Tracking of Human Groups Using Subtraction Stereo

Yuma Hoshikawa, Yuki Hashimoto, Alessandro Moro, Kenji Terabayashi, and Kazunori Umeda

Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University / CREST, JST
1-13-27, Kasuga, Bunkyo, Tokyo, Japan
{hoshika, hashimo, moro}@sensor.mech.chuo-u.ac.jp
{terabayashi, umeda}@mech.chuo-u.ac.jp

***Abstract*— In this paper, we propose a method for tracking of groups of people using three-dimensional(3D) feature points obtained by KLT method [1, 2] and a stereo camera system called "Subtraction stereo" [3]. The tracking system using the subtraction stereo, which focuses its stereo matching algorithm to foreground regions obtained by background subtraction, is realized using Kalman filter modelled tracker. The effectiveness of the proposed method is verified in the scenes whose people walk in 3D environment and are difficult to be tracked.**

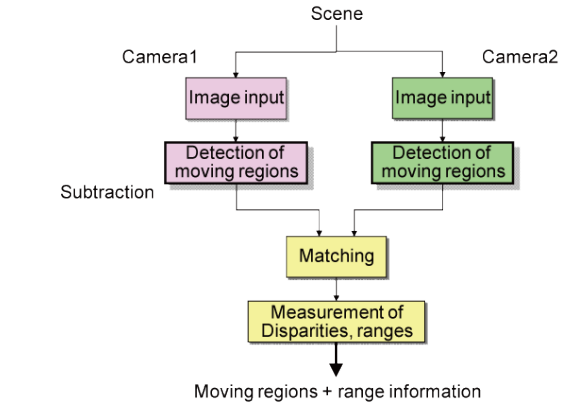
***Keywords*—stereo vision, tracking, human group, KLT, Kalman filter**

I. INTRODUCTION

A huge number of studies about sensing people for surveillance use have been carried out until now. In the scenes obtained from a surveillance camera, people sometimes create groups such as people walking in an amusement park, downtown streets, and so on. In such scenes, information that groups generate might be useful for understanding scenes. For example, if people in a group spread suddenly, something unusual might have been occurred to the group itself or the place near to the group. In order to use such kind of "group information," detection and tracking of groups are necessary.

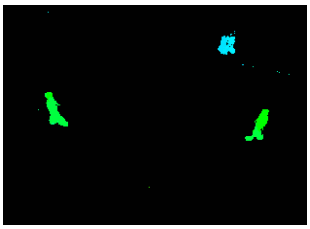
Many methods about detection and tracking of people in various situations have been proposed. Rodriguez et al. [4] realized tracking of persons in highly crowded scenes such as soccer stadium, the main entrance of a university and so on. First, they use Correlated Topic Model and create scene model learning the direction and number of the people moving in the scene, and then use this scene model to support tracking system. This method can be used even if people in a high crowded scene move several directions, not only in one direction. Sugimura et al. [5] also proposed a tracking method which realizes the tracking in highly crowded scene using an up-and-down motion of feature points on people obtained by KLT [1,2].

These days, many studies about detection and tracking of people in 3D environment using stereo camera have also been carried out [6-8]. Hoshikawa et al. [9] proposed a tracking system using Kalman filter [10] for the tracking and stereo vision method called "subtraction stereo" [3] for the 3D measurement of people. This subtraction stereo method makes the distance calculation robust by applying its stereo matching only to the extracted regions obtained by background subtraction. As this method focuses its distance calculation to foreground, it can detect persons from a disparity image in wide range with one stereo camera. Although these methods realized the tracking of individuals in a scene, they do not actually detect the "group of people" and track this group as one object using 3D information.



(a) The basic algorithm of the subtraction stereo

(b) Input image



(c) Output disparity image

Fig.1 Subtraction stereo

In this paper, we propose a tracking method which tracks groups of people as one object using the 3D feature points obtained by subtraction stereo and KLT, and the basic idea of data association proposed by Gennari [11]. The proposed method detects the group of people using the relationship between the measured 3D feature points and tracks group of people applying the Kalman filter modelled tracker.

