21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Stereo-Based Tracking of Multiple Overlapping Persons

Junji Satake Jun Miura
Department of Computer Science and Engineering
Toyohashi University of Technology

Abstract

This paper describes a method of tracking multiple persons with occlusions using stereo. We previously developed an accurate and stable tracking method using overlapping silhouette templates which considers how persons overlap in the image. It realized a fast tracking by using an approximated likelihood map based on kernel density estimation. The method, however, treated only two overlapping persons. In this paper, we propose an improvement of approximation applicable to the case where three or more persons overlap. Experimental results show that the proposed method can track persons stably even when three persons overlap in the image.

1. 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, 2, 3]. Many of them, however, use a fixed camera. In the case of using a moving camera, foreground/background separation is an important problem.

HOG [2] is currently one of the most widely used features for visual people detection [4, 5]. Moreover, the person detection methods which combined HOG and the distance information acquired using an RGB-D camera such as Microsoft Kinect sensor are also proposed [6, 7]. In these works, however, occlusions between people are not explicitly handled. In our case, since the images are taken from a camera on a mobile robot, complete occlusions frequently occur.

Ess et al. [8] 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.

Figure 1. Depth templates

To track overlapping persons, we proposed a method using overlapping silhouette templates [9, 10], which considers how persons overlap in the image. However, it treated only two overlapping persons. In this paper, we extend the method to more general parameterization and propose an improvement of approximation applicable to three or more overlapping persons.

2 Multi-person tracking using stereo

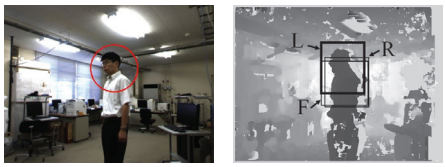
2.1 Person tracking based on distance information

To track persons with a moving camera, we use *depth templates* [9], which are the templates for human upper bodies in depth images (see Fig. 1). We made the templates manually 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 scene position is predicted using a particle filter. We thus set the foreground depth of the template to the predicted depth of the head of the person. Then we calculate the dissimilarity d between a depth template and the depth image using an SSD (sum of squared distances) criterion.

To detect a person in various orientation, we use the three templates simultaneously and take the one with the smallest dissimilarity as a matching result. An example of detection using the depth templates is shown in Figure 2.

978-4-9906441-1-6 ©2012 IAPR



(a) Input image (b) Depth image

Figure 2. Detection example using depth templates

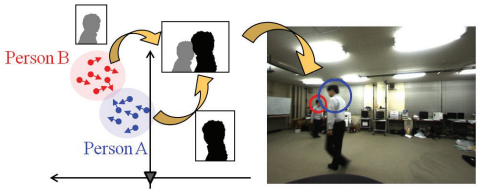


Figure 3. Procedure of tracking using an overlapping silhouette template

2.2 Estimation of 3D position using particle filter

In the robot coordinate system, the person's position at time t is defined as (X_t, Y_t, Z_t). The state variable is defined as $\boldsymbol{x}_t = [\ X_t\ Y_t\ Z_t\ \dot{X}_t\ \dot{Y}_t\ ]^T$, 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 $\boldsymbol{F}_t$ represents a linear motion with constant velocity.

We estimate the 3D position of each person by using particle filter. Likelihood L of each particle is calculated based on the dissimilarity d described in Sec. 2.1.

2.3 Multi-person tracking using overlapping silhouette templates

2.3.1 Two overlapping persons

To track overlapping persons, we make *overlapping silhouette templates* [9] which consider the overlap of persons in the image.

Each person who is isolated from others is independently tracked. When two persons, say A and B, approach each other and an overlap occurs, a new combined state vector $\boldsymbol{x}_t^{AB} = \left[\ (\boldsymbol{x}_t^A)^T\ (\boldsymbol{x}_t^B)^T\ \right]^T$ for both persons is made from the respective ones, $\boldsymbol{x}_t^A$ and $\boldsymbol{x}_t^B$. To reduce the calculation cost of template matching for the combinations, we use only particles with large likelihood values among each person's particles. Let $N_{(n)}$ be the number of particles for a track of n persons. We set $N_{(1)}$ and $N_{(2)}$ to 100 and 20×20, respectively.

Figure 3 illustrates the procedure of making combined states and overlapping silhouette templates. The template of each combined particle is made in consideration of the relative position in the image coordinates. Each overlapping silhouette template is matched to the depth image, and the likelihood L^{AB} is calculated.

When the distance between persons A and B becomes large and their templates do not overlap, the state variable $\boldsymbol{x}_t^{AB}$ is separated into $\boldsymbol{x}_t^A$ and $\boldsymbol{x}_t^B$ using the $N_{(1)}$ particles with the largest likelihood values among $N_{(2)}$ particles.

