Proc. 4th Int. Conf. on Autmatic Face and Gesture Recognition
pp. 336-341, Grenoble, France, March 2000.

Tracking a Person with 3-D Motion by Integrating Optical Flow and Depth

R. Okada
TOSHIBA Kansai Research Center
8-6-26 Motoyama-Minami-Cho,
Higashinada-Ku, Kobe 658-0015 Japan
okada@krl.toshiba.co.jp

Y. Shirai J. Miura
Department of Computer Controlled Machinery,
Osaka University, 2-1 Yamadaoka,
Suita-Shi, Osaka, 565-0871, Japan
{shirai, jun}@mech.eng.osaka-u.ac.jp

Abstract

This paper describes a method of tracking a person with 3-D translation and rotation by integrating optical flow and depth. The target region is first extracted based on the probability of each pixel belonging to the target person. The target state (3-D position, posture, motion) is estimated based on the shape and the position of the target region in addition to optical flow and depth. Multiple target states are maintained when the image measurements give rise to ambiguities about the target state. Experimental results with real image sequences show the effectiveness of our method.

1. Introduction

Visual object tracking is necessary for various applications such as autonomous vehicle navigation and human interface. While many tracking methods have been proposed, tracking using a single cue such as optical flow[1], depth[2], or edges[3], fails when the target cannot be identified based on the employed cue. Etoh et al.[4] have proposed to use multiple cues, which are color, position, and intensity gradients. Okada et al.[5] integrate optical flow and depth to extract the target region, and Yamane et al.[6] integrate optical flow and uniform intensity regions. In these methods, the target which cannot be correctly extracted from one cue can be tracked by using the other cues. However, since these methods assume that flow vectors in the target region are almost uniform, implying that the target translates almost parallel to the image plane, tracking may fail when the target moves with general 3-D motion.

In order to track a target with 3-D motions, several methods based on optical flow have been proposed[8][9][10]. However, the translation along the optical axis is less sensitive to optical flow than the translations parallel to the image plane. Also, the rotation upon an axis perpendicular to the optical axis generates the flow vectors similar to the ones generated from the translation parallel to the image plane (see Fig. 1). Therefore, it is still difficult to reliably estimate every component of general 3-D motion from optical flow alone[11]. On the other hand, the translation along the optical axis can be directly estimated from the depth. The rotation angle upon an axis perpendicular to the optical axis could also be computed if the resolution of the depth map were high enough and the 3-D target shape were known. Although such a depth map with sufficient resolution cannot be obtained by conventional methods, since the change of this rotation angle causes the shape change of the target region in the image, it can be an effective cue for estimating this rotation.

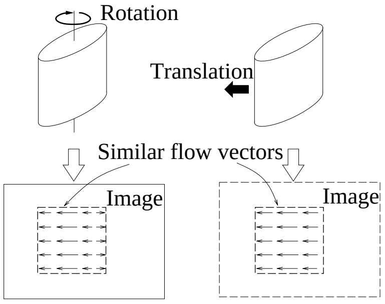
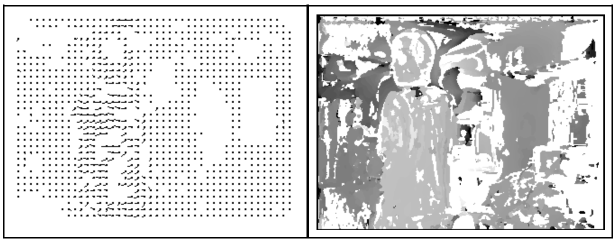


Figure 1. Similar optical flow

In order to deal with the issues stated above, we propose to extract the reliable target region by integrating optical flow and depth and then estimate the target state (3-D position, posture, and motion) using the shape of the target region, optical flow and depth. Although none of them alone can estimate the 3-D target state reliably, that they compensate for each other. In this paper the rotation axis vertical to the floor is



(a) Optical flow (Lines show flow vectors 7 pixels apart.)

(b) Disparity (Bright pixels have large disparities.)

Figure 2. Calculated flow vectors and disparities at 13th frame in Fig. 7 (flow vectors or disparities cannot be calculated in white regions)

the only one under consideration since our goal is to track a walking person. However the proposed method of integrating multiple cues is also applicable to other rotation axes.