This paper is organized as follows. In section 2, we show the outline of the subtraction stereo. In section 3, we discuss the tracking method. In section 4, we present experimental results. And then, we conclude this paper in section 5.

II. SUBTRACTION STEREO

The basic algorithm of the subtraction stereo is shown in Fig. 1(a). The subtraction stereo extracts objects in a scene by background subtraction method first, and then applies the stereo matching to the extracted regions. From the disparity image obtained by the subtraction stereo, actual heights and widths of the objects can be obtained. This size of the object can be used to identify whether extracted objects are persons

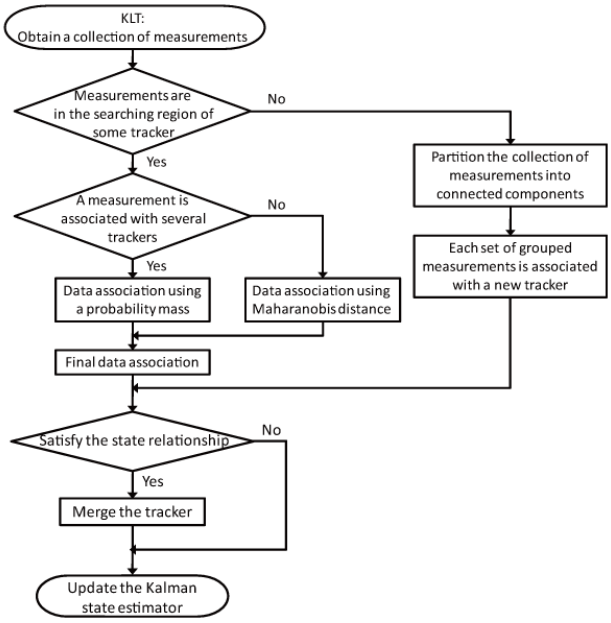


Fig.2 Flow of the group tracking method

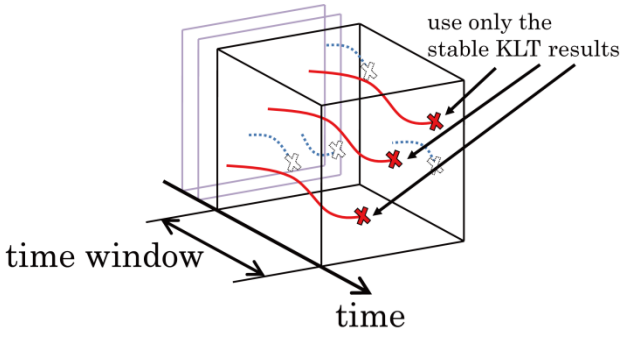


Fig.3 Example of the KLT result

or not. An example of output disparity image is shown in Fig. 1(c), obtained from an input image shown in Fig. 1(b)

III. Group Tracking Method

A. Outline of the Group Tracking Method

In this section, we explain the method to track groups of persons using a tracker modelled with Kalman filter. Although Gennari [11] uses the 2D position information of the people for the data association of the trackers and measured points, we use 3D feature points obtained using KLT and subtraction stereo for the data association. A set of feature points associated with the same tracker is detected as one group.

The flow of the tracking method is shown in Fig. 2. Firstly, KLT method is applied to the regions extracted by the subtraction stereo in order to obtain the 3D feature points of the persons. The feature points which do not satisfy the time window are removed as shown in Fig. 3, and only remained points are used as measured points.

Secondly, these feature points are initially divided into groups using the relationship of the 3D position between each point. The mean position, velocity and covariance of these initial groups are calculated, and then trackers start tracking with this information as the initial state.

Finally after the initialization of the tracking, groups of persons are tracked by associating feature points with each tracker. For the data association between measured feature points and trackers, the Maharanobis distance, calculated from the covariance of the tracker and 3D position of the each feature point, is used. In the situation which groups come closer to each other, for example crossing of groups, a feature point is associated with several trackers. In such situation, the feature point is associated considering the probability of the position and velocity calculated from the state information.

