SICE Journal of Control, Measurement, and System Integration, Vol. 4, No. 3, pp. 214–220, May 2011

Tracking of Human Groups Using Subtraction Stereo

Yuma HOSHIKAWA *, Yuki HASHIMOTO *, Alessandro MORO *,
Kenji TERABAYASHI *, and Kazunori UMEDA *

Abstract : In this paper, we propose a method for tracking groups of people using three-dimensional (3D) feature points obtained with use of the Kanade-Lucas-Tomasi feature tracker (KLT) method and a stereo camera system called "Subtraction stereo". The tracking system using subtraction stereo, which focuses its stereo matching algorithm to foreground regions obtained by background subtraction, is realized using Kalman filter based tracker. The effectiveness of the proposed method is verified using 3D scenes of people walking, which are difficult to track.

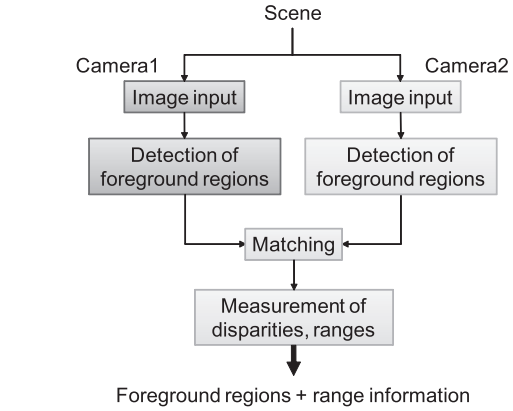
Key Words : stereo vision, tracking, human group, KLT, Kalman filter.

1. Introduction

Many studies have been undertaken on the subject of sensing people for surveillance use. When using surveillance cameras, people are frequently observed in groups, for example, walking in amusement parks and city streets. In such circumstances, the information generated by groups has the potential to aid in the interpretation of a specific scene. For example, if a group suddenly scatters, it could be assumed that something occurred within or near the group to cause that action. This type of group information requires detection and tracking system.

Many methods for the detection and tracking of people in various situations have been proposed. Rodriguez et al. [1] developed a system for tracking people in high dense crowds, e.g. in soccer stadiums and main entrances to universities. They first use correlated topic model and create a model of a scene using direction and number of their movement; the model was then used to support the tracking. This method is applicable in crowded conditions in which individuals are moving in all directions. Sugimura et al. [2] also proposed a tracking system applicable in crowded conditions; their system tracks an up-and-down motion of feature points of individuals obtained by the Kanade-Lucas-Tomasi feature tracker (KLT) [3],[4]. These feature points are clustered into groups using the similarity of frequency of up-and-down motion. Trajectories of people are obtained using gait features, local appearance, spatial proximity and motion coherency. These methods realized high tracking accuracy in crowded scene, but trajectories obtained by these methods are all Kalman filter based tracker, not 3D.

There is also a number of studies where people are tracked after being segmented to individuals using video sequences [5],[6]. Sidla et al. [7] realized detection and tracking of individuals in crowded scene. They use Ω-like shape of upper body for detection of individuals, and track people by predicting their upper body motion using KLT and Kalman filter. Yang et al. [8] took an approach of face detection for segmentation of people to individuals, and use this detection result for tracking. Using the detection result and Bayesian filtering framework with multiple cues around human faces such as color information and elliptical head model, they realized pedestrian tracking in crowded scene. Although these methods realized tracking of individuals in crowded scene, they still have a difficulty for detecting individuals in occlusion scenes.



(a) Basic algorithm of the subtraction stereo

(b) Input image (c) Output disparity image

Fig. 1 Result obtained by subtraction stereo.