2. Optical Flow and Disparity

Optical flow and disparity (depth information) are computed by obtaining the correspondences of points between two successive frames and between a pair of stereo images respectively. We use conventional SAD based method[12] to obtain the correspondences. Flow vectors and disparities are ignored at the pixels where reliable correspondences cannot be obtained. Fig. 2 shows an example of calculated optical flow and disparity. Since they are noisy and not calculated in the full region, none of them alone can estimate reliable 3-D states of the target.

3. Target State and Observations

The coordinate system and the 3-D shape model of the target person are shown in Fig. 3. We use a coarse body model rather than an accurate one in order to track a variety of people with various physical constitutions.

We estimate the translation $(t_x\ t_y\ t_z)^T$ and the rotation $(r_x\ r_y\ r_z)^T$ of the body coordinate system (fixed to the body) in the world coordinate system (fixed to the scene) and their velocities, which determine the target state. The target state vector is

$$\boldsymbol{x} = (\boldsymbol{q}^T\ \dot{\boldsymbol{q}}^T)^T, \quad \boldsymbol{q} = (r_x\ r_y\ r_z\ t_x\ t_y\ t_z)^T, \qquad (1)$$

where $\dot{\boldsymbol{q}}$ represents the differential of $\boldsymbol{q}$ about the time. Note that $r_x = r_z = \dot{r}_x = \dot{r}_z = 0$ since XZ-plane of the world coordinate system is set to be parallel to the floor and the target person is considered to be standing vertically. The transformation between the camera

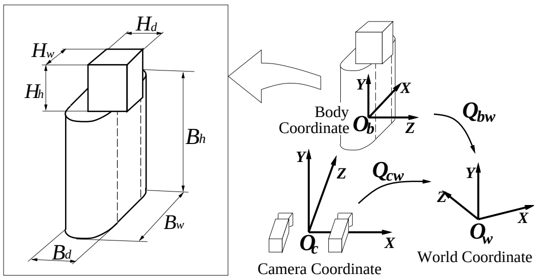


Figure 3. Coordinate system and person model (The head of the model is a parallelepiped, and the torso consists of two half cylinders and a parallelepiped.)

coordinate system (fixed to the camera) and the world coordinate system are known. The perspective projection model, whose parameters are also known, is used as the camera model.

We use flow vectors (u_i, v_i) and disparities d_i at pixels $\boldsymbol{p}_i$ in the target region in order to estimate the target state. Since the rotation upon the vertical axis cannot be reliably estimated from these observations as stated above, we employ *shoulder width* W_s in the image as the target shape to estimate this rotation. If the rotation angle upon the vertical axis and the vertical and horizontal positions are updated by simply adding their velocities to the previous states, the estimation error of the target motion is added to them and tracking may fail[9]. The shoulder width functions to correct the rotation angle which may include a large error. In addition, we employ the target position $\boldsymbol{p}_B^T$ in the image, which is the projected point of the origin $\boldsymbol{O}_b$ of the body coordinate and works for correcting the vertical and horizontal position. Thus, specifically the observation vector is represented as:

$$\boldsymbol{y}_t = (\ W_s \quad \boldsymbol{p}_B^T \quad \boldsymbol{o}_1^T \quad \cdots \quad \boldsymbol{o}_i^T \quad \cdots \quad \boldsymbol{o}_c^T\ )^T, \qquad (2)$$

where $\boldsymbol{o}_i = (u_i\ v_i\ d_i)^T$, and c represents the number of pixels in the target region.

