A Fast Stereo-Based Multi-Person Tracking using an Approximated Likelihood Map for Overlapping Silhouette Templates

Junji Satake Jun Miura
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
Email: {satake, jun}@cs.tut.ac.jp

***Abstract*—This paper describes a method of tracking multiple persons with occlusions using stereo. Many previous stereo-based systems track each person separately and do not explicitly handle such occlusions. We previously developed an accurate, stable tracking method using overlapping silhouette templates which considers how persons overlap in the image. However, because the method uses a particle filter, a lot of processing time is needed for estimating each particle's likelihood by comparing many templates with the image. In this paper, we propose a new method which can decrease the number of image comparison by using an approximated likelihood map based on kernel density estimation. Experimental results show that the proposed method is able to reduce the processing time greatly without dropping the tracking performance.**

***Index Terms*—person tracking; particle filter; stereo camera; silhouette templates; kernel density estimation;**

I. Introduction

Following a specific person is an important task for service robots. Visual person following in public spaces entails tracking of multiple persons by a moving camera. There have been a lot of works on person detection and tracking using various image features and classification methods [1]–[4]. Many of them, however, use a fixed camera. In the case of using a moving camera, foreground/background separation is an important problem.

This paper deals with detection and tracking of multiple persons which will be used on a mobile robot. Laser range finders are widely used for person detection and tracking by mobile robots [5], [6]. Image information such as color and texture is, however, sometimes necessary for person segmentation and/or identification. Omnidirectional cameras are also used [7], [8], but their limited resolutions are sometimes inappropriate for analyzing complex scenes. Stereo is also popular in moving object detection and tracking [9]–[11]. In these works, however, occlusions between people are not handled.

Ess et al. [12], [13] proposed to integrate various cues such as appearance-based object detection, depth estimation, visual odometry, and ground plane detection using a graphical model for pedestrian detection. Although their method exhibits a nice performance for complicated scenes, it is still costly to be used for controlling a real robot.

Some methods to track multiple objects by using particle filter are proposed [14]–[16]. In these methods, tracking of multiple interacting targets is realized by adding a probabilistic exclusion principle. They deal with the case where most part of each target is visible although small occlusions occur very often. In our case, since the images are taken from a camera on a mobile robot, complete occlusions frequently occur. To track overlapping persons stably, we proposed a method using overlapping silhouette templates [17], [18], which considers how persons overlap in the image.

The remainder of this paper is organized as follows. Section II describes our previous method [17], [18] using overlapping silhouette templates. In Section III, we propose an improvement of speeding up processing by using an approximated likelihood map based on kernel density estimation. Section IV presents experimental results, and finally Section V presents conclusions and future work.

II. Multi-person tracking using stereo

A. Person tracking based on distance information

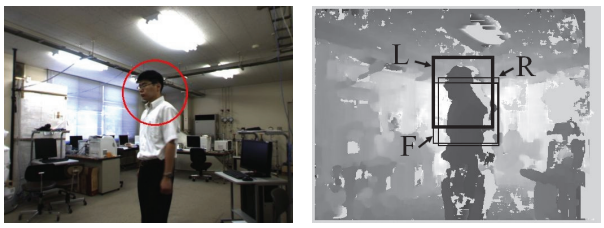
To track persons stably with a moving camera, we use *depth templates* [17], which are the templates for human upper bodies in depth images (see Fig. 1). We currently use three templates with different direction of body. We made the templates from the depth images where the target person was at 2 [m] away from the camera. A depth template is a binary template, the foreground and the background value are adjusted according to the status of tracks and input data.

For a person being tracked, his/her predicted scene position is available from the state variable (see Sec. II-B). We thus set the foreground depth of the template to the predicted depth of the head of the person.

Concerning the background depth, since it may change as the camera moves, we estimate it on-line. We make the depth histogram of the current input depth image and use the Kth percentile as the background depth (currently, $K = 90$).