B. State Model of the Tracker

In this paper, we apply a constant-velocity model for the state transition model of Kalman filter modelled tracker because the velocity of ambulation through frames can be considered as constant. The state $\mathbf{X}$ of the Kalman filter is defined as:

$$\mathbf{X} = \begin{bmatrix} x & \dot{x} & y & \dot{y} & z & \dot{z} \end{bmatrix}^T \tag{1}$$

where (x, y, z) and $(\dot{x}, \dot{y}, \dot{z})$ are the world coordinate and velocity of person in the world coordinate system. The Kalman filter predicts the state at time k+1 from the state at k as:

$$\mathbf{X}_{k+1} = \Phi \mathbf{X}_k + \Gamma_k \boldsymbol{\omega}_k \tag{2}$$

where ω_i is the process noise and Φ is the state transition model matrix. Φ is given by

$$\Phi = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

The measurement $\mathbf{Z}$ for the Kalman filter is defined as:

$$\mathbf{Z} = \begin{bmatrix} u & v & d \end{bmatrix}^T \tag{4}$$

where u and v are the image coordinate of person in an image, and d is the disparity. The relation between state $\mathbf{X}$ and measurement $\mathbf{Z}$ is represented as follows:

$$\mathbf{Z}_k = f(\mathbf{X}_k) + \mathbf{v}_k \tag{5}$$

$$f(\mathbf{X}_k) = \begin{bmatrix} \frac{x_k \cdot f}{z_k} & \frac{y_k \cdot f}{z_k} & \frac{b \cdot f}{z_k} \end{bmatrix}^T \tag{6}$$

where f and b are the focal length and baseline length of the camera respectively, and v_k is the measurement noise.

The covariance of a tracker representing a position and a velocity of the group is updated using variance of the measured feature points. The covariance of the tracker $\Sigma(k)$ is updated as:

$$\Sigma(k+1) = \alpha_s \cdot \Sigma(k) + (1-\alpha_s) \cdot C_y \qquad (7)$$

where C_y is the covariance which shows the variance of the position and velocity of the measured feature points, and α_s is a weight. With all these variables, the tracker is updated.

The tracker counts the number of the associated 3D feature points N every frame. This number N is used as a weight in case that several groups merge into a single group.

C. Data Association of Feature Points and Trackers

In this paper, a group of the persons are tracked by associating the 3D feature points with the tracker. The measured feature points are associated with the tracker when the tracker satisfies a search region of the measured feature point. The search region SR_ξ is defined as follows:

$$SR_\xi = \left\{ y_i \mid (y_i - \xi_j)^T \cdot \Sigma^{-1} \cdot (y_i - \xi_j) < \gamma \right\} \qquad (8)$$

where y_i is the 3D position of a feature point, ξ_j is the predicted 3D position of the tracker, and Σ is the covariance representing the variance of the position of the tracker. The γ is a threshold obtained experimentally every time camera arrangement is changed because the number of feature points depends on the size of the object in an image. In order to determine the γ when camera arrangement is changed, several groups of people are tracked as for the test and determine the appropriate number for γ.

When several groups come closer, some feature points might be associated with several trackers. In order to associate the feature points with the appropriate tracker, the probability of the size and velocity of the groups are considered. When the data association between the i-th feature points and j-th tracker is represented as $\theta_{i,j}$, the association probability is calculated by following equations.

$$m_p(\theta_{i,j}) = k_p \cdot p(y_i \mid \xi_j, P_{\xi_j}) \qquad (9)$$

$$m_v(\theta_{i,j}) = k_v \cdot p(\lambda_i \mid v_j, P_{v_j}) \qquad (10)$$

$$m_{Total}(\theta_{i,j}) = m_p(\theta_{i,j}) \cdot m_v(\theta_{i,j}) \qquad (11)$$

where y_i and λ_i are the position and the velocity of the i-th measured feature points respectively. ξ_j and v_j are the mean position and velocity of the tracker respectively, P_{ξ_j} and P_{v_j} represent the covariant matrix of the position and velocity of the tracker respectively. k_p and k_v represent the weight. The tracker whose probability calculated from eq. (11) is the max value is associated with the feature point.

With all these data association, the mean position of the measured feature points associated with same tracker is calculated. This mean position of the feature points is used for update of the state of the tracker. If the number of the feature points associated with a same tracker is less than threshold, the tracker updates its state without measured data.

