'_i, z'_i]^T$ are those of the corresponding point in two images, respectively. Essential matrix contains information about relative position and orientation differences between two observation points.

Equation (3) is transformed into Eq.(4),

$$\mathbf{u^T e} = \mathbf{0} \tag{4}$$

where

$\mathbf{u} = [x_i x'_i, y_i x'_i, z_i x'_i, x_i y'_i, y_i y'_i, z_i y'_i, x_i z'_i, y_i z'_i, z_i z'_i]^T$

$\mathbf{e} = [e_{11}, e_{12}, e_{13}, e_{21}, e_{22}, e_{23}, e_{31}, e_{32}, e_{33}]^T$

($e_{jk}$ is the row $j$ and column $k$ element of matrix $\mathbf{E}$). Essential matrix $\mathbf{E}$ is obtained by solving simultaneous equations for more than eight pairs of corresponding ray vectors. This means that we solve Eq.(5).

$$\min_e \| \mathbf{Ue} \|^2 \tag{5}$$

where $\mathbf{U} = [u_1, u_2, \cdots, u_n]^T$. $\mathbf{e}$ is given as the eigenvector of the smallest eigenvalue of $\mathbf{U}^T\mathbf{U}$ and then essential matix $\mathbf{E}$ is obtained.

### 3.3 Outlier Rejection

All feature points tracked along the image sequence do not behave satisfactorily as corresponding points because of image noise and so on. Feature points of mistracking should be rejected as outliers. To solve this problem, we employ a method of RANSAC (RANdom SAmple Consensus) [13].

In the procedure, we select randomly eight feature points, which are the minimum number of points for determining essential matrix $\mathbf{E}$. Let $\mathbf{E}_{rand}$ be the essential

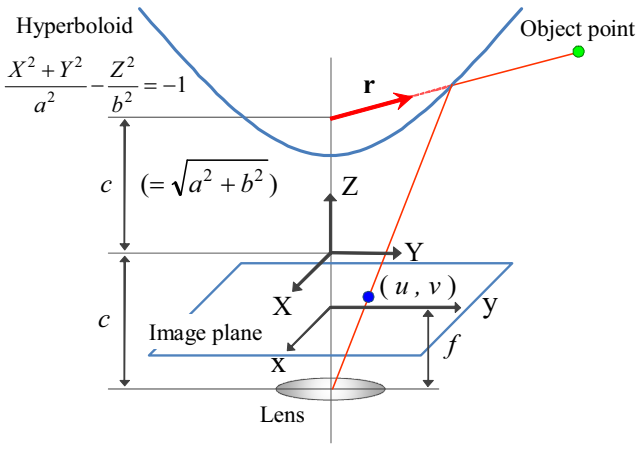



Fig. 4 Calculation of Ray Vector

matrix determined by using these feature points, and $k$ be the number of feature points satisfying Eq.(6), respectively, where $q$ is a given threshold.

$$|\mathbf{r}'^{T}_{i}\mathbf{E}_{rand}\mathbf{r}_i| < q \tag{6}$$

We repeat this process of determining essential matrix $\mathbf{E}_{rand}$ and number $k$ for predetermined times. Then we choose the case with the maximum number of $k$, and remove feature points that do not satisfy Eq.(6) as outliers. Finally, we calculate essential matrix $\mathbf{E}$ by Eq.(5) using the remaining feature points.

### 3.4 Decision of Number of Feature Points

The estimation of camera motion by using the eight point algorithm can calculate rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ if we get at least eight pairs of corresponding points. However, if there are few pairs of the correspondence point, it is difficult to estimate the camera motion precisely because of the influence of various errors in images. It is better to extract a lot of feature points for modeling environment in detail. However, when too many feature points are extracted, the precision of camera movement estimation deteriorates because the number of feature points which is characteristic visually enough in an image is limited. Then, feature points rejected as outlier by RANSAC increase. In other words, when there are a lot of these points, the estimation of camera movement is low precision. Therefore, we propose the method of automatic decision on the most suitable number of the feature points to use in measurement. We define the number of feature points to use in measurement as the number of the feature points including the maximum number of the outliers in the range that precision of camera motion estimation can be high enough. From the above, we give Eq.(7) and (8) as follows.

$$\zeta(k) = \left| \frac{w\sum_{i=k}^{k+w} i \cdot O_i - \sum_{i=k}^{k+w} i \sum_{i=k}^{k+w} O_i}{w\sum_{i=k}^{k+w} i^2 - \left(\sum_{i=k}^{k+w} i\right)^2} \right| \tag{7}$$

$$\zeta(k) < t \tag{8}$$

where $O_i$ is the number of the outliers when the number of extracted feature points is $i$, $t$ is a given threshold, $\zeta(k)$ is an increase of the number of the outliers when the number of the extracted feature points changes from $k$ to $(k+w)$. We can calculate $\zeta(k)$ stably by setting $w$ adequately.