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

Fig. 1. Depth templates



(a) Input image (b) Depth image

Fig. 2. Detection example using depth templates

For a depth template $T(x, y)$ of $H \times W$ pixels and the depth image $I_D(x, y)$, the dissimilarity d is calculated as follows.

$$d = \frac{1}{HW}\sqrt{\sum_p \sum_q [T(p,q) - I_D(x+p, y+q)]^2}\ .$$

We use the three templates simultaneously and take the one with the smallest dissimilarity as a matching result.

Figure 2 shows an example of detection using the depth templates. Three rectangles in the depth image are detection results with the three templates, and the one with the smallest dissimilarity is shown in bold line. Even when the direction of the body changed, it is possible to detect person stably by using the multiple templates.

B. Estimation of 3D position using particle filter

Figure 3 illustrates the coordinate systems attached to our mobile robot and stereo system. In the robot coordinate system, the person's position at time t is defined as (X_t, Y_t, Z_t). The state variable $\boldsymbol{x}_t$ is defined as

$$\boldsymbol{x}_t = [\ X_t \quad Y_t \quad Z_t \quad \dot{X}_t \quad \dot{Y}_t\ ]^T,$$

where $\dot{X}_t$ and $\dot{Y}_t$ denote velocities in the horizontal plane. We assume the vertical position is constant. The state equation is given by

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t \boldsymbol{x}_t + \boldsymbol{w}_t,$$

where matrix $\boldsymbol{F}_t$ is set to represent a linear uniform motion.

We estimate the 3D position of each person by using particle filter. The likelihood L of each particle is calculated based on the dissimilarity d described in Sec. II-A.

$$L = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d^2}{2\sigma^2}\right),$$

where the standard deviation σ is set up experientially. Person's position is calculated by the weighted average of particles. We use an OpenCV implementation of particle filter.

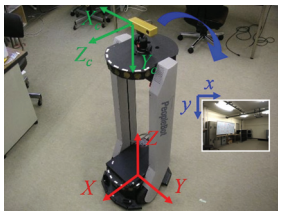


Fig. 3. Definition of coordinate systems

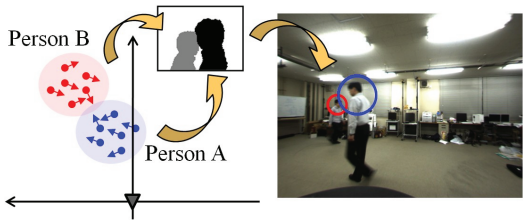


Fig. 4. Procedure of tracking using an overlapping silhouette template

C. Multi-person tracking using overlapping silhouette templates

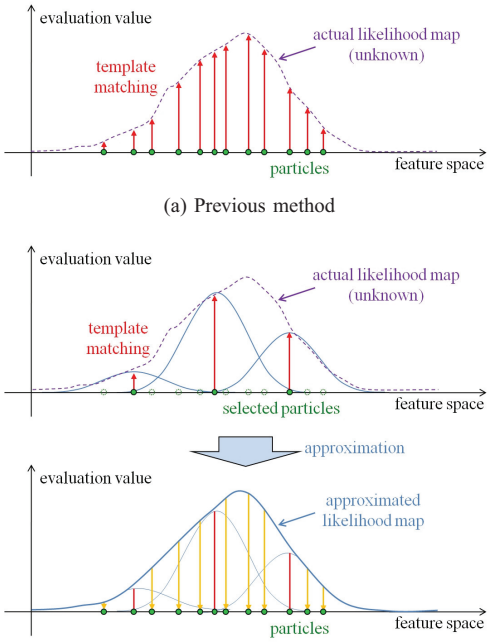
To track overlapping persons stably, we make *overlapping silhouette templates* [18] which consider the overlap of persons in the image. Each person which is isolated from other persons is independently tracked by using N particles. When two persons, say A and B, approach each other and an overlap occurs, a new combined state vector $\boldsymbol{x}_t^{AB}$ for both persons is made from the respective ones $\boldsymbol{x}_t^A$ and $\boldsymbol{x}_t^B$.

$$\boldsymbol{x}_t^{AB} = \begin{bmatrix} \boldsymbol{x}_t^A \\ \boldsymbol{x}_t^B \end{bmatrix}.$$

