Detection of Different Behavior from The Majority in A Public Space

Kohei Saito* and Hiroshi Hanaizumi,

Graduate School of Computer and Information Sciences, Hosei University,
3-7-2, Kajiocyo, Koganei, Tokyo, 184-8584, Japan
kohei.saito.7a@stu.hosei.ac.jp

Abstract

A measurement method was proposed for observing movements of persons in a public space such as a railway station. The method was based on the fact that a person occupied a 3D region and his or her sequential images captured by a security camera formed a wormhole in the spatiotemporal image space. Direction of the majority of those wormholes was estimated as movement direction of the majority by a projection algorithm. After filling the majority wormholes, remaining ones were searched and recognized as different behavior persons. A high-accuracy person extraction algorithm was also developed for preventing wormholes enlargement mainly caused by persons' shadow. An inter-frame difference technique was adopted for realizing the quick updating the background. The method was successfully applied to a security camera video for detecting different behaviors from the majority.

Keywords: spatiotemporal space, background subtraction, projection image, suspicious person detection.

1. Introduction

The importance of securing safety in a crowded public space has been increasing in recent years. Large number of security cameras have also introduced into those places. The explosion of those video data to be processed became a moment for developing methods to recognize human behavior and to detect persons automatically who took suspicious behaviors, and many methods have been developed⁽¹⁻⁴⁾. Most of them were based on the person tracking⁽⁵⁻⁷⁾, where person one-by-one detection and/or recognition were required before the tracking. They often adopted a background subtraction algorithm for the person detection in video scenes⁽⁷⁻⁹⁾.

The background subtraction, however, has some problems to be solved, especially in producing/updating the

background image. In order to extract a person exactly, the background image should be produced faithfully and be updated dynamically for preventing false area extraction due to such as daily changes of lighting condition. Actually, some of the methods use the mode of pixel density in a time window as an estimate of pixel density in the background image (here, we call it as MODE algorithm). A wider window gives us a more stable extraction but a slower updating. The slow updating is a cause of low followability for the change of lighting condition⁽¹⁰⁻¹²⁾. Unfortunately, MODE algorithm has less performance for the crowded scenes where the mode does not necessarily represent the background.

Thus, the background subtraction algorithm itself extracted persons with low precision even in a dense situation, but the conventional person detection methods were not able to process overlapped persons separately. A new person detection method with robustness for the crowded scenes has been required.

Here, we propose a new method, Detecting Different Behavior from the Majority (DDBM), with robustness against the crowded scene such as a railway station. The DDBM adopts a wormhole algorithm in a spatiotemporal space⁽¹³⁻¹⁵⁾ and represents person movements as a bunch of wormholes. A projection technique enables us to get average direction corresponding to that of the movement of the majority people. The different behavior person is detected as a wormhole having another direction from the average. Separability among the wormholes controls the performance of DDBM in detecting person one-by-one. So, an improved algorithm for the background subtraction is also introduced into DDBM. The algorithm, Background Quick Updating (BQU), is characterized by its high speed updating the background. The BQU well prevents the enlargement of the wormholes due to persons' shadows.

In this paper, we describe the principle and procedures of DDBM and BQU, respectively. Application results of

DDBM to an actual crowded scene are also shown.

2. Methodology

The proposed method DDBM is based on the fact that a person occupies a 3D region and his or her sequential images captured by a camera formed a wormhole in the spatio-temporal image space. This means that those wormholes may have contact but never merge. They are separated from each other even if in the crowded scenes. Figure 1 shows the schematic diagram of the proposed method DDBM. Video data captured by a camera at a fixed point are stored and used for the processing. Firstly, we separate moving persons as areas having value 0 from unmoved background having value 1. This process is applied to every scene in the video. Then we pile up these binary images to produce a spatiotemporal space. After searching the major direction of the moving persons in the spatiotemporal space, wormholes with the direction are filled by value 1. The residual wormholes indicate us that the different behaviors from the majority of people are detected.

Fig. 1 Schematic diagram of the proposed method DDBM.