The number of the feature points that are easy to track is different in each environment. Then, if we extract many feature points than the suitable number, the number of the outlier increases drastically. Therefore, we calculate maximum $k$ satisfying Eq.(8) as the most suitable number of the feature points to use in measurement.

### 3.5 Estimating Camera Motion

Essential matrix $\mathbf{E}$ is represented by rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t} = [t_x, t_y, t_z]^T$.

$$\mathbf{E} = \mathbf{RT} \tag{9}$$

Here, $\mathbf{T}$ is a matrix given as follows.

$$\mathbf{T} = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}$$

We calculate $\mathbf{R}$ and $\mathbf{T}$ from essential matrix $\mathbf{E}$ by singular value decomposition.

### 3.6 3D Measurement

3D coordinates of an object point which is projected as a feature point in the image are given based on triangulation with two cameras set in the geometrical relation given by rotation matrix $\mathbf{R}_m$ and translation vector $\mathbf{t}_m$, where $m$ is the number of measurement. We calculates 3D coordinates of $\mathbf{p}_{m,i}$ ($i$-th feature point) by using $\mathbf{R}_m$ and $\mathbf{t}_m$ (Fig.5).

### 3.7 Result Qualification

The accuracy of measurement is poorer as an object point lies closer to the baseline direction or it lies far from the camera. Therefore, the measurement data is a mixture of high and low accuracy. Here, by taking the differentiation of measurement result $\mathbf{p}_{m,i}$ by the image coordinates of the two feature points $[u_{m,i}, v_{m,i}]^T$ and $[u'_{m,i}, v'_{m,i}]^T$ as the estimate of the measurement accuracy, we select measurement results satisfying Eq.(10) and (11), where $h$ is a threshold.

$$\mathbf{g} = \left|\frac{\partial \mathbf{p}_{m,i}}{\partial u_{1,m,i}}\right| + \left|\frac{\partial \mathbf{p}_{m,i}}{\partial v_{1,m,i}}\right| + \left|\frac{\partial \mathbf{p}_{m,i}}{\partial u_{2,m,i}}\right| + \left|\frac{\partial \mathbf{p}_{m,i}}{\partial v_{2,m,i}}\right| \tag{10}$$

$$||\mathbf{g}|| < h \tag{11}$$

### 3.8 Bundle Adjustment

The camera motion estimated in Section 3.5 may not be always good results because the results don't consider the various errors in images. Then, we estimate the camera motion in consideration of the measurement errors in each feature point. We use bundle adjustment which is nonlinear least squares problem by the minimization of the sum of feature reprojection errors [14]. In calculation of the reprojection errors, we use the result of the camera motion estimated in Section 3.5 as initial parameter and evaluate the measurement errors to weight the calculated value in each feature point. The reprojection error is a difference between the original feature point coordinate and the reprojected coordinate. If there are few reprojection errors, the estimation of the camera motion is highly

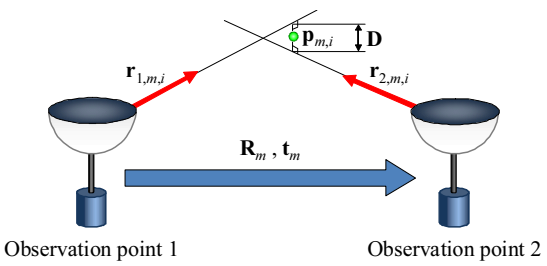



Fig. 5 Calculation of 3D Coordinates

precise. The coordinates of the reprojected feature points are calculated by Eq. (12)-(14).

$$\begin{bmatrix} u \\ v \\ -c+f \end{bmatrix} = d' \begin{bmatrix} dx \\ dy \\ dz+2c \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ -c \end{bmatrix} \quad (12)$$

$$d = \frac{cz' + b\sqrt{x'^2+y'^2+z'^2}}{\left(\frac{b}{a}\right)^2 (x'^2+y'^2) - z'^2} \quad (13)$$

$$d' = \frac{f}{dz'+2c} \quad (14)$$

We define the sum of the feature reprojection errors as follows.