Since the number of particles of each person is N, the total number of combined state is $N \times N$. To reduce the calculation cost of template matching for the combinations, we use only particles with large likelihood values among each person's particles. We set the number of each person's particles to $N = 100$, and the number of combined particles to $N^{AB} = 25 \times 25$. The initial likelihood of each combined particle is set as product of their likelihood values $L^{AB} = L^A L^B$. The state equation is as follows.

$$\boldsymbol{x}_{t+1}^{AB} = \begin{bmatrix} \boldsymbol{F}_t & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}_t \end{bmatrix} \boldsymbol{x}_t^{AB} + \begin{bmatrix} \boldsymbol{w}_t \\ \boldsymbol{w}_t \end{bmatrix}.$$

Figure 4 shows the procedure of tracking using an overlapping silhouette template. The template of each combined particle is made in consideration of the states of two persons. The relative position in the image coordinates is calculated, and the individual template of the person near the camera is overwritten on the template of the far person. The values for the foregrounds and the background are set similarly to the case of one foreground case in Sec. II-A. Only one template among three (see Fig. 1), corresponding to the estimated movement direction is used for reducing the number of combined templates. The overlapping silhouette template is matched to the depth image, and the likelihood L^{AB} is calculated.



(a) Previous method

(b) Proposed method based on kernel density estimation

Fig. 5. Estimation of each particle's likelihood

When the distance between persons A and B exceeds a threshold, the state variable $\boldsymbol{x}_t^{AB}$ is separated into $\boldsymbol{x}_t^A$ and $\boldsymbol{x}_t^B$ using the N particles with the largest likelihood values among N^{AB} particles.

Our previous work realized an accurate and robust tracking using overlapping silhouette templates [18]. However, the processing speed is about 300 [ms/frame] for estimating every particle's likelihood by comparing many templates with the image. In Section III, we propose a new method which can decrease the number of image comparison by using an approximated likelihood map based on kernel density estimation.

III. Approximation of Likelihood Map Based on Kernel Density Estimation

Figure 5 shows the outline of our proposed mehod. The previous method needs a lot of processing time because the combined template is compared with the image for all particles. Therefore, we propose a new method which selects a part of particles, compares their templates with the image, and makes an approximated likelihood map by using the comparison results based on kernel density estimation. The other particles' likelihoods are then estimated by using the approximated likelihood map.

As shown in Figure 6, the silhouette template is made based on four factors: two persons' depths X^A and X^B, horizontal distance $Y^A - Y^B$, and height difference $Z^A - Z^B$. In this paper, we consider the distance $Y^A - Y^B$ and assume that depth X^A, X^B, and height difference $Z^A - Z^B$ are constant under a short occlusion.

The position $(\bar{X}^N, \bar{Y}^N, \bar{Z}^N)$ and $(\bar{X}^F, \bar{Y}^F, \bar{Z}^F)$ denote the near and far person's position which are calculated by the

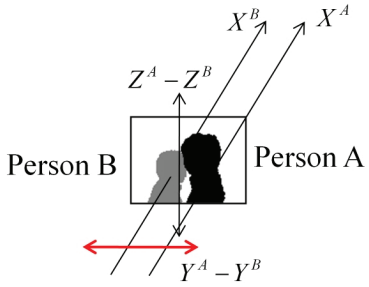


Fig. 6. Four factors of silhouette template

weighted average of particles. As shown in Figure 7(a), we examine the change of the evaluation value when horizontal positions Y^N and Y^F change; this is equivalent to examining the change of the distance $Y^A - Y^B$ and the template's horizontal position. The search range is the near person's position ± 300 [mm] and the far person's position ± 900 [mm] as follows:

