2010 International Conference on Pattern Recognition

Stereo-Based Multi-Person Tracking using 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 stereo-based person tracking method for a person following robot. Many previous works on person tracking use laser range finders which can provide very accurate range measurements. Stereo-based systems have also been popular, but most of them are not used for controlling a real robot. We previously developed a tracking method which uses depth templates of person shape applied to a dense depth image. The method, however, sometimes failed when complex occlusions occurred. In this paper, we propose an accurate, stable tracking method using overlapping silhouette templates which consider how persons overlap in the image. Experimental results show the effectiveness of the proposed method.**

***Keywords*-person tracking; mobile robot; particle filter; stereo camera; silhouette templates;**

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 for 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. Khan et al. [15] deal with tracking of multiple interacting insects. Since the targets' movements are restricted on a 2D plane and seen from above, complex occlusions hardly occur. Tweed and Calway [16] realize a tracking of many flying birds by setting links between the overlapping targets. They deal with the case where most part of each bird is visible although small occlusions between wings occur very often. In our case, since the images are taken from a camera on a mobile robot, complex and complete occlusions frequently occur. We, therefore, use distance information for discriminating and tracking multiple persons. Especially, we propose to use an overlapping silhouette template for accurately and stably tracking multiple interacting persons.

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 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.

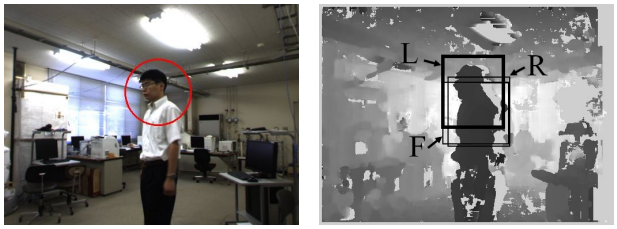
Concerning the background depth, since it may change as the camera moves, we estimate it on-line. We make the depth histogram of the current input depth image and use the Kth percentile as the background depth (currently, $K=90$).

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.

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

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1046

Figure 1. Depth templates



(a) Input image (b) Depth image

Figure 2. Detection example using depth templates

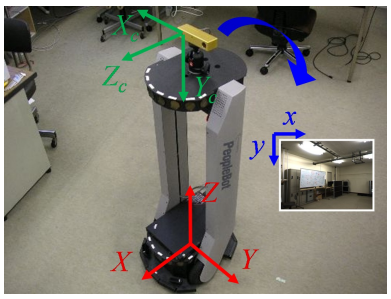


Figure 3. Definition of coordinate systems

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 highest evaluation value is shown in bold line. Even when the direction of the body changed, it is possible to detect person stably by using 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

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

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

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

Person's position is calculated by the weighted average of particles. We use an OpenCV implementation of Particle Filter.

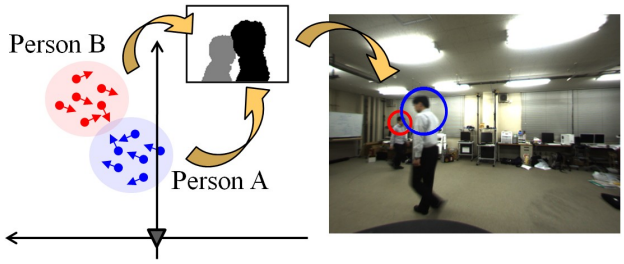


Figure 4. Procedure of tracking using an overlapping silhouette template

C. Multi-person tracking using overlapping silhouette templates

To track the close persons with stability, we make *overlapping silhouette templates* which consider overlap of persons on 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 $\boldsymbol{x}_t^B$.