$$E_f = \sum_p r_{fp} \left| \mathbf{x_{fp}} - \mathbf{x'_{fp}} \right|^2 \quad (15)$$

where $E_f$ is the sum of the feature reprojection errors in $f$-th frame, $r_{fp}$ is the weight factor for $p$-th feature point, $\mathbf{x}_{fp}$ is the original feature point coordinate, $\mathbf{x}'_{fp}$ is the reprojected coordinate, respectively. The weight factor is calculated with evaluating the error in each feature point. The norm of the vector $\mathbf{g}$ calculated by Eq.(10) evaluates the error (Eq.(16)).

$$r_{fp} = \frac{1}{\left\| \mathbf{g_{fp}} \right\|} \quad (16)$$

where $\mathbf{g}_{fp}$ is a vector which expresses the measurement accuracy of $p$-th feature point in $f$-th frame. We add larger weight to the feature points which has a small vector, namely a small error.

### 3.9 Result Combination and Integration

By the above procedure, we get individual measurement results and the geometrical relations of observation points by using pairs of stereo images that are selected in the image sequence. In order to unite these measurement results, we solve the problem of scale ambiguity among individual measurements by scale matching [15]. However, there is a mismatching of the measurement results because the errors included in feature points are different between each observation point. Therefore, after scale matching, there are more than one measurement results which show the same object points. We should integrate these measurement results into one object point. The integration method votes to voxels which are divided 3D space. We decide the weight of vote value in consideration of the measurement error (Eq. (16)) in each feature point. The 3D coordinate of the object point is the coordinate of the voxel which has largest vote value.

### 3.10 Modeling

A triangular mesh is generated from integrated measurement data by using 3D Delaunay triangulation. However, Delaunay triangulation generates a triangular mesh which contradicts a physical shape because the triangular mesh doesn't consider the shape of the measurement object. Therefore, we apply the triangular optimization method [16] to the triangular mesh (Fig.6). The method

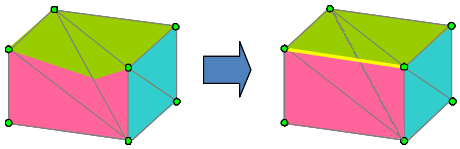


Fig. 6 Triangular Mesh Optimization

adapts the triangular mesh to physical shape by detecting a texture distortion. By texture mapping to the triangular mesh, a 3D environment model is constructed. By texture mapping to the triangular mesh, a 3D environment model is constructed.

## 4. EXPERIMENT

In the experiment we measured two environments (Fig.7(a) a passageway including an L-shape corner, (b) a room). We acquired an image sequence (5 fps) of a passageway by using an omni-directional camera installed to a mobile robot. The robot ran with 10 cm/s. We acquired an image sequence (10 fps) of a room by using the camera which a walking person has. The sizes of these image sequences are 1920 $\times$ 1080 pixels.

Figure 8 shows the top view of combined measurement result of a passageway. The blue marks in this result show the trajectory of the robot. The red marks in this result show measurement data. The result with our proposed method (a) (Section 3.3, 3.4, 3.7, 3.8) is higher precision than the result without the method (b), because of the accuracy of the passageway shape.

Figure 9 shows the bird's-eye view of 3D environment model of a passageway. Figure 10 shows the detail view of 3D environment model. Because there are little texture distortions in the model, the result shows that our proposed method can construct a model which doesn't contradict physical shape of measurement object.

Figure 11 shows the top view of combined measurement result of a room. Our proposed method rejects measurement results which are low precision.

Figure 12 shows various views of 3D environment model of a room. We can recognize object shape in an arbitrary viewpoint. The result shows that our proposed method can measure shape of objects in detail.

## 5. CONCLUSION

In this paper, we propose a method which estimates camera motion more precisely. We apply the triangular optimization method which adapts the triangular mesh to physical shape by detecting a texture distortion. Experimental results showed the effectiveness of the proposed method.

As future works, we should make the following improvements. For measurement accuracy, the camera motion should be estimated by consideration of more than three observation points. For appearance of an environment model, we must use the texture mapping method that considered the distortion of an omni-directional image.

## REFERENCES


[1] A. J. Davison: “ Real-Time Simultaneous Localisation and Mapping with a Single Camera ”, Proceedings of the 9th IEEE International Conference on Computer Vision, Vol. 2, pp. 1403-1410, 2003.

[2] H. Ishiguro, M. Yamamoto and S. Tsuji: “ Omni-Directional Stereo ”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 2, pp. 257-262, 1992.

[3] E. Schwalbe: “ Geometric Modelling and Calibration of Fisheye Lens Camera Systems ”, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 36(5/W8), 2005.