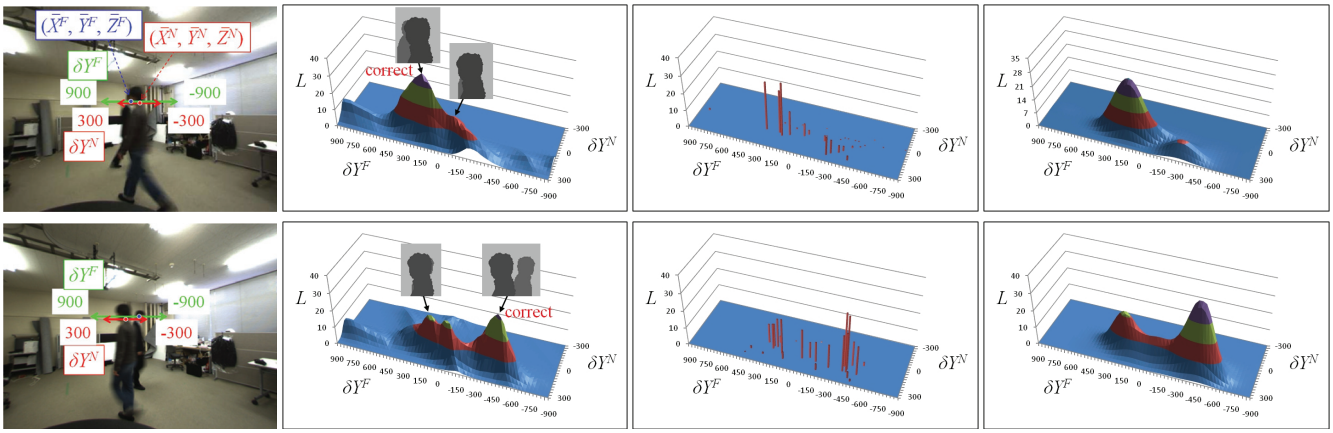
$$\begin{bmatrix} X^N \\ Y^N \\ Z^N \end{bmatrix} = \begin{bmatrix} \bar{X}^N \\ \bar{Y}^N + \delta Y^N \\ \bar{Z}^N \end{bmatrix}, \begin{bmatrix} X^F \\ Y^F \\ Z^F \end{bmatrix} = \begin{bmatrix} \bar{X}^F \\ \bar{Y}^N + \delta Y^F \\ \bar{Z}^F \end{bmatrix},$$

where $\delta Y^N = [-300, 300]$, $\delta Y^F = [-900, 900]$. To detect stably the occluded far person who appears from the side of the near person, the far person's search range is set to both sides of $\bar{Y}^N$. Figure 7 (b) shows the evaluation value of template matching. Each person's position is sampled at intervals of 30 [mm]. This likelihood map is considered a complete reference table that indicates the evaluation value for each combination of person positions. It is confirmed that the evaluation value at the correct position is the highest and falls as the position shifts. However, it needs a lot of processing time to make this reference table because 1281 $(= 21 \times 61)$ templates are matched.

To reduce the processing time, we approximate a likelihood map by a less number of template matching. At first, A certain number of particles are selected with large likelihood values at the previous frame and their templates are matched with the image to obtain the evaluation values (Fig. 7(c)). We set the number of selected particles to 30 or 50 in the experiment. The near and the far person's positions and the likelihood of particle i are denoted as (X_i^N, Y_i^N, Z_i^N), (X_i^F, Y_i^F, Z_i^F) and L_i, respectively. Using a kernel density estimation with a Gaussian distribution, an approximated likelihood map is made as follows.

$$L\left(\delta Y^N, \delta Y^F\right) = \sum_i L_i \exp\left\{ -\frac{(\delta Y^N - \mu_i^N)^2}{2\sigma^{N\,2}} - \frac{(\delta Y^F - \mu_i^F)^2}{2\sigma^{F\,2}} \right\},$$

where the averages of distributions are $\mu_i^N = Y_i^N - \bar{Y}^N$, $\mu_i^F = Y_i^F - \bar{Y}^N$. The standard deviations σ^N and σ^F are set empirically based on the size of actual person and the overlap of two persons ($\sigma^N = 40$ [mm], $\sigma^F = 100$ [mm]). Moreover, the coefficient part of distribution is omitted because only



(a) Input image and search area (b) Reference table and evaluation value (c) Evaluation value of 30 particles (d) Approximated likelihood map

Fig. 7. Approximation of likelihood map baed on kernel density estimation

the ratio is necessary for the calculation of each particle's likelihood.