2.1 A Wormhole Algorithm

We assume that persons move straight at least in the field of view of a camera used. Figure 2 (a) simplifies the movements of persons binary extracted from the video scenes. On the assumption, they form straight wormholes in the time sequence of the binarized scenes, i.e. a spatio-temporal space as shown in Fig. 2 (b). The direction of the majority of wormholes are searched, and the major wormholes are filled as shown Fig.2(c). The residual wormholes are detected as the different behaviors from the majority as shown Fig.2 (d).

The spatiotemporal model S_l has size of $W_l \times H_l \times D_l$ in the coordinates (X, Y, T) as shown in Fig. 3. Let another

Fig. 2 Extracted persons in a video scene, their wormholes in the spatiotemporal space and filtering in it.

cubic model S_c be defined in S_l . The cube S_c , whose center corresponds to that of S_l , has size of $Q_c \times Q_c \times Q_c$ in local coordinates (U, V, W), where

$$Q_c \leq \frac{1}{\sqrt{3}} \min(W_l, H_l, D_l). \quad (1)$$

In order to obtain the direction of the majority of the wormholes, we produce projection images for various directions. The major direction is obtained as one which gives the maximum contrast among them. The maximum contrast is achieved when the projection direction matches to that of the majority of wormholes. Let α , β and γ be rotation angles in axis X, Y and T, respectively. Pixel densities of the projection image $I_p(u,v)$ is calculated as

Fig. 3 Spatiotemporal models S_l and S_c .

$$I_p(u, v) = \sum_{w=0}^{Q_c} g_w(u, v) \quad (2)$$

where, $g_w(u, v)$ represents voxel density at a position (u, v, w) in S_c . The contrast is defined as

$$C = \left| \max_{u,v} I_p(u, v) - \min_{u,v} I_p(u, v) \right| + Q_c B, \quad (3)$$

where, B means number of pixels having value 0 in the projection image I_p .

Coordinates (u, v, w) also calculated from (x, y, t) as

$$\begin{pmatrix} u \\ v \\ w \\ 1 \end{pmatrix} = P_C R_T R_Y R_X P_{-C} \begin{pmatrix} x \\ y \\ t \\ 1 \end{pmatrix}, \quad (4)$$

where, R_T, R_Y, R_X are rotation matrices, and P_C, P_{-C} parallel shift ones, respectively, for example,

$$R_X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \alpha & -\sin \alpha & 0 \\ 0 & \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, P_C = \begin{pmatrix} 1 & 0 & 0 & W/2 \\ 0 & 1 & 0 & H/2 \\ 0 & 0 & 1 & D/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Angles α_m, β_m and γ_m corresponding to the direction of the majority of the wormholes is obtained by using a steepest descent method. Considering C in (3) as a function of α, β and γ , we find the maximum direction as

$$C_m(\alpha_m, \beta_m, \gamma_m) = \max_{\alpha, \beta, \gamma} C(\alpha, \beta, \gamma). \quad (5)$$

In the projection image at the angle α_m, β_m and γ_m , we call low pixel density areas as wormhole areas. We define that the wormhole area has larger extent than E_p and mean density lower than B_p . Every wormhole is filled with a polygonal prism whose base corresponds to the wormhole area recognized in the projection image (see Fig.2 (c)). After filling them, residual wormholes are detected and recognized one-by-one from the most prominent ones by applying the same procedures above.

2.2 Background Quick Updating algorithm

In a fixed camera system, we define pixels having almost the same density in two successively captured images as background pixels, though most of the conventional methods adopted the definition that pixels having the mode density in a time window were the background pixels. Figure 4 shows the processes of the BQU algorithm, where A^{t-1} and A^t are successive images, and S a switch for updating the background pixel density. The switch S is closed (ON) when the Manhattan distance M_h in the RGB feature space between

Fig. 4 Processes in the BQU.