2.3.2 Multiple overlapping persons

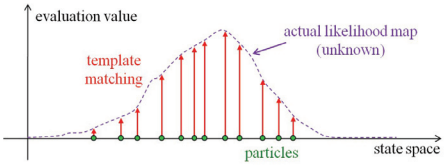
When three or more persons approach and merge in the image, a new combined state vector is made as in the case of two persons. Let $\{p_i | i = 1 \ldots m\}$ be a set of already-combined person groups which are isolated in the previous frame and merge in the current frame. Let m_j be the number of persons in the jth group ($\sum_j m_j = n$); note that in the case of an isolated person, the number is one. Then we pick up $N_{(n)}{}^{\frac{m_i}{n}}$ best particles (i.e., ones with largest likelihood values) from the jth group and make every combinations of particles from all groups to generate $N_{(n)}$ particles for simultaneously tracking all persons in the groups. For each combined particle, the corresponding overlapping silhouette template is created according to the persons' positions in the state vector.

For the overlap of three persons, for example, there are the following two cases. One is the case where three isolated persons overlap simultaneously. In this case, $N_{(3)}$ combined particles are generated from $N_{(3)}{}^{\frac{1}{3}}$ particles from each person. The other corresponds to the case where one isolated person merges to a group of two persons. In this case, $N_{(3)}{}^{\frac{1}{3}}$ and $N_{(3)}{}^{\frac{2}{3}}$ particles are picked up from the isolated person and from the group, respectively. In the experiments described below, we set $N_{(3)}$ to 729.

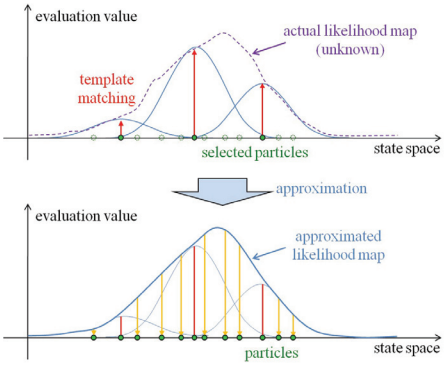
When n overlapping persons separate to a set of groups $\{p_i | i = 1 \ldots m\}$, the state variable of group j is extracted from $N_{(m_j)}$ particles with largest likelihood values among $N_{(n)}$ particles.

3 Approximation of likelihood map based on kernel density estimation

The method described in Section 2.3 needs a lot of processing time because the templates are compared with the image for all particles (Fig. 4(a)). To reduce the processing time, we proposed a method [10] which approximates a likelihood map by a less number of template matching operations based on kernel density esti-

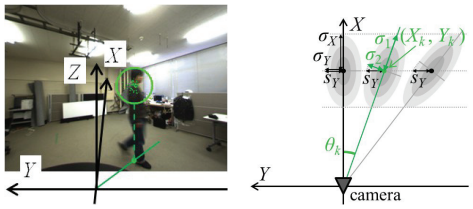


(a) Template matching for each particle

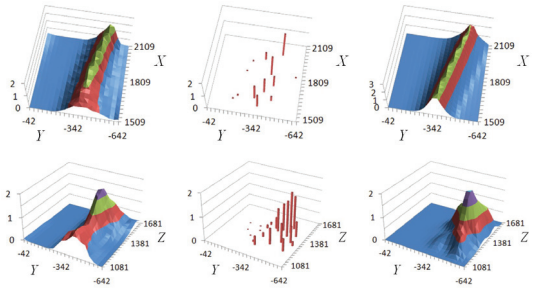


(b) Fast likelihood estimation
based on kernel density estimation

Figure 4. Estimation of each particle's likelihood



(a) Likelihood distribution considering projection



(b) Actual likelihood map (c) Evaluation value of 50 particles (d) Approximated likelihood map

Figure 5. Approximation of likelihood map based on kernel density estimation

mation (Fig. 4(b)). However, the method treated only two overlapping persons. In this paper, we propose an approximation method which applicable to three or more overlapping persons.

3.1 Approximation in the case of one person

First, we describe the case of one person which is isolated from other persons. The person is tracked using $N_{(1)}$ particles. 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 at the current frame. The likelihood of selected particle k is denoted as L_k. Using a kernel density estimation with a Gaussian distribution in X-Y-Z space, an approximated likelihood map is made as follows:

$$L(\boldsymbol{x}) = \sum_k L_k |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\Delta_k(\boldsymbol{x})^T \Sigma_k^{-1} \Delta_k(\boldsymbol{x})\right\},$$

$$\Delta_k(\boldsymbol{x}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} (\boldsymbol{x} - \boldsymbol{x}_k) = \begin{bmatrix} X - X_k \\ Y - Y_k \\ Z - Z_k \end{bmatrix},$$

$$\Sigma_k = \boldsymbol{R}_Z(\theta_k) \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \boldsymbol{R}_Z(\theta_k)^T,$$