Approximated likelihood maps are shown in Figure 7(d). Similar shapes to the reference tables of Figure 7(b) are obtained. Each particle's likelihood is estimated by using on this approximated likelihood map instead of matching between the template and the input image. We realize a faster tracking by using this approximated likelihood map which is made by a less number of template matching.

IV. Experimental Result

We have implemented the proposed method with a Bumblebee2 stereo camera (by Point Grey Research) and a note PC (Core2Duo, 3.06 [GHz]). The processed image size is 512×384. Because the method does not consider the change in depth and height of each person, we used test data sets in which persons roughly move in parallel to the image plane.

Figure 8 is a result of tracking using off-line images taken at 10 [fps]. Each pair of circles with white edges shows a tracking result by using the combined state variable. Blue lines show search area of likelihood map (see Fig. 7). Even when one person is occluded by the other, each person's position can be estimated stably.

Table I shows the comparison of tracking results for 140 occlusion cases (14 data sets were tested ten times respectively). The number of total frames is $598 \times 10 = 5980$. Each test data set is the off-line images in which two person approach each other, intersect, and then depart. Each person's position (ground truth) in each frame was given manually. We counted success cases where every person was tracked correctly at all frames and calculated the success rate. The averages of the 2D positional error and the time of processing a frame were calculated for only the success data sets.

(a) Our previous work [18] realized a stable tracking. However, it takes a lot of processing time to match the template with each $625 (= 25^2)$ particle.

(b) When the reference table (see Fig. 7(b)) is used, $1281 (= 21 \times 61)$ templates are matched. The success rate of tracking is almost the same as the method (a). The positional error slightly grows in the vertical direction because we consider the height difference is constant. For example, the far person's position shifts downward in Figure 8 #119.

(c) When the approximated likelihood map is used, the processing time is reduced greatly while keeping a similar level of success rate. In the case of 50 particles selected as reference points, the success rate is almost the same as the method (a) and (b). When 30 particles were used, the success rate fell a little. It is necessary to examine the particle selection strategy to make the likelihood map more effectively.

Figure 9 is an example of the tracking result when the persons' depths change. The tracking fails because we considered only the change of horizontal positions and assumed that the depths and the heights are constant. In the future, we should expand dimensions for depth and height, and realize a stable tracking in a more general situation.

V. Conclusion

This paper has described a method of tracking multiple persons by using overlapping silhouette templates. We proposed a method which can decrease the number of image comparison by using an approximated likelihood map based on kernel density estimation. We have reduced the number of template matching to about 1/10 without degrading the tracking performance, and the total processing time to about 100 [ms/frame]. This processing speed is practical for a person following robot. However, we considered only the change of horizontal positions and assumed that the depths and the heights are constant. In the future, we should expand dimensions for depth and height, and realize a stable tracking in a more general situation. To make the approximated likelihood map more effectively, it is necessary to examine the particle selection strategy. Moreover, it is necessary to deal with the overlapping of three persons or more.

