Proc. IEEE Int. Conf. on Multisensor Fusion and Integration
in Intelligent Systems, Baden-Baden, Germany, Aug. 2001.

Optical Flow-Based Person Tracking by Multiple Cameras

Hideki Tsutsui, Jun Miura, and Yoshiaki Shirai
Department of Computer-Controlled Mechanical Systems,
Osaka University, Suita, Osaka 565-0871, Japan
{h-tsutsu, jun, shirai}@cv.mech.eng.osaka-u.ac.jp

Abstract

This paper describes an optical flow-based person tracking method using multiple cameras in indoor environments. There are usually several objects in indoor environments which may obstruct a camera view. If we use only one camera, tracking may fail when the target person is occluded by other objects. This problem can be solved by using multiple cameras. In our method, each camera tracks the target person independently. By exchanging information among cameras, the three dimensional position and the velocity of the target are estimated. When a camera loses the target by occlusion, the target position and velocity in the image are estimated using information from other cameras which are tracking the target.

1 Introduction

To track a moving object from an image sequence is one of the most important problems in computer vision. Visual object tracking is useful for various applications such as visual surveillance and gesture recognition.

In a usual indoor environment, the target object is often occluded by other objects; therefore, a tracking method is needed which can robustly continue tracking even under occlusion.

Yamamoto et al. [2] proposed a method of multiple object tracking based on optical flow. By tracking all moving objects, the method can track a person even when he is occluded by another person. However, since object extraction uses optical flow, only moving objects can be extracted. Moreover, if the target person is occluded by a stationary object, the method cannot track him.

Several works (e.g., Rao and Durrant-Whyte [3], Utsumi et al. [4], Ukita et al. [5]) have proposed to use multiple cameras for multiple person tracking. In these works, an object region in the image is extracted by subtraction of the background image from the current image. Thus, when multiple objects overlap in one image, the system cannot distinguish them. Kato et al. [6] proposed a method of multiple persons tracking by multiple cameras based on the recognition of the target face; in this method, each camera has to be able to observe the target's face.

This paper proposes an optical flow-based person tracking method using multiple cameras. Using optical flow is an effective way to distinguish multiple moving objects. By using multiple cameras, the system can track the target even if the target is not observed by several cameras.

In this paper, we use the following terms: a *tracking camera* is the camera which is tracking the target person; a *lost camera* is the camera which has lost the target. Basically each camera tracks the target independently. When a camera loses the target due to occlusion, it (i.e., the *lost camera*) searches for the target in the image using information of the target position and velocity obtained by the other *tracking cameras*. This method of exchanging information between cameras realizes a robust person tracking in a cluttered environment.

2 Tracking Moving Object in the Image

Optical flow is calculated based on the generalized gradient method using multiple spatial filters [1]. We assume that an object region in the image has almost uniform flow vectors. An object is tracked by updating a rectangular window which circumscribes the object region.

We consider that the moving object which comes into the view first is the target, and set a window at the region where an enough number of flow vectors are obtained. We call this window the *tracking window*. The tracking proceeds by the following steps (see Figure 1):

1. The tracking window in the previous frame is shifted by the mean flow of that frame. This shifted window is called the *prediction window*. In the initial frame, the prediction window is the same as the tracking window.
2. The mean flow is calculated in the prediction window.
3. Pixels whose flow vectors are similar to the mean flow are searched for in the prediction window and its neighborhood. The object region is generated as the set of such pixels.
4. The tracking window is set to circumscribe the object region.

Figure 2 shows a result of optical flow-based person tracking.

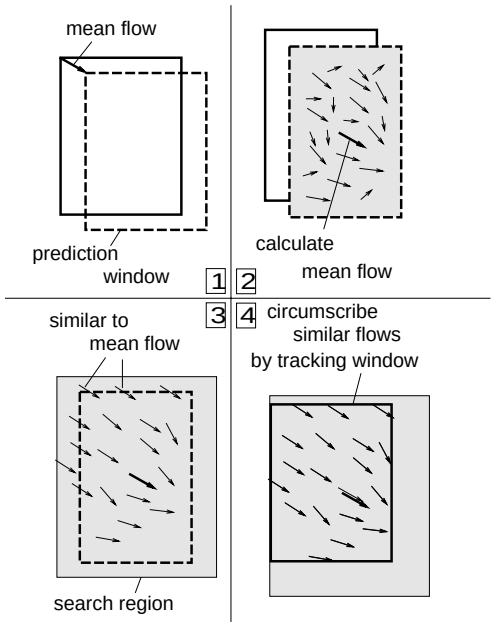


Figure 1: Tracking an object in the image.