4. Tracking Procedure

In each frame the state vector is estimated from the observation vector and the state vector in the previous frame. The tracking procedure in each frame is shown in Fig. 4. First, the current target state is predicted based on the previous state. A pair of stereo images are taken and the optical flow and the disparity are calculated. The target region is extracted by the predicted state, the optical flow and the disparity. Then the shoulder width and the target position in the image are computed out of the extracted target region. The

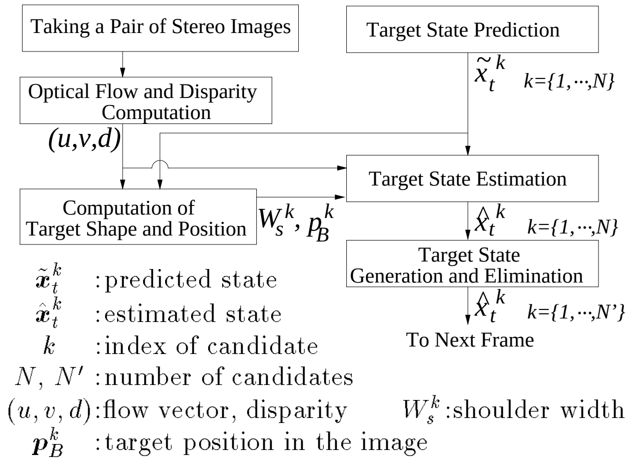

Taking a Pair of Stereo Images
Target State Prediction
$\tilde{x}_t^k$ k={1,···,N}
Optical Flow and Disparity Computation
(u,v,d)
Target State Estimation
Computation of Target Shape and Position
W_s^k, p_B^k
$\hat{x}_t^k$ k={1,···,N}
Target State Generation and Elimination
$\hat{x}_t^k$ k={1,···,N'}
To Next Frame
$\tilde{\boldsymbol{x}}_t^k$:predicted state
$\hat{\boldsymbol{x}}_t^k$:estimated state
k :index of candidate
N, N' :number of candidates
(u,v,d):flow vector, disparity
W_s^k:shoulder width
$\boldsymbol{p}_B^k$:target position in the image

Figure 4. Flow of tracking procedure

target state is estimated from these observations and the predicted target state. Initially, since the previous state vector is not available, we obtain the state vector assuming that the moving object is the target person. If the target state cannot be determined uniquely, possible state candidates are generated. The above tracking procedure is applied to each candidate. If the state candidates inconsistent with the current observation are detected, they are eliminated.

4.1. Target Shape and Position

Although many tracking methods use black background to simplify the segmentation of the target[13], we extract the reliable target region even in a complex background by integrating optical flow and depth. The extracted target region is used for computing the shoulder width and the position of the target in the image.

Target Region The target region is determined to be a set of pixels which have flow vectors and disparities similar to the predicted flow vectors and the disparities generated from the predicted target state. But the target region cannot be reliably extracted based on these observations alone because the obtained flow vectors and disparities are noisy (see Fig. 2). The target region should also be located at the position consistent with the past target states.

We deal with these conditions together by calculating the probability of each pixel $\boldsymbol{p}_i$ belonging to the target, which is called *target probability.* The target probability is calculated by using Bayes' theorem:

$$P(\boldsymbol{p}_i \in T|\boldsymbol{o}_i) = \frac{P(\boldsymbol{o}_i|\boldsymbol{p}_i \in T)P(\boldsymbol{p}_i \in T)}{P(\boldsymbol{o}_i)}, \qquad (3)$$

where T represents a set of pixels belonging to the target, $P(\boldsymbol{p}_i \in T)$ is the prior probability of a pixel $\boldsymbol{p}_i$ belonging to the target, which works for eliminating the pixels far from the predicted target region, $P(\boldsymbol{o}_i|\boldsymbol{p}_i \in T)$ is the likelihood of an observation, which works for integrating optical flow and disparity and $P(\boldsymbol{o}_i)$ is the probability of observing $\boldsymbol{o}_i$. These three terms are calculated as follows.

The likelihood $P(\boldsymbol{o}_i|\boldsymbol{p}_i \in T)$ is calculated from the probability distribution of predicted flow vectors and predicted disparities of the target, which is approximated by the normal distribution. The expectation $\tilde{\boldsymbol{o}}_i = (\tilde{u}_i\ \tilde{v}_i\ \tilde{d}_i)^T$ of the normal distribution is determined to be the flow vector and the disparity generated from the target 3-D model having the predicted target state $\tilde{\boldsymbol{x}}_t$. Since the error (variance $\tilde{X}_t$) between the observation and the expectation occurs from prediction error of the target states and the image noise, the covariance matrix $\tilde{O}_i$ of the normal distribution should be determined by taking these two errors into account. Given that the error caused by the image noise is a white noise with zero mean and variance W_{it} whose components are independent of each other,