[4] R. Bunschoten and B. Krose: “ Robust Scene Reconstruction from an Omnidirectional Vision System ”, IEEE Transactions on Robotics and Automation, Vol. 19, No. 2, pp. 351-357, 2003.

[5] C. Geyer and K. Daniilidis: “ Omnidirectional Video ”, The Visual Computer, Vol. 19, No. 6, pp. 405-416, 2003.

[6] J. Gluckman and S. K. Nayar: “ Ego-motion and Omnidirectional Cameras ”, Proceedings of the 6th International Conference on Computer Vision, pp. 999-1005, 1998.

[7] J. Takiguchi, M. Yoshida, A. Takeya, J. Eino and T. Hashizume: “ High Precision Range Estimation from an Omnidirectional Stereo System ”, Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.263-268, 2002.

[8] M. Tomono: “ 3-D Localization and Mapping Using a Single Camera Based on Structure-from-Motion with Automatic Baseline Selection ”, Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 3353-3358, 2005.

[9] J. Meguro, T. Hashizume, J. Takiguchi and R. Kurosaki: “ Development of an Autonomous Mobile Surveillance System Using a Network-based RTK-GPS ”, Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp.3107-3112, 2005.

[10] J. Meguro, Y. Amano, T. Hashizume and J. Takiguchi: “ Omni-Directional Motion Stereo Vision Based on Accurate GPS/INS Navigation System ”, Proceedings of the 2nd Workshop on Integration of Vision and Inertial Sensors, 2005.

[11] S. Wei, Y. Yagi and M. Yachida: “ Building Local Floor Map by Use of Ultrasonic and Omni-Directional Vision Sensor ”, Proceedings of the 1998 IEEE International Conference on Robotics and Automation, pp. 2548-2553, 1998.

[12] J. Y. Bouguet: “ Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm ”, OpenCV, Intel Corporation, 2000.

[13] M. A. Fischler and R. C. Bolles: “ Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography ”, Communications of the ACM, Vol. 24, No. 6, pp. 381-395, 1981.

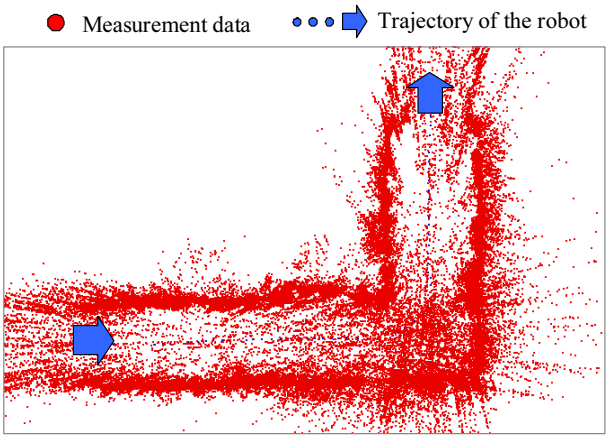
[14] B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon: “ Bundle Adjustment -A Modern Synthesis ”, Vision Algorithms: Theory & Practice, Springer-Verlag LNCS 1883, 2000.

[15] T. Harada, A. Yamashita and T. Kaneko: “ Environment Observation by Structure from Motion with an Omni-directional Camera ”, Proceedings of International Workshop on Advanced Image Technology 2006, pp.169-174, 2006.

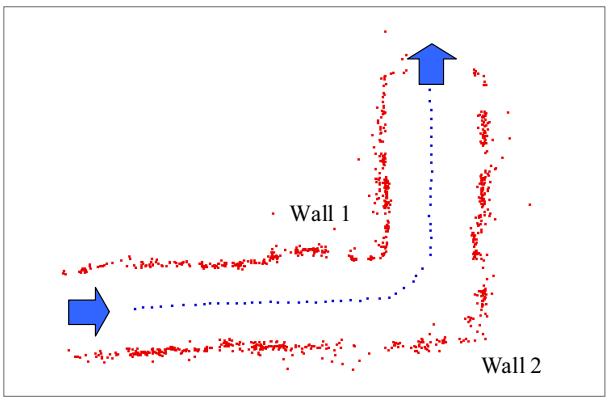
[16] D. D. Morris and T. Kanade: “ Image-Consistent Surface Triangulation ”, Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol.1, pp.332-338, 2000.