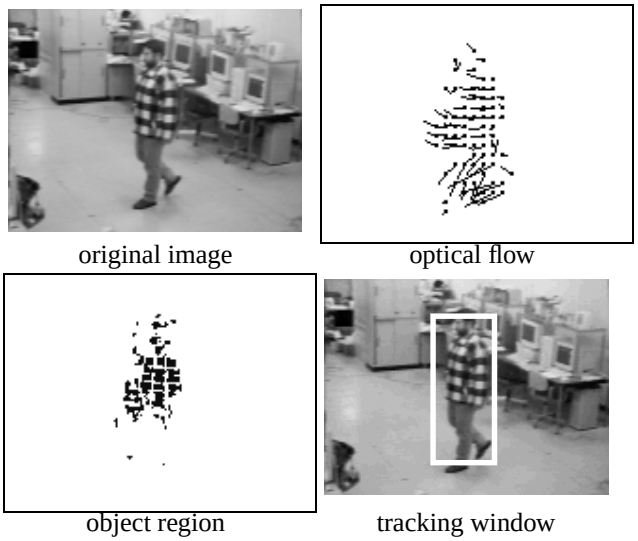


Figure 2: Extraction of the object region.

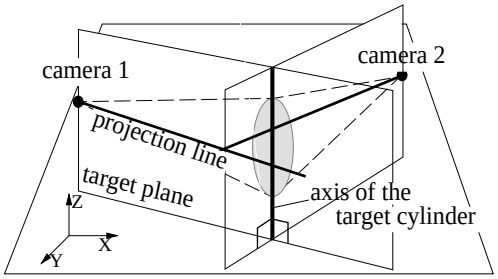


Figure 3: Estimating the target position.

3 Tracking Target by Multiple Cameras

3.1 Target Position Estimation by Multiple Cameras

We model a person by a vertical cylinder and its height and radius are set to the height and the width of the person.

Let $\boldsymbol{X} = [X, Y, Z, 1]^t$ denote a point in the world coordinate system, and $\boldsymbol{x} = [x, y, 1]^t$ denote the projected point of $\boldsymbol{X}$ in the image. The following equation is satisfied:

$$h\boldsymbol{x} = \boldsymbol{C}\boldsymbol{X}, \tag{1}$$

where h denote a scale factor and $\boldsymbol{C}$ denote the camera parameter, which are calibrated in advance:

$$\boldsymbol{C} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \end{bmatrix}. \tag{2}$$

If the object region is correctly extracted, the center of gravity of the object region $\boldsymbol{x}$ is determined. The corresponding three dimensional position $\boldsymbol{X}$ of the target satisfies the equation (1). If there are at least two tracking cameras, the target position is obtained as the intersection of projection lines from such tracking cameras. However, the vertical position of the center of gravity in the image is not reliable; this is because the most part of the target's feet has different velocity from that of his body, as shown in Figure 2, and is often excluded from the object region. We, therefore, use only the horizontal position of the center of gravity to estimate the target position.

Instead of a projection line, we consider the plane which contains the projection line and is perpendicular to the floor. We call it a *target plane*. The axis of the target cylinder is obtained on this plane. A target plane is given by the following equation:

$$\boldsymbol{a} \begin{bmatrix} X \\ Y \end{bmatrix} = b, \tag{3}$$

where

$$\boldsymbol{a} = \begin{bmatrix} (C_{11} - C_{31}x)(C_{23} - C_{33}y) - (C_{13} - C_{33}x)(C_{21} - C_{31}y) \\ (C_{12} - C_{32}x)(C_{23} - C_{33}y) - (C_{13} - C_{33}x)(C_{22} - C_{32}y) \end{bmatrix}^t, \tag{4}$$

$$b = (C_{13} - C_{33}x)(C_{24} - C_{34}y) - (C_{14} - C_{34}x)(C_{23} - C_{33}y). \tag{5}$$

When there are at least two tracking cameras, the axis of the target cylinder is obtained as the intersection line of such target planes (see Figure 3). If we have more than two target planes, the axis position is calculated by the least squares method as follows:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{B}, \tag{6}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1 \boldsymbol{a}_2 \boldsymbol{a}_3 \cdots \boldsymbol{a}_m]^t$ $\boldsymbol{B} = [b_1 b_2 b_3 \cdots b_m]^t$, and $\boldsymbol{a}_i$ b_i denote parameters of the target plane for the ith camera, m is the number of tracking cameras.