To date, many studies have been conducted using stereo cameras to detect and track people in 3D environments [9]–[11]. Hoshikawa et al. [12] proposed a tracking system using Kalman filter [13] for tracking; they used a stereo vision system called "subtraction stereo" [14] for the 3D measurement of people's movement. This subtraction stereo method makes the distance calculation robust by applying its stereo matching only to the extracted regions obtained by background subtraction. Distance calculation is focused on the foreground, therefore individuals can be detected from disparity images over a wide

* Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University / CREST, JST, 1-13-27 Kasuga, Bunkyo, Tokyo, Japan
E-mail: hoshika@sensor.mech.chuo-u.ac.jp
(Received August 12, 2010)
(Revised January 11, 2011)

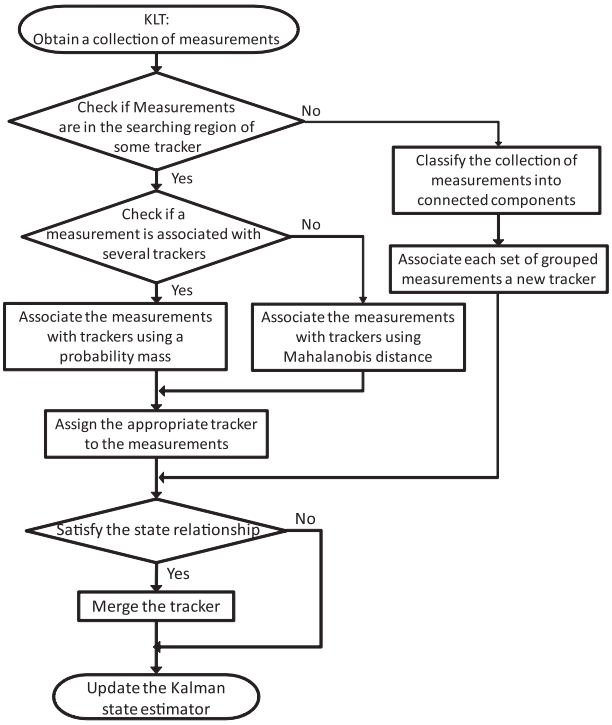


Fig. 2 Flow of the group tracking method.

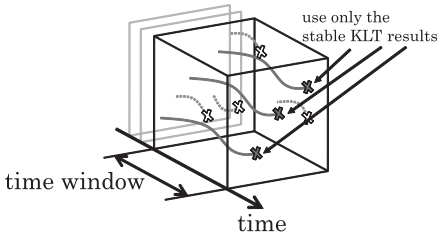


Fig. 3 Example of the KLT result.

range with one stereo camera. Although these methods realize the tracking of individuals in a scene, they do not actually detect a "group of people" and track this group as one object using 3D information.

In this study, we propose a method that tracks groups as one object using the basic idea of data association proposed by Gennari [15] and the 3D feature points obtained by subtraction stereo and KLT. In the Gennari's study [15], a probabilistic data association method which allows the tracking of groups is proposed. The results shown in the Gennari's work are obtained using 2D position on images, and the experiments were undertaken only in one scene, which is not three dimensional. Our proposed method permits the detection of a group of people using the relationship among the measured 3D feature points and the tracking of the group applying the Kalman filter based tracker. The proposed method allows groups of people to be detected and tracked in a three dimensional environment.

This paper is organized as follows. Section 2 is an outline of subtraction stereo. Section 3 is a discussion of the tracking method. Section 4 is a presentation of the experimental results, and Section 5 is the conclusion.

2. Subtraction Stereo

The basic algorithm of the subtraction stereo is shown in Fig. 1 (a). The subtraction stereo first extracts objects in a scene by the background subtraction method 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 determine whether or not extracted objects are persons. An example of an output disparity image is shown in Fig. 1 (c), obtained from an input image shown in Fig. 1 (b). The color in Fig. 1 (c) represents the distance from camera; green means near and blue means far.

3. Group Tracking Method

3.1 Outline of the Group Tracking Method

In this section, we explain the method used to track groups of persons using a tracker modeled with Kalman filter. Although Gennari [15] uses 2D position information of people for the data association of the trackers and measured points, we use 3D feature points obtained with 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. First, the KLT method is applied to the regions extracted by the subtraction stereo to obtain the 3D feature points of the individuals. The feature points that do not satisfy the time window are removed as shown in Fig. 3, and only remaining points are used as measuring points.

