Robust Human Tracking Using Occlusion-free Images from Multiple Video Cameras

Kentaro TSUJI, Mingxie ZHENG, Eigo SEGAWA, Morito SHIOHARA, Takashi MORIHARA
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211-8588, Japan.
tsuji.kentarou@jp.fujitsu.com

Abstract

In this paper we present a robust human tracking method that can detect and track multiple individuals in crowded conditions. The method we have developed uses multiple ceiling-mounted cameras to overcome occlusion problems, making simple and robust image processing techniques effective. Tests performed under various conditions and times in a retail store environment have proven the method to show a high accuracy of detection and tracking capability. The simplicity of the method allows for real-time functionality with a single PC to process 16 camera data outputs covering the entire store area.

1. Introduction

In recent years, there is a growing demand from various fields for technology to automatically and quantitatively monitor human behavior in ways that thus far could only be monitored through human observation. For security applications, the ability to automatically track human behavior would make it possible to monitor comings and goings at restricted-access locations. In retail marketing, such technology could be used to analyze the in-store purchasing behavior of customers, including the frequency and amount of time spent in specific store areas. Consequently, technologies that automatically extract the motion paths of individuals from video data output, is one viable method to quantify such human behavior.

Boosting [1] is widely used for not only face detection but for human body detection as well. Though the method has proven its performance in face detection, it has shown difficulty in detecting a target's deformations and occlusions. Conversely, particle filtering [2] is well known for its capability in human tracking under conditions where occlusions exist. However, other detection methods must be employed to detect occluded targets at the first stage of the tracking process. Therefore, to realize stable human detection and tracking in real fields such as retail store environments, the most difficult problem is occlusion.

Examples of recent methods employed such as [3] and [4] assume using a single video camera or multiple cameras shooting at a diagonal angle. Such camera configurations can cause occlusions where a person is hidden behind other persons or objects. The occlusions increase relative to the density of people and width of the camera's FOV. To reduce the existence of occlusions dramatically, we propose using a configuration with multiple narrow FOV cameras ceiling-mounted and affixed to shoot the area directly below. It is reasonable to use multiple ceiling-mounted cameras because small and inexpensive cameras are available in recent years. Our human detection and tracking method functions in real-time while using multiple cameras to cover a wide area in monitoring range with the proposed camera configuration.

2. Video Camera Configuration

Fig.1 shows the video camera configuration to reduce occlusions. Considering the fact that people move along the ground surface plane and not in a vertical direction, there is no occlusion at the image center from the camera shooting directly below. Thus by using multiple cameras and setting the distance between them near enough, one of the images shot by these cameras can be occlusion-free for a particular target person. In such conditions, the cameras FOVs overlap with the others. This makes it easy to detect and track all target persons in an area. To best utilize the multiple camera configuration, all cameras should be calibrated in advance.

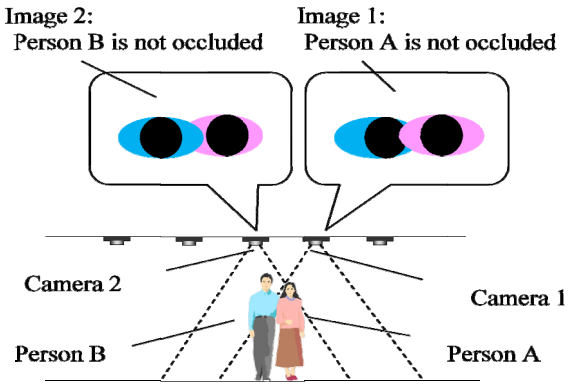

Figure 1. Video camera configuration.

3. Human Detection and Tracking

Fig. 2 shows the basic flow of our algorithm. The algorithm consists of three parts, taking advantage of the multiple images: background subtraction, hypothesis generation, and hypothesis verification (performed as particle filtering [2].) A sample of particle filtering corresponds to a hypothesis. Background subtraction, hypothesis generation and hypothesis verification cor-

respond to observation, drift and diffusion, and likelihood calculations respectively.