$$\tilde{\boldsymbol{o}}_i = \boldsymbol{h}_i(\tilde{\boldsymbol{x}}_t), \quad \tilde{O}_i = H_i\tilde{X}_tH_i{}^T + W_{it}, \qquad (4)$$

where $\boldsymbol{h}_i$ is the vector function transforming the target state to flow vector and disparity, $H_i = \left.\frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{x}}\right|_{\tilde{\boldsymbol{x}}_t}$. Note that the errors in the flow vector components and that of the disparity are independent of each other because $H_i\tilde{X}_tH_i{}^T$ has non-zero covariances. W_{it} is determined based on the contrast because the reliabilities of the flow vector and the disparity depend on the contrast. $\tilde{\boldsymbol{x}}_t$ and $\tilde{X}_t$ are obtained from the prediction stage of the Kalman filter.

The prior probability is calculated by

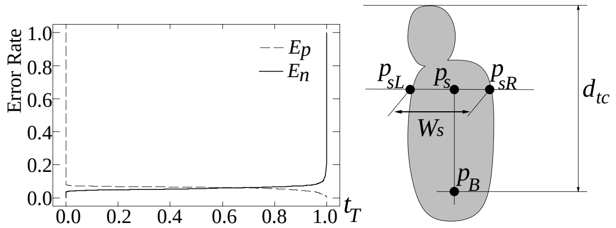
$$P(\boldsymbol{p}_i \in T) = \int_D r_i(\boldsymbol{x})p(\boldsymbol{x})\boldsymbol{dx}, \qquad (5)$$

$$\begin{cases} r_i(\boldsymbol{x}) = 1 & if\ \boldsymbol{p}_i \in R(\boldsymbol{x}) \\ r_i(\boldsymbol{x}) = 0 & else \end{cases}$$

where D represents a set of all the predicted target states, $R(\boldsymbol{x})$ is a target region generated by projecting the target model whose state is $\boldsymbol{x}$ to the image plane and $p(\boldsymbol{x})$ is the probability density of the predicted target state being $\boldsymbol{x}$. The probability distribution of the predicted target state is assumed to be the normal distribution with mean $\tilde{\boldsymbol{x}}_t$ and variance $\tilde{X}_t$ which are obtained from the prediction stage of the Kalman filter.

$P(\boldsymbol{o}_i)$ is calculated by Eq. (6), supposing the probability of every possible observation is equivalent if a pixel $\boldsymbol{p}_i$ does not belong to the target.

$$P(\boldsymbol{o}_i) = P(\boldsymbol{o}_i|\boldsymbol{p}_i \in T)P(\boldsymbol{p}_i \in T) + \mathcal{U}(\boldsymbol{o}_i)(1 - P(\boldsymbol{p}_i \in T)), \qquad (6)$$



(a)Relation between t_T and error rate (b)Shoulder width and target position

Figure 5. Target Region

where $\mathcal{U}$ represents the uniform distribution whose range is determined assuming that objects do not move too fast and are neither too close to the camera nor too far from the camera.

The target region S_t is determined to be a set of pixels with the target probability exceeding a threshold t_T. There are two kinds of error in extracting the target region, i.e., the error of determining the pixels in the background belonging to the target (represented by the error rate E_p) and the error of determining the pixels on the target belonging to the background (E_n). Fig. 5(a) shows these two error rates in a representative frame for various t_T. Since both E_n and E_p are small and insensitive to t_T while $0.2 < t_T < 0.8$, t_T is determined to be a value in this range.

Shoulder Width and Target Position $\boldsymbol{p}_{sL}$ and $\boldsymbol{p}_{sR}$ denote the boundary points of the target region on the horizontal line through the predicted shoulder position (see Fig. 5(b)). The shoulder width W_s is determined to be the distance between $\boldsymbol{p}_{sL}$ and $\boldsymbol{p}_{sR}$.

The target position in the image is defined to be the point which is on the vertical line through the midpoint of $\boldsymbol{p}_{sL}\boldsymbol{p}_{sR}$ and is d_{tc} below the top of the head, where d_{tc} is the distance in the image between the top of the head and the origin of the body coordinate.