Secondly, these feature points are 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 the trackers start tracking with this information as the initial state.

Finally, after the initialization of the tracking, groups of individuals are tracked by associating feature points with each tracker. For the data association between measured feature points and trackers, the Mahalanobis distance, calculated from the covariance of the tracker and 3D position of each feature point, is used. In a situation in which groups come closer to each other, for example, at a crossing, 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.

3.2 State Model of the Tracker

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

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

where (x, y, z) and $(\dot{x}, \dot{y}, \dot{z})$ are the world coordinate and velocity of a person in the world coordinate system.

The Kalman filter predicts the state at time k+1 from the state at k as follows:

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

where $\boldsymbol{\omega}_k$ is the process noise and Φ is the state transition model matrix. Φ is given by

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

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

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

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

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

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

where f and b are the focal length and baseline length of the camera, respectively, and $\mathbf{v}_k$ is the measurement noise.

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

$$\mathbf{\Sigma}(k+1) = \alpha_s \cdot \mathbf{\Sigma}(k) + (1 - \alpha_s) \cdot \mathbf{C}_y \tag{7}$$

where $\mathbf{C}_y$ is the covariance showing the variance of the position and the 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 in every frame. This number N is used as a weight in case several groups merge into a single group.

3.3 Data Association of Feature Points and Trackers

In this paper, a group of 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:

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

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 the camera arrangement is changed because the number of feature points depends on the size of the object in an image. To determine the γ when the camera arrangement is changed, several groups of people are tracked for the test, and they determine the appropriate number for γ.

When several groups get close to each other, some feature points might be associated with several trackers. To associate the feature points with the appropriate tracker, the probability of the size and velocity of the groups is considered. When the data association between the i-th feature points and the j-th tracker is represented as $\theta_{i,j}$, the association probability is calculated with the following equations:

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

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

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

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, and 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 this data association, the mean position of the measured feature points associated with the same tracker is calculated. The mean position of the feature points is used to update the state of the tracker. If the number of the feature points associated with the same tracker is less than the threshold, the tracker updates its state without measured data.

3.4 Initial Grouping of the Feature Points

Feature points, not associated with any trackers, are grouped initially in such a scene when a group appears in 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:

$$y_i \ R_0 \ y_j \Leftrightarrow SR_{y_i} \cap SR_{y_j} \neq \emptyset \tag{12}$$

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 feature points that overlap in the search region is over threshold, the feature points are sorted as a group. The mean position, velocity, and covariance of these initially grouped feature points are given to the tracker as an initial state of the group, and the tracker starts tracking from that moment.

3.5 Group Merging

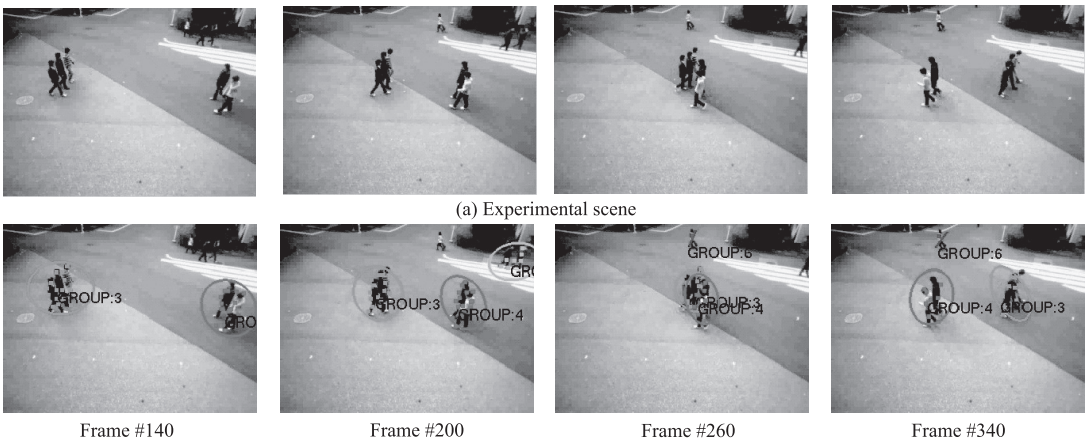
Groups sometimes merge and create a new group. In this paper, the trackers determine whether or not groups have merged by using a state similarity. For state similarity, the Mahalanobis distance between the states of each tracker is used. Groups are merged if the search regions using the Mahalanobis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