3.1. Background subtraction

We adopt a simple background model that represents a background as a single image. The background model is updated for each incoming frame with the static pixels as follows.

$$B_{i,j}(t+1) = B_{i,j}(t) + k\, f(B_{i,j}(t) - I_{i,j}(t)) \quad (1)$$

$$f(x) = \tanh(x) \quad (2)$$

$B_{i,j}(t)$ and $I_{i,j}(t)$ represents the background model and incoming frame at coordinates (i, j) at time t. Updated speed is controlled by the weight k.

The background model can be easily replaced with a more complex one such as a mixture of the Gaussian model as needed. Change detection is performed on each incoming frame. We use normalized correlation coefficients as the measure of difference between incoming frames and the background model for obtaining insensitive results under illumination changes. The pixels with coefficients small enough are classified as foreground pixels.

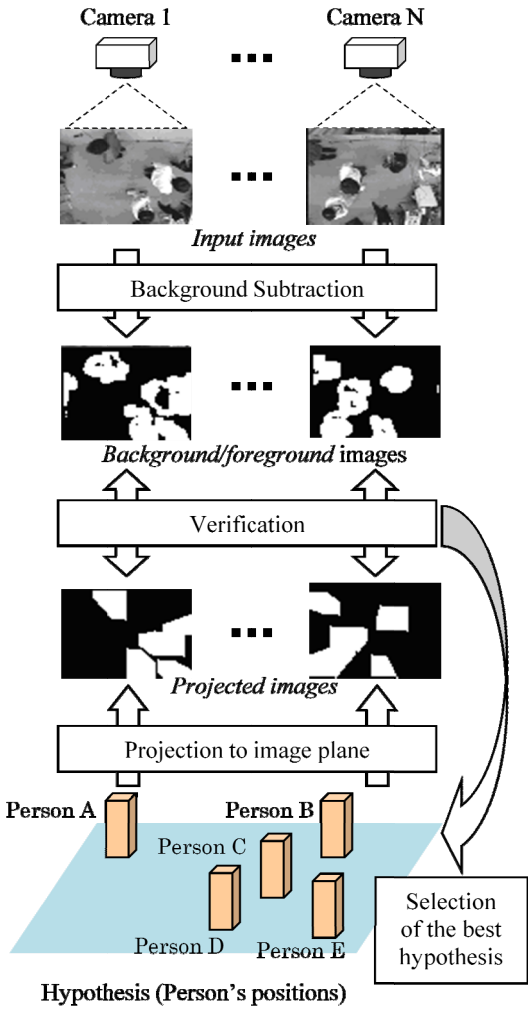

Figure 2. Basic flow of proposed method.

3.2. Hypothesis generation

We assume that target persons move on a single plane where the position of a person is represented as (x, y) coordinates on the plane. In addition, we model a human as a cube (shown in Fig. 2) for two reasons. One reason is the low computational cost to create projected images. The other is the insensitivity to pose changes in target persons. Under these assumptions, the state vector s_t at time t, that corresponds to a hypothesis in our algorithm and a sample of particle filter, is expressed by (x_t^i, y_t^i, h_t^i) where x_t^i, y_t^i and h_t^i represents i-th target person's position and height.

$$St = \{ x_t^1, y_t^1, h_t^1, x_t^2, y_t^2, h_t^2, \ldots, x_t^n, y_t^n, h_t^n \} \quad (3)$$

The state vector consists of the position of each target person taking into account the positional relationship between persons. For example, the positional relationship can be used to reject a hypothesis that two persons are too close.

Initial state estimation: As the initializing process, positions of target persons are determined based on background subtraction results as follows.

1) Background subtraction for all input images
2) Labeling of the result images
3) Check the sizes of each label near the image center
4) If the size is deemed an appropriate human size, the label is registered as a person and the position is calculated using the coordinates of the label center and camera parameters.

Heights of persons are set to the default value.