D. Initial Grouping of the Feature Points

Feature points, not associated with any trackers, are grouped initially in such a scene when a group appears into an image. The initial grouping is done using the relationship between the positions of each measured feature points. The relationship between the positions of each feature points is represented as follows:

$$\mathbf{y}_i \; R_0 \; \mathbf{y}_j \quad \Leftrightarrow \quad SR_{\mathbf{y}_i} \cap SR_{\mathbf{y}_j} \neq \emptyset \qquad (12)$$

where SR_{y_i} represents the search region of each feature point. This search region SR_{y_i} is set as a sphere that has a radius obtained experimentally. If the number of the feature points which have overlapped search region each other is over threshold, these feature points are sorted as a group. The mean position, velocity, and the covariance of this initially grouped feature points are given to the tracker as an initial state of the group, and the tracker start tracking from this moment.

E. Group Merging

Groups sometimes merge together and create new group. In this paper, the trackers identify whether groups are merged or not by using the state similarity. For the state similarity, the Maharanobis distance between the states of each tracker is used. Groups are merged if the search regions using the Maharanobis distance overlap as follows:

$$\mathbf{x}_i \; R \; \mathbf{x}_j \quad \Leftrightarrow \quad MR_{\mathbf{x}_i} \cap MR_{\mathbf{x}_j} \neq \emptyset \qquad (13)$$

where $\boldsymbol{x_i}$ and MR_{x_i}represent a state of the i-th tracker and search region with the Maharanobis distance respectively. The number of associated feature points N is used as a weight for deciding the state of new merged groups. This N becomes large when the group is consisted of many people and less when the group is consisted of less people. The weight is used to decide the state of the merged group generated. Let I_n and G_i be a set of the groups satisfying the condition (13) and elements of this set I_n respectively. For this set I_n, each element is merged by applying the following equations:

$$\mathbf{x}_{m,j} = \sum_{G_i \in I_n} \overline{N_{i,j}} \cdot \mathbf{x}_i \qquad (14)$$

$$\boldsymbol{\Sigma}_{m,j} = \sum_{G_i \in I_n} \left[\overline{N_{i,j}} \cdot \left\{ \boldsymbol{\Sigma}_i + (\mathbf{x}_i - \mathbf{x}_{m,j}) \cdot (\mathbf{x}_i - \mathbf{x}_{m,j})^{\mathrm{T}} \right\} \right] \qquad (15)$$

where Σ_i represent covariance of the i-th tracker, and $\overline{N_{i,j}}$ is defined as $\overline{N_{i,j}} = N_i / \sum_{G_i \in I_n} N_i$. This N_i is the number of the measured points associated with the i-th tracker.

IV. Experiment

The effectiveness of the proposed tracking method is evaluated in three scenes. In the first scene, two groups walk across the image and cross each other. In this scene, data associations of the each feature point become unstable because of the partial occlusion. In the second scene, two groups merge into a single group. In the third scene, a group

in which walks up a stairs is tracked. This experiment shows the effectiveness of the proposed tracking method which can be used in such a 3D environment.

Each experiment was done with following settings. A stereo camera used for the experiments is Point Grey Research Bumblebee2 (color, f=3.8[mm]). The camera was set at the height of 5.1[m] and 40[^{o}] downward tilt in first and second experiment. For the third experiment, the camera was set at the 2.1[m] and 40[^{o}] downward tilt. The process noise, measurement noise and the threshold γ defined in eq. (8) were set to the appropriate values estimated experimentally. For the initial grouping condition defined in eq. (12), the search region SR_{y_i} of each feature point was set as a sphere which has a radius of 1.0[m] for first and second experiment, and 0.5[m] for third experiment. The measured feature points are used for state update of the tracker when there were more than two feature points associated with the tracker.

A. *Groups Crossing Each Other*

The effectiveness of the proposed method is evaluated in a scene in which two groups move along opposite directions and cross. In the scene, groups are partially occluded and the associations of each feature point become unstable.

The experimental result is shown in Fig. 4(b), and the 3D trajectories of each group are shown in Fig. 5. In Fig. 4(b), each dot on the groups represents the measured feature points, and the ellipse shows the groups. In Fig. 5, trajectories are plotted in the world coordinate whose origin point is just under the camera. Although one of the groups is occluded behind the other group, each group is tracked correctly and trajectories are obtained well. Using the velocity information, these groups are not merged into a single group.