$$x_i \ R \ x_j \Leftrightarrow MR_{x_i} \cap MR_{x_j} \neq \emptyset \tag{13}$$

where x_i and MR_{x_i} represent a state of the i-th tracker and the search region with the Mahalanobis distance, respectively. The number of associated feature points N is used as a weight for determining the state of the new merged groups. This N becomes large when the group consists of many individuals and smaller when the group consists of fewer individuals. 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 the elements of this set I_n respectively. For this set I_n, each element is merged by applying the following equations:

$$X_{m,j} = \sum_{G_i \in I_n} \overline{N_{i,j}} \cdot x_i \tag{14}$$

$$\Sigma_{m,j} = \sum_{G_i \in I_n} [\overline{N_{i,j}} \cdot \{\Sigma_i + (X_i - X_{m,j}) \cdot (X_i - X_{m,j})^T\}] \tag{15}$$



(a) Experimental scene

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

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

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

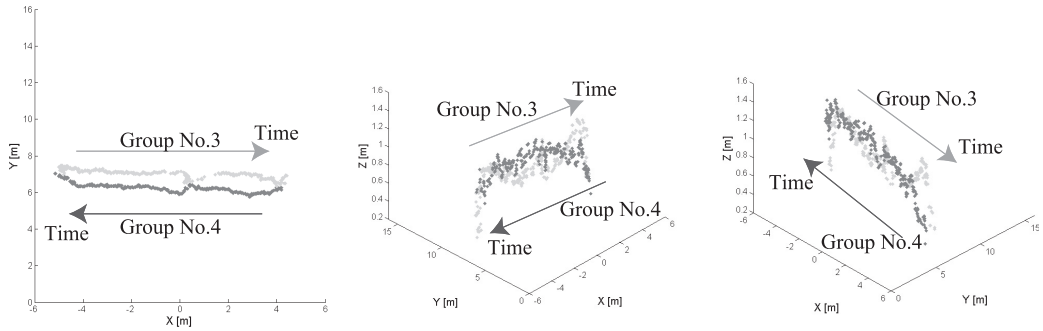


Fig. 5 Measured 3D trajectories obtained from the crossing experiment (cyan: Group No.3, magenta: Group No.4).

where Σ_i represents 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.

4. Experiment

The effectiveness of the proposed tracking method is evaluated in three scenes. In the first, two groups walk toward one another and then across. In this scene, data associations of each feature point become unstable because of the partial occlusion. In the second, two groups merge into one. In the third, a group is tracked walking up stairs. This experiment shows the effectiveness of the proposed tracking method, which can be used in a 3D environment. The evaluation is undertaken using the offline movies taken at 30 frames per second for each scene. The number of frames used for the first and second scenes is 400, and 500 for the third scene. The movies for the first and second scenes are taken at outdoors on a sunny day. And for the third scene, the movie is taken indoors.

Each experiment was conducted in the following settings. A stereo camera used for the experiments is Point Grey Research Bumblebee2 (color, f=3.8[mm], 320 × 240). The camera was set at a height of 5.1[m] with a 40[deg] downward tilt in the first and second experiments. For the third experiment, the camera was set at a height of 2.1[m] with a 40[deg] downward tilt. The process noise, measurement noise, and 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 with a radius of 1.0[m] for the first and second experiments and 0.5[m] for the third experiment. The size of search region is experimentally decided depending on a scene. The measured feature points are used for the state update of the tracker when there were more than two feature points associated with the tracker. The ellipses shown in experimental results (Fig. 4 (b), Fig. 6 (b), Fig. 8 (b)) are drawn to involve every feature point associated as a same group.