Drift: For the drift process of particle filtering, we use two types of drift, to utilize both the data that is obtained from observation, and that which comes from the knowledge about a target's motion. Drift based on observation is performed by using optical flow for determining the next position of a sample. The other drift is executed with a Kalman Filter assuming that a target moves at constant velocity. Half of samples are drifted by the two methods mentioned here respectively.

3.3. Hypothesis verification

To determine the best hypothesis, all generated hypotheses are projected on to all camera image planes, resulting in generation of the projected images as shown in Fig. 2. All projected images are checked against all foreground and background images to take advantage of the multiple cameras. Likelihood L_k of a hypothesis is defined as follows.

$$L_k = \sum_i \frac{(S_i - P_{k,i})}{A} + \alpha \cdot D_k \quad (4)$$

$$D_k = \begin{cases} 1 & \text{if } \min d^k{}_{i,j} < th \\ 0 & \text{else} \end{cases} \quad (5)$$

S_i represents the image result of background subtraction of i-th camera. $P_{k,i}$ represents the projected image of k-th hypothesis for i-th camera A is the number of pixels of an image. D_k is the penalty value for impractical positional relationship between persons of k-th hypothesis.

The value is 1 when any of the distances between two persons of k-th hypothesis is smaller than a predefined value. Otherwise, the value is 0. α is the weight value.

4. Experiments

We evaluated the proposed method using video data collected at a convenience store. To perform the evaluation, 16 cameras were ceiling-mounted to cover the entire area of the store, roughly 50m^2 of floor space. All camera views were set to shoot directly below with each camera FOV set at a horizontal angle of 67 degrees and a vertical angle of 53 degrees. The monitoring area and camera positions are shown in Fig. 3. Sample camera images are shown in Fig. 4.

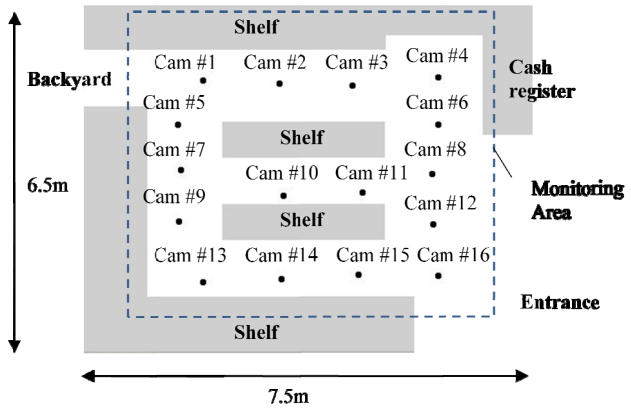

Figure 3. Store layout and camera configuration.

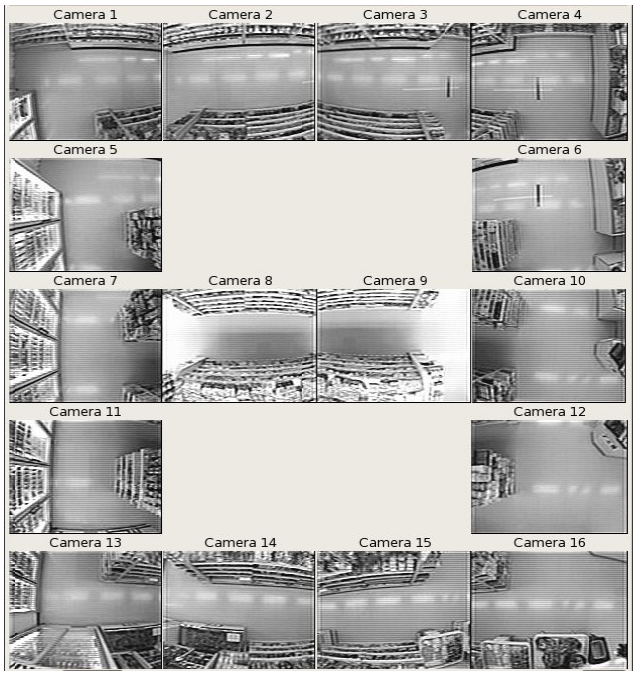

Figure 4. Sample images taken by 16 cameras.