where the distribution is rotated about Z-axis in consideration of the perspective projection, so that its principal axis is aligned to the viewing direction (Fig. 5(a)). The rotation angle θ_k is calculated by $\theta_k = \tan^{-1}\frac{Y_k}{X_k}$. The standard deviations σ_X, σ_Y, and σ_Z along X, Y, and Z-axis when the person is right in front of the camera ($\theta_k = 0$) are set empirically based on the size of actual person. When the positional difference between the depth image and the template becomes larger, likelihood L becomes smaller. This is independent of the position in the image. Since the perspective projection is assumed, horizontal displacement Δu in the image is proportional to ΔY in 3D space. Therefore, the standard deviation s_Y of the section which cuts the center of distribution in Y direction is constant (Fig. 5(a)). The deviations σ_1, σ_2 and σ_3 along the principal axes of the rotated distribution are calculated as follows using θ_k:

$$\sigma_1 = \frac{\sigma_X}{\cos\theta_k}, \quad \sigma_2 = s_Y \cos\theta_k, \quad \sigma_3 = \sigma_Z.$$

Figure 5(d) shows an example of approximated likelihood map based on 50 selected particles (Fig. 5(c)). Similar shape to actual likelihood map (Fig. 5(b)) which indicates the evaluation value for each position is obtained. The number of particles used for the approximation was experientially set so that a sufficiently approximated map is obtained.

3.2 Approximation in the case of n persons

In the case of n overlapping persons, the dimension of feature space is extended, and a likelihood map is approximated in $3n$ dimensional space. We assume that the influence of a gap in template matching is independent at each person, and approximate the likelihood map as follows:

$$L_{(n)}(\boldsymbol{x}^n) = \varsigma(\boldsymbol{x}^n) \times \sum_k L_k \left|\Sigma_k^n\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\Delta_k^n(\boldsymbol{x}^n)^T(\Sigma_k^n)^{-1}\Delta_k^n(\boldsymbol{x}^n)\right\},$$

where

$$\Delta_k^n(\boldsymbol{x}^n) = \left[\begin{bmatrix} X^{p_1}-X_k^{p_1} \\ Y^{p_1}-Y_k^{p_1} \\ Z^{p_1}-Z_k^{p_1}\end{bmatrix}^T \cdots \begin{bmatrix} X^{p_n}-X_k^{p_n} \\ Y^{p_n}-Y_k^{p_n} \\ Z^{p_n}-Z_k^{p_n}\end{bmatrix}^T\right]^T,$$

$$\Sigma_k^n = \begin{bmatrix} \Sigma_k & & 0 \\ & \ddots & \\ 0 & & \Sigma_k \end{bmatrix}.$$

We multiply a likelihood map by the coefficient $\varsigma(\boldsymbol{x}^n)$ as an exclusion principle. It is because multiple persons do not occupy the same position in 3D space, although the motion of each person is independently predicted by using particle filter. The coefficient $\varsigma(\boldsymbol{x}^n)$ is calculated as a product of sigmoid functions of distance δ between two persons:

$$\varsigma(\boldsymbol{x}^n) = \prod_{i\neq j} \frac{1}{1+e^{-a\{\delta(i,j)-b\}}} \quad (i,j \in \{1,\cdots,n\}),$$

$$\delta(i,j) = \sqrt{(X^{p_i}-X^{p_j})^2+(Y^{p_i}-Y^{p_j})^2}.$$

Figure 6 is the sigmoid function ($a = 0.02$, $b = 500$) used in the experiment. These parameters were set based on the actual person's size.

Figure 7 shows examples of approximation in the cases of two and three overlapping persons[1]. Even when multiple persons overlap, similar shape to actual likelihood map (Fig. 7(b)) is obtained (Fig. 7(d)). Although one mode is missing in the second approximated likelihood map, it does not influence tracking because particle does not exist in that position.

4 Experimental result

We have implemented the proposed method with a Bumblebee2 stereo camera (by Point Grey Research) and a note PC (Core i7, 2.7 [GHz]). The processed image size is 512×384. In this paper, we tested the method for the cases where at most three persons overlap.

[1] The figure shows the change of the likelihood values for parameters Y^A and Y^B. For the other parameters, the weighted average values over particles are used.

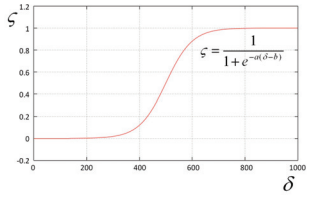


Figure 6. Sigmoid function used for exclusion principle

Figure 8 shows a result of tracking using off-line images taken at 10 [fps]. Each circle with white or red edge shows a tracking result by using the combined state variable for two or three persons, respectively. Even when three persons overlap in the image, each person's position can be estimated accurately.