Fig. 8. Experimental result of tracking

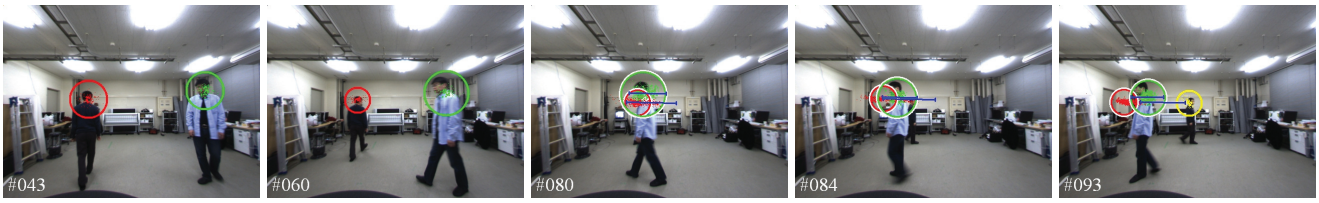


Fig. 9. Example of tracking failure because of change in depth

TABLE I
COMPARISON OF TRACKING RESULTS

tracking method	template matching	processing time (matching)	processing time (total)	positional error	success rate
(a) previous method [18]	625 [times/frame]	247 [ms/frame]	328 [ms/frame]	7.18 [pixel]	97.8 [%]
(b) using reference table	1281 [times/frame]	475 [ms/frame]	555 [ms/frame]	8.08 [pixel]	97.1 [%]
(c) proposed method	50 [times/frame]	20 [ms/frame]	101 [ms/frame]	8.27 [pixel]	97.1 [%]
	30 [times/frame]	11 [ms/frame]	93 [ms/frame]	8.36 [pixel]	91.4 [%]

ACKNOWLEDGMENT

A part of this research is supported by Kurata Memorial Hitachi Science and Technology Foundation, NEDO Intelligent RT Software Project, and JSPS KAKENHI 23700203.

REFERENCES

[1] P. Viola, M.J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," Int. J. Computer Vision, vol. 63, no. 2, pp. 153–161, 2005.
[2] N. Dalal and B. Briggs, "Histograms of oriented gradients for human detection," IEEE Conf. Computer Vision and Patttern Recognition, pp. 886–893, 2005.
[3] B. Han, S. W. Joo, and L. S. Davis, "Probabilistic fusion tracking using mixture kernel-based Bayesian filtering," 11th Int. Conf. Computer Vision, 2007.
[4] S. Munder, C. Schnorr, and D. M. Gavrila, "Pedestrian detection and tracking using a mixture of view-based shape-texture models," IEEE Trans. ITS, vol. 9, no. 2, pp. 333–343, 2008.
[5] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers, "People tracking with a mobile robot using sample-based joint probabilistic data association filters," Int. J. Robotics Research, vol. 22, no. 2, pp. 99–116, 2003.
[6] N. Bellotto and H. Hu. "Multisensor data fusion for joint people tracking and identification with a service robot," IEEE Int. Conf. Robotics and Biomimetics, pp. 1494–1499, 2007.
[7] H. Koyasu, J. Miura, and Y. Shirai, "Realtime omnidirectional stereo for obstacle detection and tracking in dynamic environments," IEEE/RSJ Int. Conf. Intelligent Robots and Sysetms, pp. 31–36, 2001.
[8] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, "People tracking and following with mobile robot using omnidirectional camera and a laser," IEEE Int. Conf. Robotics and Automation, pp. 557–562, 2006.
[9] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continuous detection," 7th Int. Conf. Computer Vision, 1999.
[10] A. Howard, L. H. Matthies, A. Huertas, M. Bajracharya, and A. Rankin, "Detecting pedestrians with stereo vision: safe operation of autonomous ground vehicles in dynamic environments," Int. Symp. Robotics Research, 2007.
[11] D. Calisi, L. Iocchi, and R. Leone, "Person following through appearance models and stereo vision using a mobile robot," VISAPP Workshop on Robot Vision, pp. 46–56, 2007.
[12] A. Ess, B. Leibe, and L. V. Cool, "Depth and appearance for mobile scene analysis," 11th Int. Conf. Computer Vision, 2007.
[13] A. Ess, B. Leibe, K. Schindler, and L. V. Cool, "A mobile vision system for robust multi-person tracking," IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[14] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," 7th Int. Conf. Computer Vision, pp. 572–578, 1999.
[15] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," 8th European Conf. Computer Vision, pp. 279–290, 2004.
[16] D. Tweed and A. Calway, "Tracking many objects using subordinated Condensation," British Machine Vision Conf., pp. 283–292, 2002.
[17] J. Satake and J. Miura, "Robust stereo-based person detection and tracking for a person following robot," ICRA Workshop on People Detection and Tracking, 2009.
[18] J. Satake and J. Miura, "Stereo-Based Multi-Person Tracking using Overlapping Silhouette Templates," 20th Int. Conf. on Pattern Recognition, pp. 4304–4307, 2010.