4.1 Groups Crossing Each Other

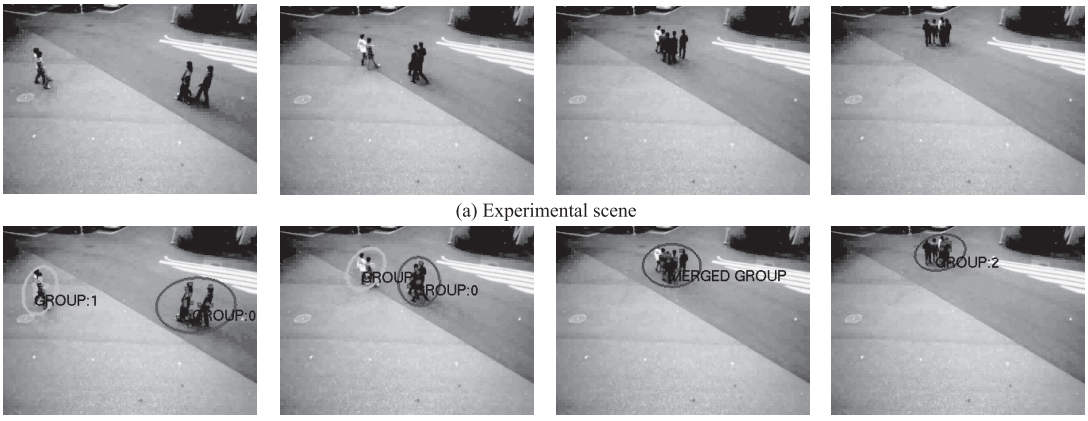
The effectiveness of the proposed method is evaluated in a scene in which two groups move along opposite directions and cross. In the scene, groups are partially occluded, and the associations of each feature point become unstable.

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.

4.2 Groups Merging into a Single Group

We verified that two groups merge into a single group with our proposed 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 center of the image.

The experimental result is shown in Fig. 6 (b), and the 3D



(a) Experimental scene

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

(b) Experimental result (two groups merge)

Fig. 6 Experimental result in which groups merge into one.

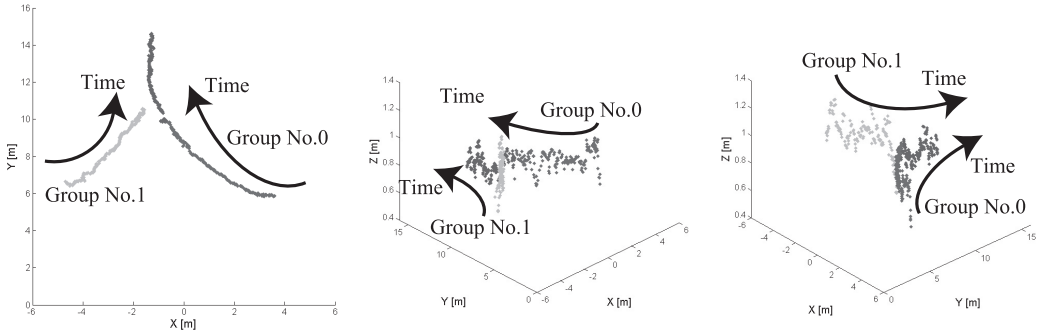
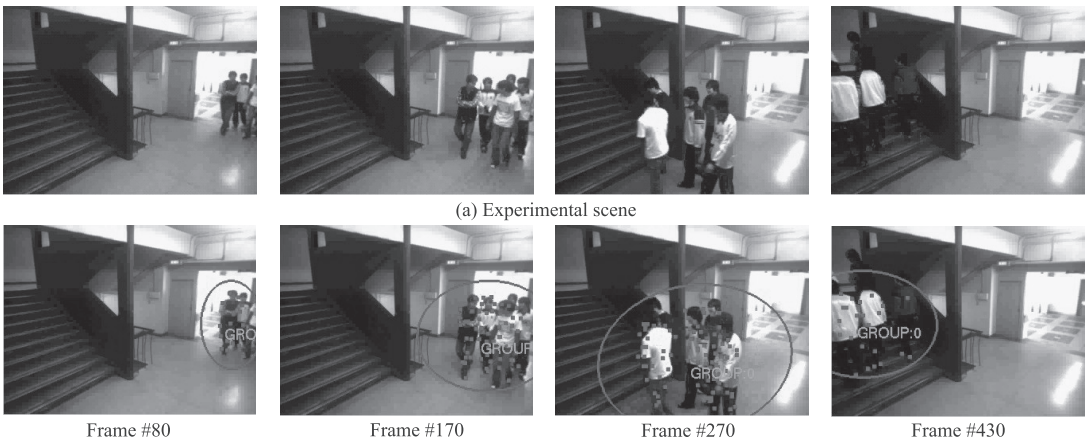