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

When the number of particles of each person is N, the total number of combined state is $N \times N$. Because the calculation cost for all the combinations is too large, 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 $N^{AB} = 25 \times 25$ in the experiment. An initial likelihood of the combined particle is set as $L^{AB} = L^A L^B$. 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 that 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

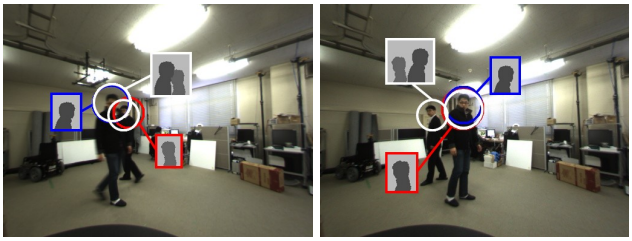


Figure 5. Comparison of tracking results with overlapping silhouette templates (drawn by white frame) and with individual templates (drawn by blue or red frame)

combined templates. The overlapping silhouette template is matched to the depth image, and the likelihood L^{AB} is calculated.

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.

Figure 5 shows the comparison of tracking results with overlapping silhouette templates and with individual templates. In the tracking with individual templates, the estimated position of the far person (drawn by red) is not accurate because a part of his silhouette had hidden by the near person (Fig. 5 left). In addition, when one person was fully occluded, the wrong one was tracked (Fig. 5 right). On the other hand, the persons were tracked accurately with overlapping silhouette templates because these consider the overlap of persons.

III. Experimental Result

We have implemented the proposed method on a PeopleBot (by Mobile Robots) with a Bumblebee2 stereo camera (by Point Grey Research) for the experiments (see Fig. 3). A note PC (Core2Duo, 3.06 [GHz]) performs all processes including stereo calculation, person detection and tracking, and robot motion control. The processed image size is 512×384.

Figure 6 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. Even when the directions of movement of a person changed while occluded by another person (for instance, person B at #180 and person A at #318), he was tracked correctly. This is most probably because a sufficient variety of hypotheses were generated and evaluated in consideration of the overlap of persons.

Table I shows the comparison of tracking results for 60 occlusion cases (12 case were tested five times respectively). Each test data set is the off-line images in which two person approach each other, intersect, and then part. These include various changes in direction and speed of movement. 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. Using overlapping silhouette templates makes the tracking far more accurate and stable than the individual template-based tracking.

Table I
Comparison of Tracking Results
Concerning the Number of Particles

(a) using only individual templates

number of particles	success rate	positional error	processing time
$N = 100$	73.3 [%]	9.56 [pixel]	248 [ms]
$N = 200$	75.0 [%]	9.95 [pixel]	383 [ms]

(b) using overlapping silhouette templates

number of particles	success rate	positional error	processing time
$N^{AB} = 15 \times 15$	86.7 [%]	7.09 [pixel]	226 [ms]
$N^{AB} = 20 \times 20$	90.0 [%]	6.74 [pixel]	314 [ms]
$N^{AB} = 25 \times 25$	93.3 [%]	6.64 [pixel]	436 [ms]
$N^{AB} = 30 \times 30$	90.0 [%]	6.53 [pixel]	568 [ms]
$N^{AB} = 100 \times 100$	91.7 [%]	6.43 [pixel]	4993 [ms]

Figure 7 shows a result on control of mobile robot following a specific person. The robot moved toward person A who was detected first. Even when person B passed between the robot and person A, the target person was correctly tracked.

IV. Conclusion

This paper has described a method of tracking multiple persons by using distance information obtained by stereo. We realized an accurate and stable tracking using overlapping silhouette templates. In future work, we should speed up processing in order to deal person's quick movement. We will also deal with more complex scenes where a new person appears from the behind of the tracked persons by enhancing the overlapping silhouette templates.

Acknowledgment

A part of this research is supported by Kurata Memorial Hitachi Science and Technology Foundation, Tateisi Science and Technology Foundation, and NEDO (New Energy and Industrial Technology Development Organization, Japan) Intelligent RT Software Project.

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.

Figure 6. Experimental result of tracking using off-line images

Figure 7. Result on a person following control

[3] B. Han, S. W. Joo, and L. S. Davis, "Probabilistic fusion tracking using mixture kernel-based Bayesian filtering," 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," 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," 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," 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," 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.