Contour-based Pedestrian Detection with Foreground Distribution Trend Filtering and Tracking

Kiat Siong Ng, Min-Chun Hu, Yu-Jung Hsiao, Kuang-Yu Nien, and Pei-Yin Chen
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan City, Taiwan (R. O. C)
jason.ng@mail.csie.ncku.edu.tw, anita_hu@mail.ncku.edu.tw, yami.hsiao@gmail.com,
kevin11030@gmail.com, pychen@mail.ncku.edu.tw

***Abstract*— In this work, we propose a real-time pedestrian detection method for crowded environments based on contour and motion information. Sparse contour templates of human shapes are first generated on the basis of a point distribution model (PDM), then a template matching step is applied to detect humans. To reduce the detecting time complexity and improve the detection accuracy, we propose to take the ratio and distribution trend of foreground pixels inside each detecting window into consideration. A tracking method is further applied to deal with the short-term occlusions and false alarms. The experimental results show that our method can efficiently detect pedestrians in videos of crowded scenes.**

Keywords—pedestrian detection; sparse contour; foreground distribution trend; crowded scene

I. INTRODUCTION

Pedestrian detection is essential for intelligent surveillance systems. Knowing the number of pedestrians in a scene can benefit demographic statistics and have many applications. For example, a retail store manager may assign more staff to serve customers if the surveillance system detects more people in a certain area. Moreover, by analyzing the clothing appearance of a detected person enabling us to search the same person among an enormous collection of surveillance videos. There have been many types of research on pedestrian detection with different focuses: 1) How to detect pedestrians of various sizes and poses, 2) how to make the detection system robust under different lighting conditions, and 3) how to solve the shadow and occlusion problems especially in crowded scenes [1, 2]. For example, based on a Multi-Task model to jointly consider the commonness and differences, Yan [3] conducted resolution aware transformations to map local features from different resolutions to a common subspace. Huang [4] proposed the image-range fusion system (IRFS) in which a dynamically illuminated object (DIO) detector is designed to overcome the problem caused by uncertain partial lighting condition. Liao [5] utilized the spatiotemporal background extractor to detect moving objects and applied Random Forest to learn the shad model with shadow features. The trained Random Forest shadow detector is further used to remove the shadow from the result of the spatiotemporal background extractor. Tang [6] analyzed typical failures of human trackers and trained a detector explicitly using these cases to detect human with partial occlusion. Utilizing both motion and shape clues, Hu [7] targeted real-time pedestrian detection in crowded environments and showed that their method is efficient. Using the foreground filtering step proposed by Hu, the detection accuracy would be low if the background subtraction result is too noisy. In this work, the shadow detection method [8], which pre-selects shadow pixels based on chromaticity and then calculates the correlation of the gradient direction between the given frame and the background, was utilized to decrease false alarms caused by shadows. Furthermore, we modified the filtering step and smoothed out the detected results by using a short-term tracking procedure, which can more accurately detect pedestrians without significant processing overhead.

II. PROPOSED METHOD

The proposed system framework is shown in Fig. 1. Since pedestrians have similar shapes, we first trained the sparse contour templates of pedestrians based on the point distribution model (PDM) [9]. Then the contour templates were used to detect pedestrians in a template matching step. However, applying template matching to all scanning windows results in heavy computation time. To reduce the number of template matching procedures, the background subtraction technique was first applied to extract moving foreground pixels, and a candidate filtering step was used for reducing the number of detecting windows, where the template matching should have been performed. The candidate filtering step was designed according to the ratio/distribution of foreground pixels inside each detecting window. Moreover, a tracking method was applied to deal with short-term occlusions and noises.

A. Templates Generation and Matching

Sparse contour representation and PDM are applied to generate representative templates of human shapes. Fig. 2 shows 30 generated human templates, and each template consists of 13 line segments. This human template generation method has been proven to be effective for tolerating shape variations by Beleznai [9].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

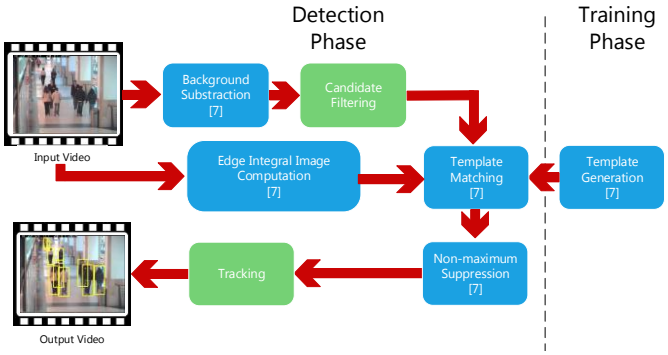

Fig. 1. Overview of the proposed system.