There are two cases where the above method does not work. One is the case where there is only one tracking camera. The other is the case where the target planes make small angles; in this case, the determinat $|\boldsymbol{A}^t \boldsymbol{A}|$ is very small and the estimated position is not reliable. The next section describes how to cope with such two cases.

3.2 Target Area Estimation by Single Camera

By considering the size of the object region in the image and the human model, the target area can be estimated to some extent by using the information from only one tracking camera.

Estimating Range of Depth As described in Section 2, usually a part of the target is extracted and is circumscribed by the tracking window. in the image, the top of the head is not lower than that of the tracking window, and the bottom of the foot is not higher than that of the tracking window. These facts are used to estimate the possible depth range of the target.

In the target plane (see Figure 4), let L_h and L_f denote the projection line through the top and the bottom of the tracking window, respectively.

First, we consider the plane $Z = Z_{wh}$, which indicates the height of the target person, and calculate the intersection M_h of this plane and L_h. Since the top of the head is not lower than that of the tracking window in the image, the person should be farther than M_h. Thus, plane P_h, which includes M_h and is perpendicular to both the target plane and the floor, represents one constraint on the target depth. Similarly, considering the condition on the foot position, we can obtain the other constraining plane P_f from line L_f and plane $Z = 0$. The target exists between P_h and P_f.

Estimating Range of Width Since the tracking window circumscribes a part of the target, the leftmost and the rightmost edge of the target never be inside the tracking window. Figure 5 shows the top view of Figure 4. When the rightmost edge of the target is on line $A_r B_r$, the axis of the human model is on line $C_l D_l$, such that these lines are in parallel and the distance between the lines is equal to the radius of the model. Line $C_l D_l$ represents the left boundary. Similarly, we can obtain line $C_r D_r$ as the right boundary.

The target area which is constrained by the ranges of depth and width becomes a tetragon $C_r C_l D_l D_r$. Figure 6(a) is an example of target area estimation.

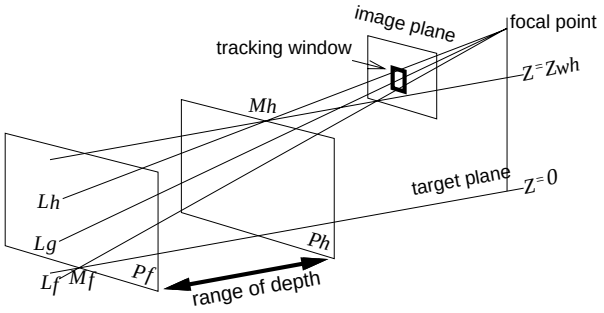


Figure 4: Estimating the range of depth.

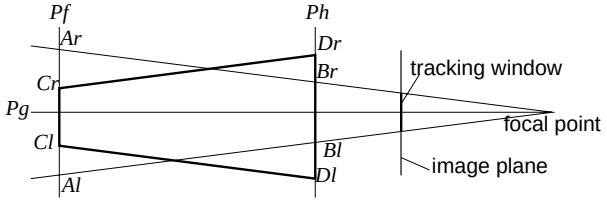


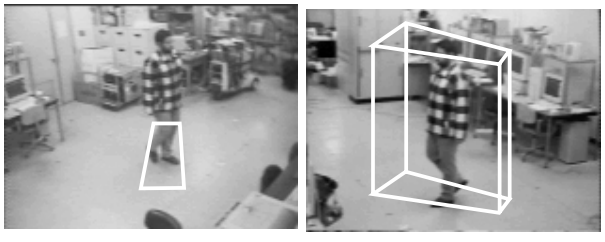
Figure 5: Estimating the range of width.

3.3 Determination of Estimated Window for a Lost Camera

For a lost camera, we estimate the area in the image where the target will be projected using the target position on the floor obtained by tracking cameras. This area is called *the estimated window*. A lost camera searches the estimated window for the target.

When the target position is estimated by other multiple tracking cameras (see Section 3.1), we put the human model at the target position and calculate its projection to the image of a lost camera by equation (1); the window circumscribing the projected region is considered as the estimated window.

When the target area is estimated by a single tracking camera (see Section 3.2), we put the human model at every possible position inside the target area and calculate its projection to the image of the lost camera. Then the



camera A camera B

Figure 6: Target area and its projection.

window circumscribing the union of all projected regions is considered as the estimated window. Figure 6(b) shows the projected region for camera B which is generated from the target area estimated by camera A.