Table 1 shows the comparison of tracking results for 360 occlusion cases (total 15830 frames). 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 error of horizontal position and the time of processing a frame were calculated for only the success cases. In the tracking with individual templates, the occluded person is often missed because the overlap is not considered. Although a stable tracking is realized by using overlapping templates, it takes a lot of processing time to match the template with each particle. When the approximated likelihood map is used, the processing time is reduced greatly while keeping a similar level of success rate.

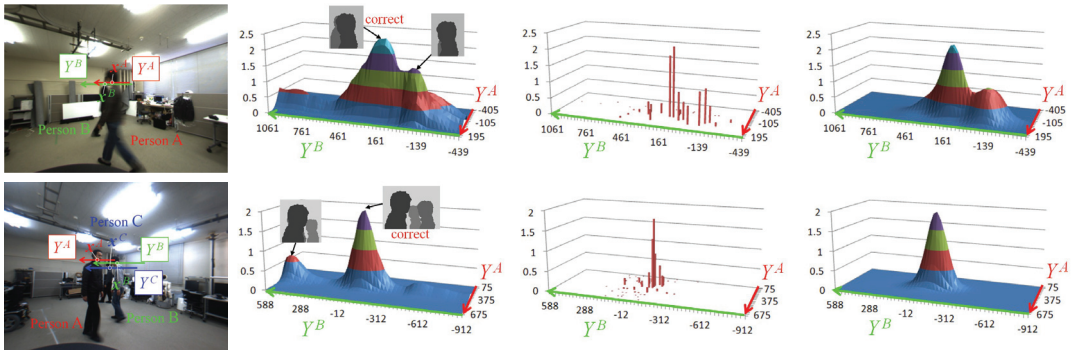
5 Conclusion

This paper has described a method of tracking multiple overlapping persons. We extended our previous method, which considers only overlaps of two persons, to be applicable to general, three or more persons cases. We tested the method using real image sequences where at most three persons overlap and showed it can achieve about 10% increase of success rate in difficult, three-persons cases, compared with the tracking using individual templates.

Some of future work are to test the method for more complicated cases with more overlapping persons and to apply the method to RGB-D cameras which usually provide more reliable depth data.

Acknowledgment

A part of this research is supported by JSPS KAKENHI 23700203 and NEDO Intelligent RT Software Project.



(a) Input image (b) Actual likelihood map (c) Selected particles (d) Approximated likelihood map

Figure 7. Examples of approximation in the case of two and three overlapping persons

Figure 8. Experimental result of tracking

Table 1. Comparison of tracking results

(a) two overlapping persons (240 occlusion cases)

tracking method	template matching	processing time	positional error	success rate
individual templates	200 [times/frame]	146.6 [ms/frame]	6.24 [pixel]	80.8 [%]
each overlapping template	400 [times/frame]	195.0 [ms/frame]	5.30 [pixel]	95.0 [%]
using approximation (proposed)	100 [times/frame]	65.9 [ms/frame]	5.40 [pixel]	94.6 [%]

(b) three overlapping persons (120 occlusion cases)

tracking method	template matching	processing time	positional error	success rate
individual templates	300 [times/frame]	202.5 [ms/frame]	4.75 [pixel]	73.7 [%]
each overlapping template	729 [times/frame]	363.3 [ms/frame]	4.38 [pixel]	86.7 [%]
using approximation (proposed)	200 [times/frame]	128.7 [ms/frame]	4.54 [pixel]	85.2 [%]

References

[1] P. Viola, *et al.* Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2): 153–161, 2005.

[2] N. Dalal and B. Briggs. Histograms of oriented gradients for human detection. *CVPR*, 886–893, 2005.

[3] S. Munder, *et al.* Pedestrian detection and tracking using a mixture of view-based shape-texture models. *IEEE Trans. ITS*, 9(2): 333–343, 2008.

[4] M. Enzweiler and D. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. *PAMI*, 31(12): 2179–2195, 2009.

[5] P. Dollar, *et al.* Pedestrian Detection: A Benchmark. *CVPR*, 304–311, 2009.

[6] L. Spinello and K.O. Arras. People Detection in RGB-D Data. *IROS*, 3838–3843, 2011.

[7] J. Salas and C. Tomasi. People Detection Using Color and Depth Images. *MCPR*, 127–135, 2011.

[8] A. Ess, *et al.* Object detection and tracking for autonomous navigation in dynamic environments. *IJRR*, 29(14): 1707–1725, 2010.

[9] J. Satake and J. Miura. Stereo-Based Multi-Person Tracking using Overlapping Silhouette Templates. *ICPR*, 4304–4307, 2010.

[10] J. Satake and J. Miura. A Fast Stereo-Based Multi-Person Tracking using an Approximated Likelihood Map for Overlapping Silhouette Templates. *ACPR*, 392–396, 2011.