System configuration: NTSC signal channel outputs from the 16 cameras were inputted to a 16-split multi viewer. The multi viewer then outputted a single channel of NTSC signal containing 16 video images of 1/16th the original size respectively. The single NTSC signal output was then inputted to a single PC, which processed the 16 video images of 80 by 60 pixels simultaneously.

4.1. Video data for evaluation

We collected store video data at various times to reflect a variety of store conditions, such as total numbers of people and people densities. Types of video data collected for evaluation are shown in Table 1.

Table 1. Video data for evaluation.

Data no.	1	2	3
Time	8:15-8:30	11:30-11:45	19:00-19:15
Total # [person(s)]	89	85	63
Ave. # [person(s)]	5.8	5.4	7.6
Max. # [person(s)]	13	12	12
Ave. density [person(s)/m^2]	1.1	1.1	1.2
Max. density [person(s)/m^2]	4	4	3

4.2. Results

We evaluated the performance of the proposed method by comparing the results with those obtained manually. For detection, a result was considered correct if the two bounding rectangles detected by our method and through manual detection overlapped by a ratio of more than 50%. Tracking was considered successful if all detections from a frame appearance to a frame disappearance were correct. Table 2 shows the results of the performance evaluations.

Table 2. Performance evaluation.

Data no.	1	2	3
# of persons	89	85	63
# of correctly detected persons	89	85	63
# of false positives	18	6	34
# of correctly tracked persons	75	74	42

Fig. 5 shows an example of detection for a certain time of data type no.1. Detected people were overlaid as cubes on the images. There were 13 people in the store and all of them were correctly detected though the density of people in front of the cash register was high (roughly 3 persons/m^2). Fig. 6 shows an example of tracking for the same time of Fig. 5. Lines in Fig. 6 represent motion paths of detected people for last 30 seconds. Motion paths starting from the entrance shows the continuous tracking was done in crowded conditions.

Fig. 7 shows the processing times using this method. Because processing time varies depending on the number of cameras and number of detected and tracked targets, we measured the time variation for different numbers of targets. All results were obtained by using a single PC

(Core 2 Duo, 2.4GHz) to process the 16 camera data outputs. Figure 7 shows that our method is effective at 30fps for tracking up to 11 targets when using 16 cameras.

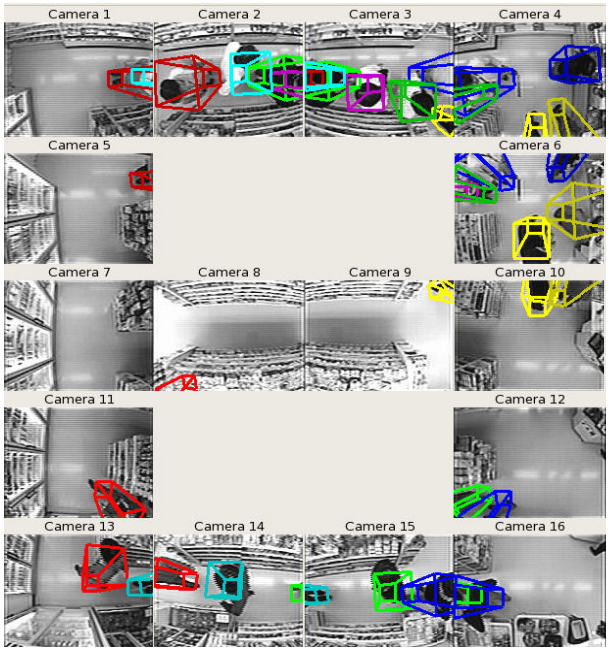

Figure 5. Example of detection.