3.4 Target Identification by Target Velocity

A lost camera searches the estimated window for the target. However, if there is another moving object in this window, the lost camera may mistakenly track him as the target. To discriminate the target from other moving persons, we use the target velocity.

3.4.1 Estimation of Target Velocity

The velocity in an image is derived by differentiating equation (1):

$$\boldsymbol{v} = -\frac{\dot{h}}{h^2}\boldsymbol{C}\boldsymbol{X} + \frac{1}{h}\boldsymbol{C}\boldsymbol{V} \tag{7}$$

where $\boldsymbol{V} = \dot{\boldsymbol{X}} = [U, V, W]^t$ denote a three dimensional velocity, $\boldsymbol{v} = \dot{\boldsymbol{x}} = [u, v]^t$ denote a velocity in the image.

When there are at least two tracking cameras, a set of the constraint equations about $\boldsymbol{V}$ is derived by substituting each mean flow which is estimated by tracking cameras for $\boldsymbol{v}$ at equation (7). The target velocity $\boldsymbol{V}$ is then calculated by applying the least squares method to the equations.

When there is only one tracking camera, the target velocity $\boldsymbol{V}$ cannot be calculated by the above method. In this case, by assuming that the target have no vertical velocity, we calculate $\boldsymbol{V} = [U, V, 0]$ by solving equation (7).

3.4.2 Estimation of Target Velocity in the Image

For lost cameras, the target velocity $\boldsymbol{v}$ in the image is predicted from $\boldsymbol{V}$ using equation (7). Each lost camera searches for flow vectors which are similar to $\boldsymbol{v}$. Even if another object is moving in the estimated window at a different velocity, the camera can discriminate it from the target.

Figure 7 is an example of effective use of target velocity. In this figure, "+" shows the center of gravity of the object region. A solid window shows a tracking window, and a dashed window shows an estimated window in a lost camera.

In camera C, the target is occluded by an obstacle. However, an appropriate position is searched by using the estimated window which is given by the other cameras. At the same time, there is another person with a leftward velocity in the image. Camera C correctly judges that the person is not the target because the predicted velocity $\boldsymbol{v}$ is rightward.

4 Evaluating the Reliability of Tracking for Robust Information Integration

If a camera mistakenly tracks a person other than the target, the integration result of information from multiple tracking cameras will be degraded even if the other cameras are correctly tracking the target. Therefore, we evaluate the reliability of tracking of each camera on-line and, if the tracking of a camera is not reliable, we do not use the information from that camera for estimating the position and the velocity of the target. In addition, for such an *unreliable* camera, information from the other tracking cameras are used for prediction, as in the case of lost cameras described in the previous section.

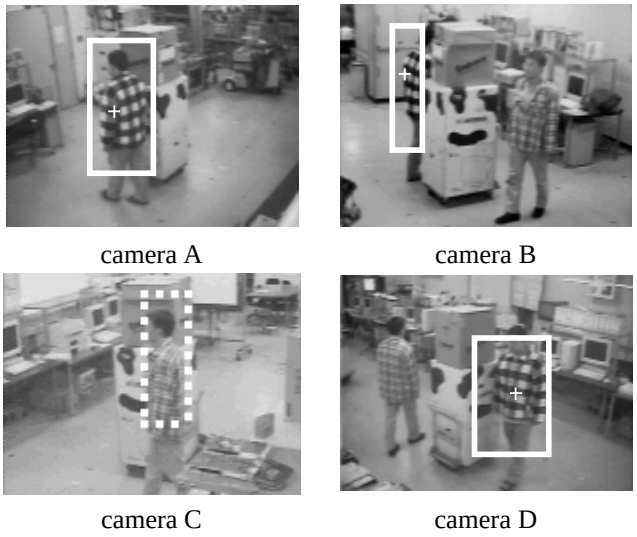


Figure 7: Discriminate the target from another person using velocity information.

If the velocity of the target and those of other persons are different in the image, they can easily be distinguished as shown in Figure 7. Difficult situations, therefore, arise when the target and another person have similar velocities. If both persons overlap with each other in such a situation, we judge that the current tracking is unreliable because the target position in the image is unreliable.