Fig. 7 Measured 3D trajectories obtained from the merging experiment (red: Group No.0, green: Group No.1, blue: merged Group No.2).



(a) Experimental scene

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

(b) Experimental result (3D environment)

Fig. 8 Experimental result in which individuals are measured in a 3D environment.

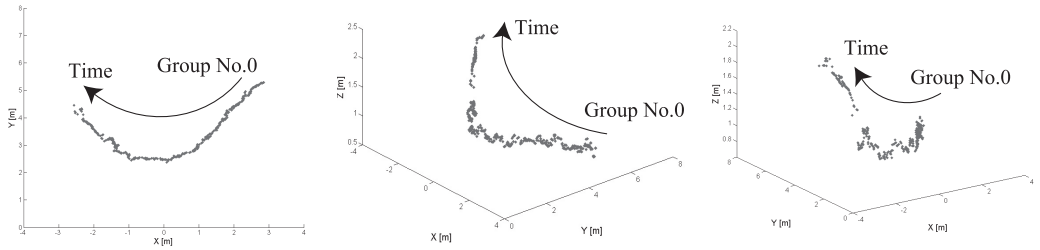


Fig. 9 Measured 3D trajectories of the group going up stairs in a scene.

trajectories of each group are shown in Fig. 7. As shown in Fig. 6 (b), two groups (Group 0 and Group 1) are merged into a single group (Group 2), and the effectiveness of the proposed method is verified. Compared with the experimental result obtained from the crossing experiment, this result also shows that groups merge only when they have a similar state. Although

Table 1 Position estimation of group tracking.

	Average error [m]	Standard deviation
Experimental scene 1 (Fig. 4)	0.119	0.0765
Experimental scene 3 (Fig. 8)	0.105	0.0756

(a) Crowded scene (b) Tracking result

Fig. 10 An example of group tracking result in a crowded scene.

two groups have similar positional information during crossing, they do not merge into a single group.

4.3 Tracking of a Group in a 3D Environment

To verify the effectiveness of the proposed method in a 3D environment, a group is tracked in a scene including stairs. As shown in Fig. 8 (a), the group walks on a plane floor first and then 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 a 3D environment. Since the tracking methods with a single lens camera such as [4] and [5] are difficult to use in such a 3D environment, our proposed method would be useful for tracking in a 3D environment.

The position estimation error of the group in the three dimensional scene is shown in Table 1. Comparing with the position estimation error in the normal situation (experimental scene 1), the position estimation error tends to be similar. The result shows that the proposed method is able to estimate group position stably even in a three dimensional environment.

5. Conclusions and Future Work

In this paper, we have proposed a new tracking method using a Kalman filter based tracker with 3D feature points obtained by subtraction stereo and KLT. The effectiveness of the new tracking system was verified in the following experiments. In the first one, in which groups of individuals moved in opposite directions and crossed each other, we verified that the proposed method could be used even when there were partial occlusions in a scene using velocity and position information for data association. In the second one, in which groups of people merge into a single group, we have verified the effectiveness of the proposed method for the merger of two groups. This experiment also showed that the velocity of groups is useful for the merging of groups and the data association of the tracking. In the third experiment, in which people walk up stairs, we verified that the tracking method can be used in a complex 3D scene. Although the effectiveness of the method was verified in these experiments, it is still difficult to apply in crowded conditions because the features used for data association are limited to group position and velocity. An example of a tracking failure in a crowded condition is shown in Fig. 10. In this crowded scene, people are too close to each other and they are difficult to be grouped correctly. Therefore, for future work, the authors are going to add other features to the data association and apply the method in crowded conditions, such as real city environments.