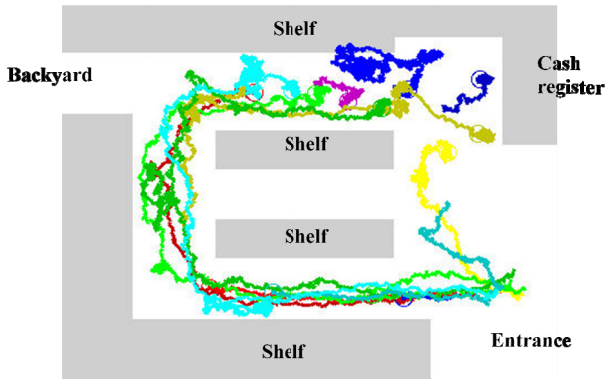

Figure 6. Example of tracking.

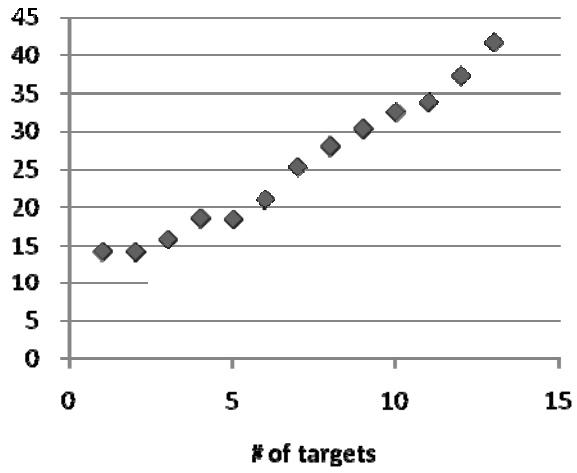

Figure 7. Processing time [ms].

4.3. Discussion

As shown in Table 2, the human detection rate was 100% from the store collected video data. However, the number of false positives was large, especially for data type no.3. The main cause of the false positives was due to background changes caused by the appearance and movement of non-human objects. Fig. 8 shows examples of false positives. In the left figure an opened refrigerator door was detected as a human. In the right figure a tray put in front of a shelf for restocking was detected. One way to potentially solve this problem is to combine more sophisticated human modeling techniques that can discriminate a human from other objects such as [1]. Another way is to use the knowledge about the monitoring area such that a human appears only at the entrance. Adding more cameras can help reducing false positives as well. False tracking results were related to false positives, because in many cases switching, which is the tracking of correct persons to others, occurs with false positives. Thus decreasing the number of false positives is expected to improve tracking accuracy.

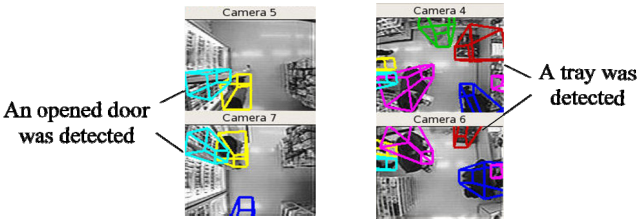

Figure 8. Example of false positives.

5. Conclusion

We have developed a robust human tracking method that can detect and track multiple individuals in crowded store conditions. Our method uses multiple ceiling-mounted video cameras to overcome occlusion problems, making simple and robust image processing techniques effective. Performance evaluations proved the method to be highly accurate at human detection and tracking. The simplicity of the method allows for real-time functionality with a single PC to process 16 camera data outputs covering the entire store area.

References

[1] P. Viola and M. Jones: "Rapid Object Detection Using a Boosted Cascade of Simple Features," *International Conference on Computer Vision and Pattern Recognition*, vol.1, pp.511-518, 2001.

[2] M. Isard and A. Blake: "CONDENSATION - Conditional Density Propagation for Visual Tracking," *International Journal of Computer Vision*, vol.29, no.1, pp.5-28, 1998.

[3] I. Haritaoglu, et al.: "W4: Real-time Surveillance of People and Their Activities," IEEE Trans. PAMI, vol.22, no.8, pp.809-830, 2000.

[4] T. Zhao and R. Nevatia: "Tracking Multiple Humans in Complex Situations," IEEE Trans. PAMI, vol.26, no.9, pp.1208-1221, 2004.