B. *Groups Merging into a Single Group*

We verified that two groups merge into a single group with our proposing method. The experimental scene is shown in Fig. 6(a). Two groups walk from each side of the image and merge into a single group at the centre of the image.

The experimental result is shown in Fig. 6(b), and the 3D trajectories of each group are shown in Fig. 7. As shown in Fig. 6(b), two groups (Group0 and Group1) are merged into a single group (Group2) and verified the effectiveness of the proposing method. Compared with the experimental result obtained from the crossing experiment, this result also shows that groups are merged only when groups have similar state. Although two groups have similar position information during crossing, these groups do not merge into a single group.

C. *Tracking of a Group in 3D Environment*

In order to verify the effectiveness of the proposing method in 3D environment, a group is tracked in a scene including a stairs. As shown in Fig. 8(a), the group walks on a plane floor first, and the walks up the stairs.

The experimental result is shown in Fig. 8(b), and the 3D trajectory of the group is shown in Fig. 9. As shown in the experimental result, the group can be tracked correctly in 3D environment. Since the tracking method using single lens camera is difficult to be used in such kind of complex 3D environment, we can say our proposed method is useful for tracking in 3D environment.

V. Conclusions and Future Work

In this paper, we have proposed a new tracking method using Kalman filter modelled tracker with 3D feature points obtained by subtraction stereo and KLT. The effectiveness of the new tracking system was verified by the following experiments. In first experiment whose groups of people move along opposite directions and cross each other, we verified that the proposed method could be used even if there are partial occlusions in a scene using velocity and position information for data association. In second experiment, whose groups of people merge into a single group, we have verified the effectiveness of the proposed method for the merge of two groups. This experiment also showed that the velocity of groups is useful for merge of groups and the data association of the tracking. In third experiment whose people walk up a stairs in a scene, we have verified that this tracking method can be used in complex 3D scene. Although effectiveness of the method was verified in these experiments, the method is still difficult to be applied for crowded scenes because the features used for data association are only position and velocity of groups. Therefore, as for the future work, we are going to add other features for the data association and apply the method for crowded scenes such as real city environment.

References

[1] C. Tomasi and T. Kanade: “Detection and tracking of point features”, *Technical Report CMU-CS-91-132*, 1991.

[2] C. Tomasi and J. Shi: “Good features to track,” in *Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[3] K. Umeda, T. Nakanishi, Y. Hashimoto, K. Irie and K. Terabayashi, “Subtraction Stereo -A Stereo Camera System That Focuses On Moving Regions -,” in *Proc. of SPIE-IS&T Electronic Imaging*, 2009, Vol.7239 Three-Dimensional Imaging Metrology, 723908.

[4] M. Rodriguez, S. Ali and T. Kanade, “Tracking in Unstructured Crowded Scenes,” in *Proc. of the IEEE ICCV2009*, 2009, pp.1389-1396.

[5] D. Sugiura, K. M. Kitani, T. Okabe, Y. Sato and A. Sugimoto, “Using Individuality to Track Individuals: Clustering Individual Trajectories in Crowds using Local Appearances and Frequency Trait,” *Proc. of the IEEE ICCV2009*, 2009, pp.1467-1474.

[6] R. M. Salinas, E. Aguirre, and M. García-Silvente, “People detection and tracking using stereo vision and color,” *Image and Vision Computing*, Vol.25, Issue 6, pp.995-1007, 2007.

[7] S. Bahadori, L. Iocchi, G. R. Leone, D. Nardi and L. Scozzafava, “Real-time people localization and tracking through fixed stereo vision,” *Applied Intelligence*, Vol.26, No.2, pp.83-97, 2007.

[8] T. Zhao, M. Aggarwal and T. Germano, “Toward a sentient environment: real-time wide area multiple human tracking with identities,” *Machine Vision and Application*, Vol.19, No.5, pp.301-314, 2008.