References

[1] M. Rodriguez, S. Ali, and T. Kanade: Tracking in unstructured crowded scenes, *Proc. IEEE ICCV2009*, pp. 1389–1396, 2009.
[2] 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. IEEE ICCV2009*, pp. 1467–1474, 2009.
[3] C. Tomasi and T. Kanade: Detection and tracking of point features, *Technical Report*, CMU-CS-91-132, Carnegie Mellon University, 1991.
[4] C. Tomasi and J. Shi: Good features to track, *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.
[5] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool: Robust tracking-by-detection using a detector confidence particle filter, *Proc. IEEE ICCV2009*, pp. 1515–1522, 2009.
[6] A. Ess, K. Schindler, B. Leibe, and L. van Gool: Improved multi-person tracking with active occlusion handling, *Proc. IEEE ICRA2009 Workshop on People Detection and Tracking*, pp. 54–59, 2009.
[7] O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer: Pedestrian detection and tracking for counting applications in crowded situations, *Proc. IEEE AVSS2006*, pp. 70–75, 2006.
[8] M. Yang, F. Lv, W. Xu, and Y. Gong: Detection driven adaptive multi-cue integration for multiple human tracking, *Proc. IEEE ICCV2009*, pp. 1554–1561, 2009.
[9] R.M. Salinas, E. Aguirre, and M. Garcia-Silvente: People detection and tracking using stereo vision and color, *Image and Vision Computing*, Vol. 25, Issue 6, pp. 995–1007, 2007.
[10] 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.
[11] 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.
[12] Y. Hoshikawa, K. Terabayashi, and K. Umeda: Human tracking using color information and subtraction stereo, *Proc. AWSVCI2009*, pp. 5–8, 2009.
[13] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT press, 2006.
[14] K. Umeda, T. Nakanishi, Y. Hashimoto, K. Irie, and K. Terabayashi: Subtraction stereo —A stereo camera system that focuses on moving regions—, *Proc. SPIE-IS&T Electronic Imaging*, Vol. 7239, 723908, 2009.
[15] G. Gennari and G.D. Hager: Probablistic data association methods in visual tracking of groups, *Proc. IEEE CVPR2004*, Vol. 2, pp. 876–881, 2004.

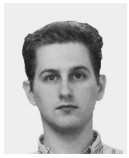
Yuma Hoshikawa

He received his B.S. degree from Chuo University, Japan, in 2009. From 2009, he is in the master course of Chuo University. His research interests include image processing. He is a member of JRM and JSPE.

Yuki HASHIMOTO

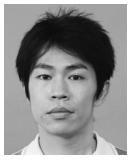
He received his B.S. and M.S. degrees from Chuo University, Japan, in 2008 and 2010, respectively. His research interests include image processing.

Alessandro MORO



He received his M.S. degree from University of Udine, Italy, in 2006. From 2007, he is in the doctor course of University of Trieste. His research interests include robotics and image processing.

Kenji TERABAYASHI



He received his B.S. and M.S. degrees from Hokkaido University, Japan, in 2002 and 2004, respectively. He recieved his Ph.D. from the University of Tokyo in 2008. In 2008, he joined the faculty of Chuo University, where he is currently an Assistant Professor of Department of Precision Mechanics. His research interests include human machine interface, robot vision, and virtual reality. He is a member of IEEE, RSJ, JSPE, VRSJ, and so on.

Kazunori UMEDA

He received his B.S., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 1989, 1991, and 1994, respectively. In 1994, he joined the faculty of Chuo University, where he is currently a Professor of Department of Precision Mechanics. His research interests include robot vision, range image processing, and human machine interface. He is a member of IEEE, RSJ, JSPE, JSME, and so on.