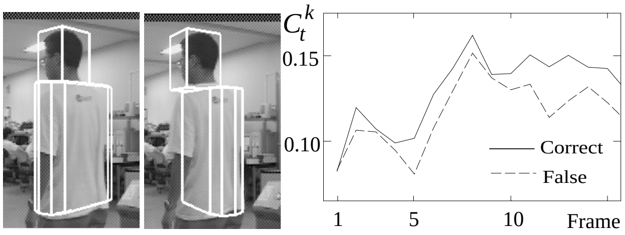
4.2. Target State Estimation and Prediction

In each frame, the target state is estimated from the observations and should also be consistent with the past target states. In order to achieve this behavior through the system, we utilize the Kalman filter. Although the employment of the Kalman filter is not the major issue in this paper, the implementation is discussed briefly.

Provided that the target velocity is almost constant between two successive frames, the system equation is

$$\boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + \boldsymbol{u}_t, \quad A \equiv \begin{bmatrix} I_6 & I_6 \\ 0 & I_6 \end{bmatrix}, \tag{7}$$

where I_6 represents the 6×6 unit matrix, $\boldsymbol{u}_t$ represents the prediction error, which is assumed to be a white nose with zero mean and variance U.



(a) Correct (b) False (c) Criterion of candidates

Figure 6. Criterion of candidates

The observation equation is

$$\boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{x}_t) + \boldsymbol{w}_t, \tag{8}$$

where $\boldsymbol{h}$ is the observation function transforming the state vector to the observation vector, $\boldsymbol{w}_t$ represents the observation error, which is assumed to be a white noise with zero mean and variance W_t. W_{it} in Eq. (4) is a submatrix of W_t and the variances of shoulder width and the target position are determined based on the smoothness of the target region boundary. Since $\boldsymbol{h}(\boldsymbol{x})$ is not linear function, we use the extended Kalman filter.

If the segmentation of the target region fails, the incorrect observations which is the shoulder width and target position in the image is obtained. But since the failure of the segmentation often causes the lack of smoothness of the target region boundary, the variances of the incorrect observations become large and they hardly affect the estimation of the target state. Thus the failures of the segmentation for several frames do not cause the serious tracking failure.

4.3. State Generation and Elimination

Multiple rotation angles upon the vertical axis are possible for a shoulder width because the observation function of shoulder width is not monotone. However, only one candidate, which may be incorrect, can be obtained because of the linearization of the observation function $\boldsymbol{h}$. This problem happens when the estimated rotation angle is around extrema. Therefore, when the estimated rotation angle comes out of the neighborhood of an extremum, another possible state candidate is generated to have symmetrical rotation angle to the original one with respect to the extremum. The covariance matrix of the new state candidate is determined to be equal to that of the original target state.

It is conceivable that the probability of each observation being derived from the false candidate is lower than that from the correct candidate. The probability of the jth element y_{jt}^k of the current observation

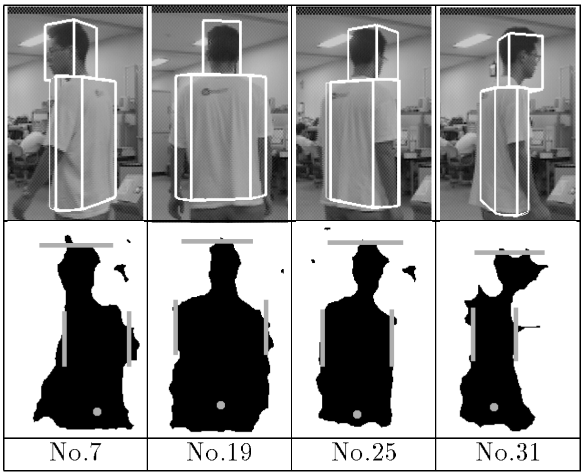


Figure 7. Tracking result (Upper figures show the model with estimated state. Lower ones show extracted target region (black), top of the head (gray line), shoulder width (gray lines) and target position (gray point)).

vector being derived from a candidate k is represented as $P(y^k_{jt}|\tilde{y}^k_{jt},(\sigma^k_{jt})^2)$, whose probability distribution is assumed to be the normal distribution with mean $\tilde{y}^k_{jt}$ and variance $(\sigma^k_{jt})^2$, where $\tilde{y}^k_{jt}$ and $(\sigma^k_{jt})^2$ are the jth element of the predicted observation vector $\tilde{\boldsymbol{y}}^k_t = \boldsymbol{h}(\tilde{\boldsymbol{x}}_t)$ and its variance respectively. We evaluate the consistency of a candidate k by the mean of these probabilities C^k_t for all observations of the pixels belonging to the target.