To evaluate the reliability of tracking, we have to detect the beginning and the end of an overlap. We assume that another person gradually approaches and parts from the target before and after an overlap; we, therefore, do not deal with the case where, for example, another person suddenly appears near the target from behind an obstacle. Under this assumption, the period of an overlap is recognized as follows:

1. Detect the beginning of an overlap by tracking other persons near the target.
2. Detect the end of the overlap by:

   (a) a large difference between the size of the tracking window and that of the estimated one in the case where the target is not occluded by an obstacle.

(b) a large difference between the position of the tracking window and that of the estimated one in the case where the target is occluded by an obstacle.

However, since the camera cannot know if the target is occluded during the overlap, the camera checks both conditions actually.

We will examine each step in the following.

4.1 Detecting Overlap with Other Persons

In order to detect an overlap with other persons, we also track them. However, since tracking all persons in the image is costly, we track other persons only when they are near the target being tracked or the estimated window. In addition, we do not calculate their three-dimensional position and velocity since this tracking of other persons is only for evaluating the reliability of tracking of the target.

For tracking other persons, two search areas are set on the both side of the target as shown in Figure 8; each search area is the width of the target apart from the tracking window so as not to extract the target's flow. In each search area, the mean flow is calculated. If the total area of the regions which have the similar flow to the mean flow is large enough, another person is considered to appear in the search area and the tracking for the person starts. This tracking terminates when another persons has gone away from search areas.

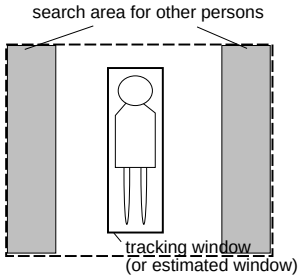


Figure 8: Search areas for other persons.

Note that after an overlap is detected, the tracking continues as usual although it is considered to be unreliable, because the information of the tracking window is later used for detecting the end of the overlap.

4.2 Examining Size Difference between Tracking Window and Estimated Window

If the target and another person overlap with each other, the tracking window is set to circumscribe both persons. If the camera continues the usual tracking process (see Section 2), however, the size of the tracking window grows as the two persons get apart from each other. So we use the width of the tracking window to determine if the window circumscribe only the target (see Figure 9).

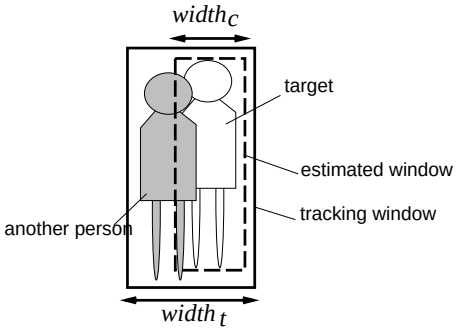


Figure 9: Tracking window for overlapping persons.

Let $width_t$ denote the width of the tracking window and $width_c$ denote that of the estimated window. When the ratio

$$\frac{width_t - width_c}{width_c} \tag{8}$$

exceeds a certain threshold (currently, 1), the camera searches only the estimated window for the target, and if the target is extracted separately from the other person, then the tracking of the camera is considered to become reliable again.

4.3 Examining Positional Difference between Tracking Window and Estimated Window

When the target is occluded by an obstacle and another person exists in the estimated window, that person will be tracked. In such a case, the camera continuously checks the distance between the estimated position and the position of the tracking window. When the distance becomes large (see Figure 10), the camera searches only the estimated window for the target, and if the target is extracted separately from the other person, then the tracking of the camera is considered to become reliable again.

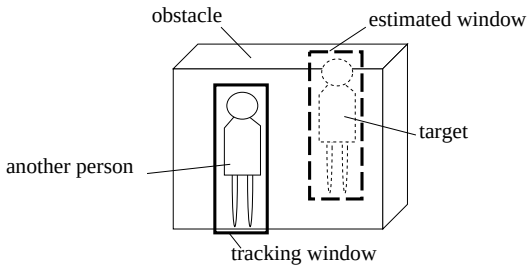


Figure 10: Tracking window in the case where the target is occluded.

5 Experiments

The proposed method has been tested with videotaped image sequences captured in our laboratory. Figure 11 shows a tracking result. There are a target and another walking person who has similar velocity in the image, and other three persons who are standing. The figure shows the images taken by camera A (see Figure 12). At the first frame, the target is at the upper left of the image, circumscribed by a white rectangle. At the 60th frame, the target and another person overlap and the tracking window circumscribes both. As the two person get apart from each other, the size of the tracking window grows. At the 101st frame, since the value of equation (8) exceeds the threshold, the camera starts trying to find the target in the estimated window, and then at the 130th frame, the target is found again and the reliable tracking restarts. Figure 12 shows the trace of the target position in this tracking.