[9] Y. Hoshikawa, K. Terabayashi and K. Umeda, “Human Tracking Using Color Information and Subtraction Stereo,” in *Proc. AWSVCI2009*, 2009, pp.5-8.

[10] S. Thrun, W. Burgard and D. Fox, *Probabilistic Robotics*, MIT press, 2006.

[11] G. Gennari and G. D. Hager, “Probablistic Data Association Methods in Visual Tracking of Groups,” in *Proc. of the IEEE CVPR2004*, 2004, Vol. 2, pp.876-881.

(a) Experimental scene

Frame #140 Frame #200 Frame #260 Frame #340

(b) Experimental result (two groups cross each other)

Fig.4 Experimental result whose groups cross each other with partial occlusion

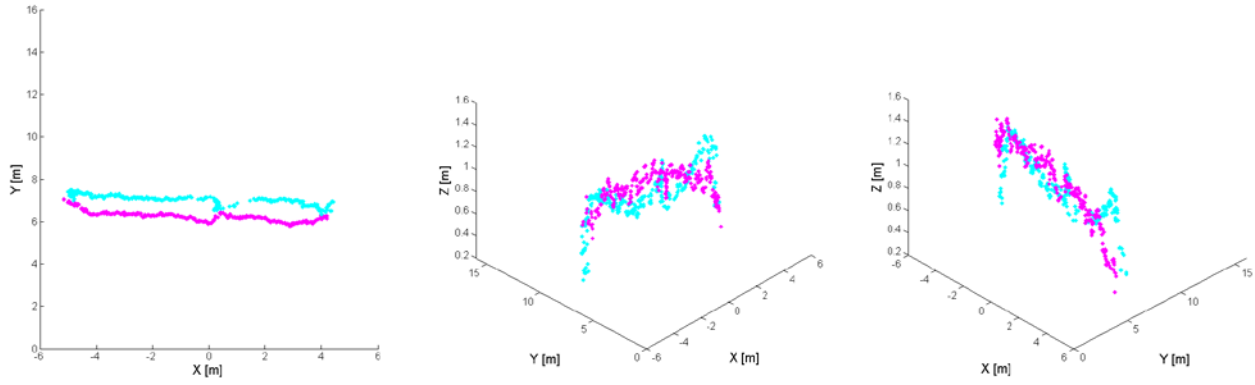


Fig.5 Measured 3D trajectories obtained from the crossing experiment

(Cyan : Group No.3, Magenta : Group No.4)

(a) Experimental scene

Frame #100 Frame #180 Frame #250 Frame #330

(b) Experimental result (two groups merge)