The correct and false candidates at the 13th frame of Fig. 7 are shown in Fig. 6(a) and (b) respectively. Fig. 6(c) shows the transition of C^k_t for these candidates and shows that C^k_t of the correct candidate is larger than that of the false candidate. Therefore, we eliminate the candidate k if C^k_t is apparently smaller than the other candidates.

4.4. Initialization and Failure Recovery

At the initial frame or at the frames when the tracking failures have been detected, a moving object region in the image is searched for and the region having disparities similar to the mean disparity calculated inside this moving object region is determined to be the initial target region. The tracking failure is detected when the extracted target region suddenly becomes small. This happens when the target moves too fast.

The shoulder width and the target position in the image is computed from the initial target region and the multiple initial target state is calculated as described above.

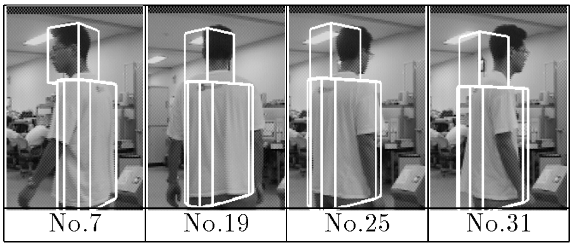


Figure 8. Tracking without shoulder width

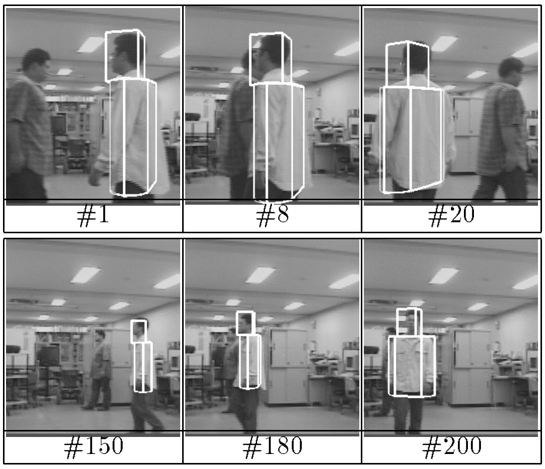


Figure 9. Tracking result in the existence of another object with similar flow vector or disparity

5. Experimental Result

Fig. 7 shows the result of tracking a person moving with 3-D motion including rotation upon the vertical axis. In this sequence, there are two candidates from the 1st frame to the 4th frame and from the 23th frame to the 29th frame and the rotation upon this axis can be correctly estimated.

Fig. 8 shows the result of tracking without using the shape of the target region, which means that we exclude the shoulder width W_s from Eq. (8). In this result, the rotation upon the vertical axis cannot be correctly estimated because the accurate estimation of angular velocity using optical flow is difficult and estimation error is accumulated.

In the sequence shown in Fig. 9, the target person and the other person walk closely in the opposite direction around the frame #8 and they walk apart in the same direction around the frame #180. In these cases, although the target person cannot be distinguished from the other person based on the disparity or the optical flow alone, the target person can be distinguished by integrating the optical flow and the depth thus the target with 3-D motion is correctly tracked.