densities of pixels at the same position (x, y) in the successive images as

$$M_h(A^{t-1}(x, y), A^t(x, y)) < 3. \quad (6)$$

On the other hand, when the Manhattan distance is equal to or greater than 3, the switch S is opened (OFF) and the background is not updated. The BG is used as the background and difference as the result from the background subtraction is obtained as FG. In the RGB feature space, we measure the Euclidian distance E_d between a background pixel $BG(x, y)$ and the current observation $A^t(x, y)$, where x and y means position of the pixel. We regard that a moving person is detected when the distance exceeds a threshold R , as

$$FG(x, y) = \begin{cases} 0 & E_d(BG(x, y), A^t(x, y)) > R \\ 1 & E_d(BG(x, y), A^t(x, y)) \leq R \end{cases}. \quad (7)$$

The zeros in $FG(x, y)$ produce a part of the wormhole in the spatiotemporal model S_i .

3. Experiments and Discussion

In order to evaluate the performances of the BQU and the DDBM, a 20 minute-length video captured at a railway station was processed. The video was divided into color images with size of 720×480 and stored in a folder. A Windows-10 PC with 3.60 GHz clock CPU and 32 GB memories was used for the processing in the experiments. In the experiments, instead of reading images from a camera, images were sequentially read from the folder.

3.1 Performance Evaluation of BQU algorithm

The performance of the Background Quick Updating algorithm was evaluated by comparing the that of the conventional MODE algorithm, especially in extracting accuracy and in eliminating foot shadows. In the MODE, we set the time window as 16.7 [s] ($= 1/30$ [s] $\times 500$ frames).

Fig. 5 An image in the video used.

(a) Person areas

(b) BQU

(c) MODE.

Fig. 6 Manually selected parson areas (black) (a), those extracted by BQU (b) and by MODE (c).

Figure 5 shows an image in the video captured at a crowded railway station we processed. Red box at the left-hand side indicates the area in which we numerically evaluated the performance of BQU and MODE as shown in Figs. 6. We manually separated persons (black area) from the background (white) as shown in Fig. 6 (a), and used it as the truth. Figures 6 (b) and (c) were the areas extracted as moving persons by BQU and by MODE, respectively. The precision and the recall were calculated from those images as shown in Table 1. We see the BQU have better performance in both the precision and the recall than the MODE does.

On the other hand, red box at the right-hand side in Fig.5 indicates the area in which we can see the superiority of the BQU as shown in Figs. 7. Actually, the BQU clearly extracted two persons without shadowing at their feet, though the MODE extracted them with the shadows. The shadow decreased the performance in recognizing persons

Table 1 Numerical evaluation of the person extraction.

	DDMB	MODE
PRECISION	0.878	0.835
RECALL	0.745	0.700

(a) persons extracted by BQU (b) ones by MODE

Fig. 7 Comparing the performance in prevention of the shadowing at feet between the BQU and the MODE.

separately. Thus, the performance was confirmed that the BQU was able to extract moving persons satisfying the conditions necessary for the DDBM.

3.2 Performance Evaluation of DDBM

Figures 8 (b) shows a 3D spatiotemporal model S_l built from the binary images (see Fig.8 (a)) in which moving persons were extracted by the BQU. Considering the symmetry, the size of the 3D model was determined as $480 \times 480 \times 480$ [pixel³], where the size was limited by the smaller one of the width and the height of the original images. The smaller model S_c was also defined with size of $270 \times 270 \times 270$ [pixel³], and used for searching the moving direction of the majority of persons captured in the video scenes.

The direction searching process was applied to the model, and the projection image was obtained as shown in Fig. 9, which maximized the contrast C in (3). In the process, a lattice point search using the model S_c was applied firstly, then the maximum contrast lattice point was used as the starting point in the steepest descent method. Actually, the contrast $C(\alpha, \beta, \gamma)$ was calculated for a total of 27 combinations of α , β and γ selected from 45, 90 and 135 degrees. The steepest descent method gave us the maximum contrast at angles $(\alpha_m, \beta_m, \gamma_m) = (134.02, 44.55, 89.37)$ degrees, respectively. Figure 10 (a) shows the 3D view of the model S_l from the direction of $(\alpha_m, \beta_m, \gamma_m)$ with removing duct of the ceiling for better understanding the inside. Black regions are the so-called wormholes yielded by movements of the majority persons. In order to fill the

(a) A binary Image of persons

(b) 3D spatiotemporal model

Fig. 8 A binary image of persons extracted by BQU (a) and the 3D spatiotemporal model built by piling the images (b) with coordinates (X, Y, T).