Fig.6 Experimental result whose groups merge into a single group

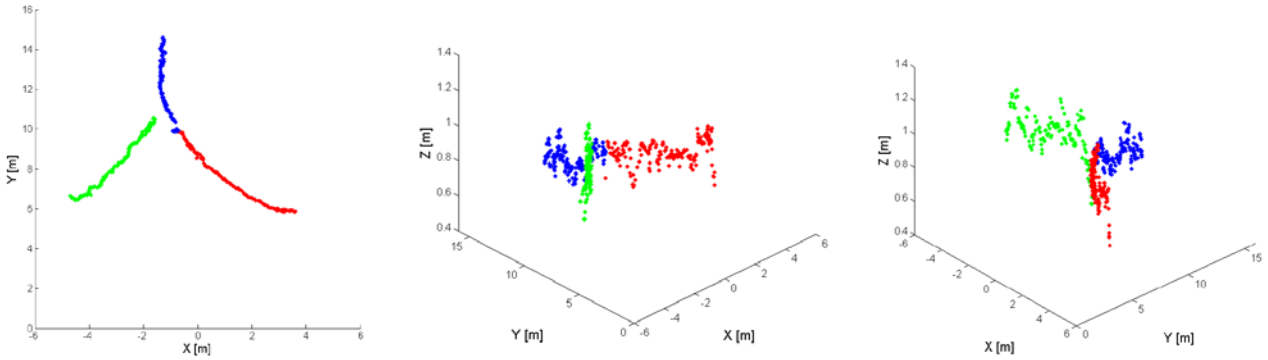
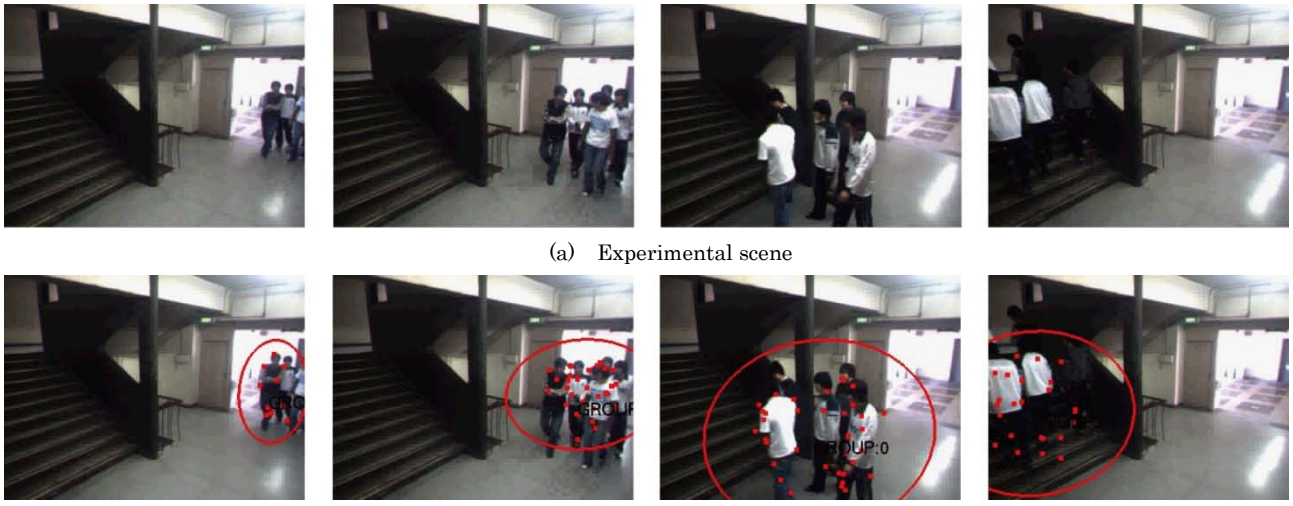


Fig.7 Measured 3D trajectories obtained from the merging experiment

(Red : Group No.0, Green : Group No.1, Blue : Merged Group No.2)



(a) Experimental scene

Frame #80 Frame #170 Frame #270 Frame #430

(b) Experimental result (3D environment)

Fig.8 Experimental result whose persons are measured in 3D environment

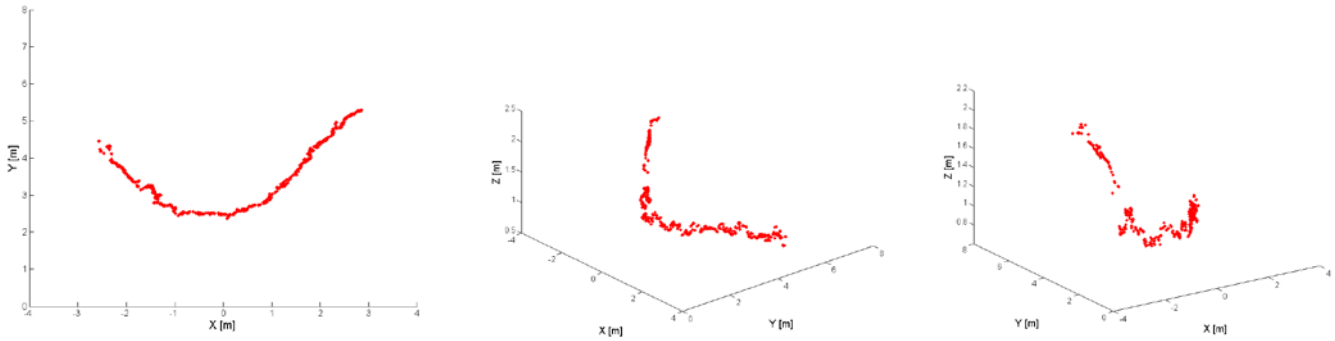


Fig.9 Measured 3D trajectories of the group going up the stairs in the scene