The trajectory of the target person projected on the

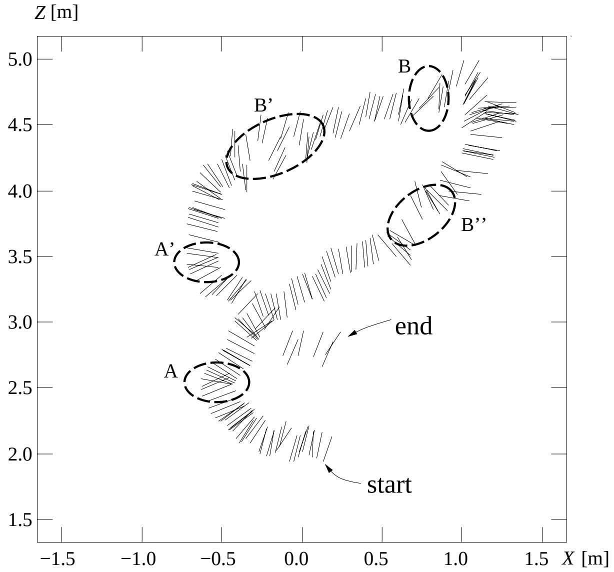


Figure 10. Top view of target trajectory (a line segment represents the shoulder direction. The target position is the midpoint of the segment.)

floor is shown in Fig. 10. There is a sudden change of the rotation angle at A and A', because a new state candidate is generated in this frame and the original one is eliminated in the subsequent frame. At B, B', and B" the estimated states are unstable because the background region is extracted as a part of the target region due to the error of the disparities caused by mistakes of correspondences.

6. Discussion and Conclusion

When the target stops walking or the target walks near the background objects, similar situations occur as shown in Fig. 9. The target can be correctly tracked in these cases. If the target walks near an object with the same velocity, the extracted target region includes both of them. Even in this case the target position is still tracked and the fact that two objects are tracked can be detected when the shoulder width becomes much larger than the predicted one. Although the estimation of the rotation angle becomes incorrect, it becomes correct when the two objects get separated.

In this paper, we proposed to integrate the shape and the position of the target region, optical flow, and depth so that they compensate for each other to track a person with 3-D motion. The target region was reliably extracted by integrating the optical flow and the depth in complex backgrounds. Furthermore multiple state candidates are maintained for robust tracking.

References

[1] S. Yamamoto, Y. Mae, Y. Shirai, and J. Miura, "Realtime Multiple Object Tracking Based on Optical Flows", *R&A*, Vol. 3, pp. 2328–2333, 1995.

[2] D. Coombs and C. Brown, "Real-time Smooth Pursuit Tracking for a Moving Binocular Robot", *CVPR*, pp. 23–28, 1992.

[3] D. P. Huttenlocher and J. J. Noh and W. J. Rucklidge, "Tracking Non-Rigid Objects in Complex Scenes", *4th ICCV*, pp. 93–101, 1993

[4] M. Etoh, Y. Shirai, "Segmentation and 2D Motion Estimation by Region Fragments", *4th ICCV*, pp. 192–198, 1993

[5] R. Okada, Y. Shirai, and J. Miura, "Object Tracking Based on Optical Flow and Depth", *Proc. of Int. Conf. on Multi-sensor Fusion and Integration for Intelligent Systems*, pp. 565–571, 1996.

[6] T. Yamane, Y. Shirai, and J. Miura, "Person Tracking by Integrating Optical Flow and Uniform Brightness Regions", *R&A*, pp. 3267–3272, 1998

[7] S. Araki, T. Matsuoka, H. Takemura, and N. Yokoya, "Real-time Tracking of Multiple Moving Objects in Moving Camera Image Sequence Using Robust Statistics", *14th ICPR*, Vol. 2, pp. 1433–1435, 1998.

[8] S.X. Ju, M.J. Black, and Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion", *Proc. of Int. Conf. on Face and Gesture Recognition*, pp. 461–465, 1996

[9] M. Yamamoto, A. Sato, S. Kawada, T. Kondo and Y. Osaki, "Incremental tracking of human actions from multiple views", *CVPR*, pp. 2–7, 1998.

[10] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps", *CVPR*, pp. 8–15, 1998

[11] G. Adiv, "Inherent Ambiguities in Recovering 3-D Motion and Structure from a Noisy Flow Field", *PAMI*, Vol. 11, No. 5, pp. 477–489, 1989.

[12] H. Inoue and T. Tachkawa and M. Inaba, "Robot Vision System with a Correlation Chip for Real-time tracking, Optical Flow and Depth Map Generation", *R&A*, pp. 1621–1626, 1992

[13] J. M. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects", *5th ICCV*, pp. 612–617, 1994