Fig. 9 Projection image along the major direction.

regions, the projection image (see Fig.9) was binarized by using the threshold of the mean density. The binary image was back-projected onto the model S_i as shown Fig.10 (b). The white blocks filled the dark regions in S_i , i.e. movements of the majority persons, as shown in Fig.10 (c). Remaining black regions in the model S_i are the part of trajectories of persons taking different behavior from the majority to be

(a) model with removing duct

(b) regions majority persons moving

(c) remaining regions

Fig. 10 Three dimensional views of the spatiotemporal model built from the binarized images with removing the white block due to duct of the ceiling in the video scene.

detected.

In order to recognize the different behavior person from the majority, we applied the same procedures as the previous majority movement search. Figure 11 (a) shows the projection image giving us the maximum contrast for the remaining wormholes. The projection image suggested us

(a) possible trajectories (b) one of them with binarization
Fig.11 Projection image for finding possible holes of person moving.

that dark regions were separated from each other, and therefore, we could recognize the person one-by-one. We selected a patch marked by an ellipse, and applied the steepest descent method so that the area of the patch had the minimum extension. Figure 11 (b) shows the patch minimized, then binarized with the mean density. The direction angles were obtained as $(\alpha_m, \beta_m, \gamma_m) = (-110.53, -0.37, 39.89)$ degrees, respectively. The patch was projected back onto S_l , and we got the 3D region corresponding to the trajectory of the person in the spatiotemporal space. The 3D region was used as a window in the spatiotemporal space. We regarded the 3D region as the 2D mask moving at a constant speed in the original video scenes, and considered that a person was recognized when the mask included him/her in the video scenes during his or her appearance. Figures 12 show three scenes of the original video with superimposing the mask. We see that the mask includes an identical person in every scene. Thus, the proposed method DDBM correctly recognized a different behavior person from the majority.

The proposed method was currently inefficient. The method took 95 [s] for extracting moving persons and building the 3D spatiotemporal model S_l from the video scenes, 255 [s] for the lattice point search using the model S_c , and 3403 [s] for the fine search for the movement direction of the majority persons using the steepest descent method with 24 iterations. After filling the model, it took 1103 [s] for getting the direction of a different behavior person, and 126 [s] for producing the 2D mask for marking up the person recognized.

In the results shown in Fig. 12, the shape of the mask is not similar with that of the person recognized. We consider the phenomenon is caused by the complexity of persons walking in a crowded area. They may change the walking direction or stop to avoid colliding with others. Those

(a) at time t_1

(b) at time t_2

(c) at time t_3

Fig. 12 A person recognized as one taking different behavior from the majority ($t_1 < t_2 < t_3$).

behaviors make their trajectories different from lines, though the proposed method DDBM assumes linear trajectories. So the estimated trajectory patch becomes wider polygonal

prism enclosing the entire complex trajectory. The projection technique also may include a part of the trajectory yielded by another person's body.

4. Conclusions

As a preliminary step of detecting suspicious individuals in crowded places, a measurement method was proposed for observing persons' movements in a public space such as a railway station. The method was based on the fact that a person occupied a 3D region and his or her sequential images captured by a security camera formed a wormhole in the spatiotemporal image space. The proposed method was divided into two sub-methods; Background Quick Updater (BQE) and Detecting Different Behavior from the Majority (DDBM). The former introduced an inter-frame difference technique into producing the background and well extracted moving persons from the original video scenes with avoiding their shadows. The latter built a 3D spatiotemporal model from the extracted areas with binarization, in which the persons' movements were recorded as so-called wormholes. On the assumption that persons in a video scene move straightly with a constant speed, the direction of the majority of the wormholes was searched so that the contrast in the projection image reached the maximum by using a projection technique. After filling wormholes along the major direction, remaining ones were searched and recognized as the different behavior persons. The proposed method successfully detected a person crossing the stream of the majority persons in an actual video captured at a railway station.