(a) Passageway (b) Room

Fig. 7 Experiment Environment

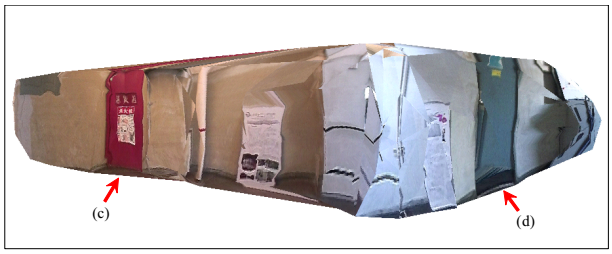



(a) Result without Our Method

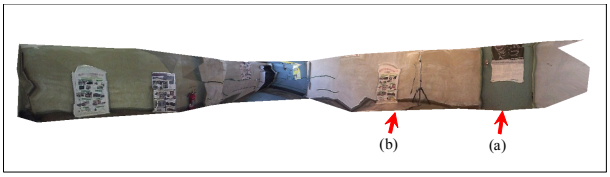



(b) Result with Our Method

Fig. 8 Integrated Measurement Data of Passageway

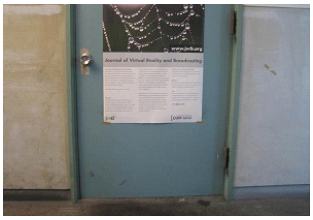


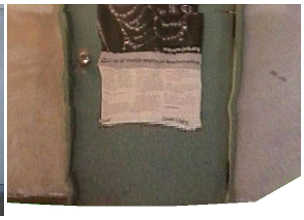

(a) Wall 1



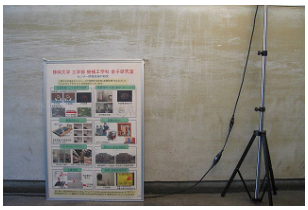

(b) Wall 2

Fig. 9 Bird's-eye View of 3D Environment Model





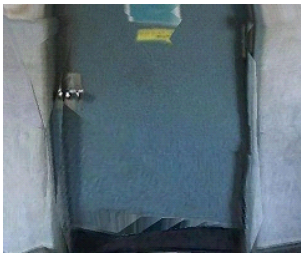
(a) Door 1



(b) Penel

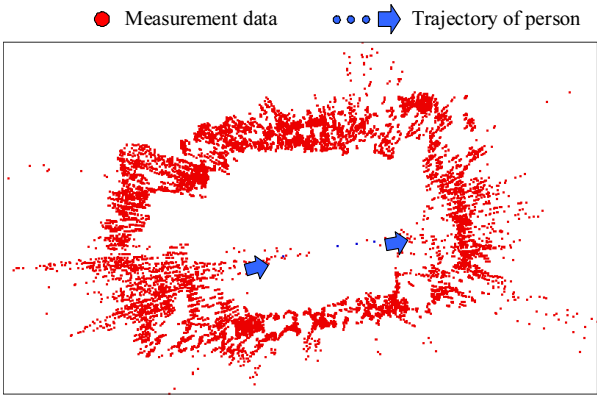


(c) Fire Hydrant

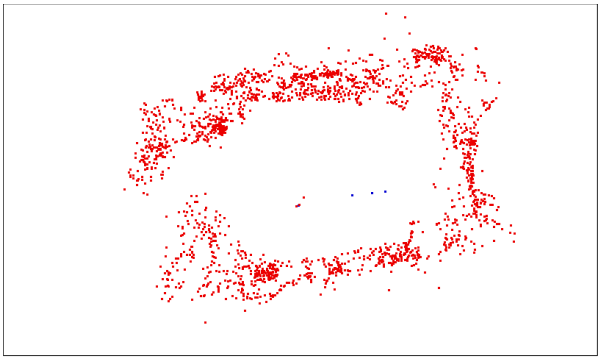


(d) Door 2

Fig. 10 Detail View



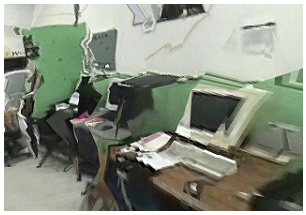

(a) Result without Our Method



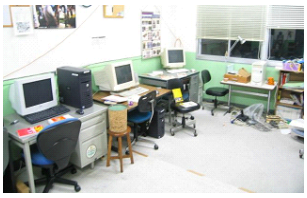
(b) Result with Our Method

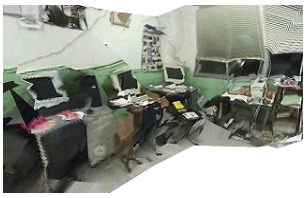
Fig. 11 Inregrated Measurement Data of Room

(a)



(b)





(c)

Fig. 12 3D Model of Room