When we estimate the target position using more than one tracking cameras (see Section 3.1), we assume that the horizontal position of the center of gravity of the object region is reliable. However, if the object region is not completely extracted due to, for example, partial occlusion or image noise, the target position may be less reliable; this is the cause of most of random perturbations in the trace in Figure 12. One way to deal with such perturbations is to apply the target area extraction (see Section 3.2) to multiple camera cases; that is, the target area can also be calculated as the intersection of target areas obtained by all tracking cameras.

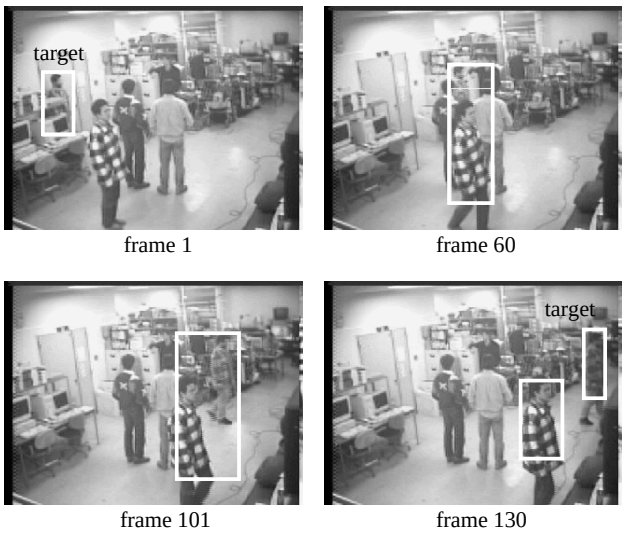


Figure 11: Tracking result.

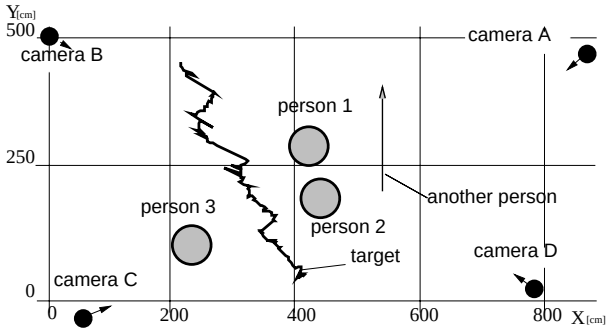


Figure 12: Top view of person's trace.

6 Conclusion and Future Works

This paper has described a person tracking method by multiple cameras based on optical flow. In this method, since information on the target is shared among cameras, a camera which failed to track the target can get information from the other tracking cameras to predict the target position and velocity in the image. We have also described the conditions to judge that the tracking by a camera is not reliable even when the tracking in the image seems correct. The experimental results show that a target person can be tracked robustly without mixing him up with other persons even in case of occlusion.

We have already developed a realtime tracking system for multiple targets using single camera [2]. One future work is to realize a realtime implementation of the proposed method. Another future work is to extend the method so that multiple persons are tracked at the same time.

References

[1] H. Chen, Y. Shirai, and M. Asada: Detecting Multiple Rigid Image Motions from an Optical Flow Field Obtained with Multi-Scale, Multi-Orientation Filters. IEICE Trans. Vol. E76-D No. 10, pp. 1253-1262 (1993).

[2] S. Yamamoto, Y. Mae, Y. Shirai, and J. Miura: Realtime Multiple Object Tracking Based on Optical Flows, Proc. Int. Conf. on Robotics and Automation, pp. 2328-2333 (1995).

[3] B. S. Rao and H. Durrant-Whyte: A Decentralized Bayesian Algorithm for Identification of Tracked Targets, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 23, No. 6, (1993).

[4] A. Utsumi, H. Mori, J. Ohya, M. Yachida: Multiple-Human Tracking using Multiple Cameras. Proc. Third IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 498-503 (1998).

[5] N. Ukita, T. Nagao, and T. Matsuyama: Versatile Cooperative Multiple-Object Tracking by Active Vision Agents. Proc. MVA2000, pp. 353-358 (2000).

[6] T. Kato, Y. Mukaigawa, and T. Shakunaga: Cooperative Distributed Tracking for Effective Face Registration. Proc. MVA2000, pp. 353-358 (2000).