To improve the method especially in efficiency, to develop an automated algorithm for separately recognition of remaining wormholes after filling the majority persons' trajectories and to extend the method from detection of different behavior persons to finding suspicious individuals are the subjects for a future study. Development of an intelligent algorithm for the background subtraction is also one of future tasks.

References

- (1) Wassima Aitfares, Abdellatif Kobbane, and Abdelaziz Kriouile, "Suspicious Behavior Detection of People by Monitoring Camera," Proc. ICMCS-2016, p113-117, October, 2016.
- (2) Xinyi Cui, Qingshan Liu, Mingchen Gao, and Dimitris N. Metaxas, "Abnormal Detection Using Interaction Energy Potentials," Proc. CVPR-2011, p3161-3167, June, 2011.
- (3) NikoLaus Viertl and Csaba Belezai, "A pedestrian detection system combining motion detection, spatial grouping and classification," Proc. ISPA-2009, p194-199, September, 2009.
- (4) Min Sun, Dongping Zhang, Leyi Qian, and YeShen, "Crowd Abnormal Behavior Detection Based On Label Distribution Learning," Proc. ICICTA-2015, p345-348, June, 2015
- (5) CHI Haihong, LIU Lei, and SONG Hanlin, "Particle Filter Based Moving Object Tracking in Dynamic Scene," Proc. CCC-2015, p3969-3974, July, 2015.
- (6) Shengluan Huang and Jingxin Hong, "Moving Object Tracking System Based On Camshift And Kalman Filter," Proc. CECNet-2011, p1423-1426, April, 2011.
- (7) Dongliang Jin, Songhao Zhu, Xian Sun, Zhiwei Liang, and Gupzheng Xu, "Optical Flow and Spatio-temporal Gradient Based Abnormal Behavior Detection," Proc. CCDC-2016, p1532-1537, May 2016.
- (8) Birmohan Singh Deptt, Dalwinder Singh, Gurwinder singh, and Neeraj Sharma Deptt, Vicky Sibbal Centre, "Motion Detection for Video Surveillance," Proc. ICSPCT-2014, p578-584, July, 2014.
- (9) Milin P. Patel and Shankar K. Parmar, "Moving Object Detection with Moving Background using Optic Flow," Proc. ICRAIE-2014, May, 2014.
- (10) Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew, "On the Removal of Shadows from Images," Proc. TPAMI-2005, Vol28, p59-68, November, 2005
- (11) Olivier Barnich and Marc Van Droogenbroeck, "A Universal Background Subtraction Algorithm for Video Sequences," Proc. IEEE IMPR-2010, Vol.20, p1709-1724, December 2010.
- (12) Surendra Gupte, Osama Masoud, Robert F. K. Martin, and Nikolaos P. Papanikolopoulos, "Detection and Classification of Vehicles," Proc. IEEEITS-2002, Vol. 3, p696-702, August, 2002
- (13) Kazuhiro Otsuka, Tsutomu Horikoshi, and Satoshi Suzuki, "Image Velocity Estimation from Trajectory Surface in Spatiotemporal Space," Proc. IEEECS-1997, p200-205, June, 1997.
- (14) Per Ola Kristeansson, Nils Dahlback, Daniel Anundi, Marius Bjoernstad, Hanna Gillberg, Jonas Haraldsson, Ingrid Martensson, Mathias Nordvall, and Josefine Stahl, "An Evaluation of Space Time Cube Representation of Spatiotemporal Patterns," Proc. TVCG-2008, Vol. 15, p696-702, November, 2008.
- (15) Minh Vo, Srinivasa G. Narasimhan, and Yaser Sheikh, "Spatiotemporal Bundle Adjustment for Dynamic 3D Reconstruction," Proc. CVPR-2016, p1710-1718, June, 2016