Images were scanned with a sliding detecting window, and a template matching procedure was used to determine whether the content inside the current window was similar to any of the generated templates. The size of the detecting window was adaptively determined by the regression method proposed by Hu [7], and the templates were normalized to the same size as the detecting window before matching. The input image was denoted as I, the content in the current detecting window as $I_{x,y}$, and the compared template as T. The matching score between $I_{x,y}$ and T was determined according to the consistency of edge points in $I_{x,y}$ and contour points in T. Given I, we first generated eight corresponding edge maps by applying Gabor filters, and each map $EM(\theta)$ recorded edge points in one specific orientation defined in $\Theta = \{0°, 26.57°, 45°, 63.43°, 90°, 116.57°, 135°, 153.43°\}$. Given a template T, each line segment L_i was adjusted to the nearest orientation θ among Θ, and the new line segment was denoted as $\hat{L}_i$. The matching score between $I_{x,y}$ and T was defined as

$$S(I_{x,y},T) = \prod_i \frac{EI(EM(\theta),\hat{L}_i)}{Length(\hat{L}_i)} \quad (1)$$

where $EI(EM(\theta),\hat{L}_i)$ was the edge integral value [10], along the line segment $\hat{L}_i$ on the edge map $EM(\theta)$. Note that a small value ϵ will be assigned to $EI(EM(\theta),\hat{L}_i)$ if it is zero, and therefore this formula can deal with partial occlusions. A non-maximum suppression step was applied to more accurately locate the positions of pedestrians in each frame.

B. *Candidate Filtering*

Exhaustively matching all sliding windows with the 30 templates is inefficient. Since pedestrian regions should contain enough foreground pixels, Hu [7], proposed to utilize the foreground ratio inside the detecting window to reduce the

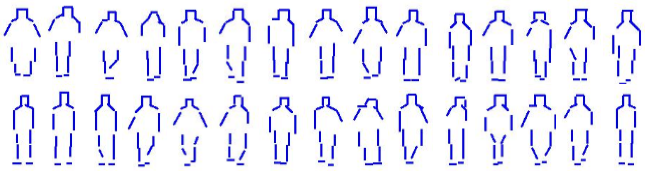

Fig. 2. The generated human sparse contour templates.

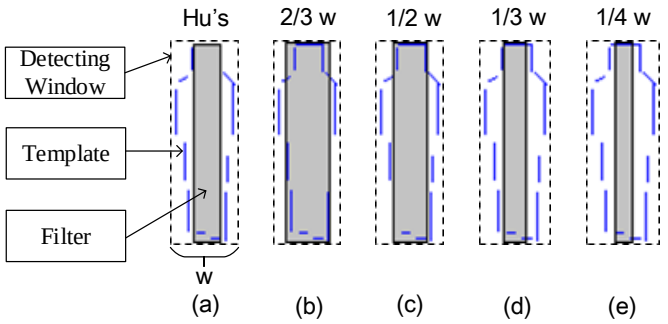

Fig. 3. Examples of different center-body regions: (a) Center-body used in [7]. (b-e) use 2/3, 1/2, 1/3, and 1/4 of the detecting window to define the center-body region, respectively.

matching number. This method also diminishes false alarms occurring when more than two subjects are gathered in a window and resulting a human-like shape between the two pedestrians. However, this filtering method is too naïve and does not work well when the background subtraction result is not robust. In this work, we modified the filtering step in order to find more reliable pedestrian candidates for further template matching by considering both the ratio and the distribution trend of foreground pixels inside the detecting window. In the method proposed in [7], the center-body region (as shown in Fig. 3(a)) was determined according to the head width of the template, and the template matching step will not be applied to the window having few foreground pixels inside the center-body region. Unfortunately, as shown in Fig. 2, the head width of the generated templates might be quite small such that the center-body region would be too narrow for providing discriminative foreground ratio information. We defined the center-body region in different ways (as shown in Fig. 3(b)-(e)), and empirically found that if the width of the detecting window is w, then using the middle $w/3$ area of the detecting window as the center-body region would result in better accuracy.

The horizontal/vertical foreground distribution trend inside the detecting window is also an important clue for candidate filtering. As shown in Fig. 4, we horizontally divided the detecting window into 10 equal regions and computed the foreground ratio in each bin. Fig. 4(a)-(c) show the horizontal foreground distribution trend of a real pedestrian. Fig. 4(d)-(f) show that even though a detecting window might have many foreground pixels inside the center-body region because it contains many human parts from multiple pedestrians, we can still distinguish it from real humans according to the foreground distribution trend. Compared to the filtering method described in [7], our method took an additional step to calculate the foreground distribution trend. It is worth noting that the additional step did not increase the overall detection time because it significantly reduced the number of candidates for the template matching procedure.

C. *Short-term Tracking*

After frame-by-frame detection, we further applied a short term tracking process to smoothen the detection results. We defined 30 consecutive frames as a tracking unit, and maintained several tracking lists for each tracking unit. The lists were used to record the detected pedestrians from different frames. If two

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

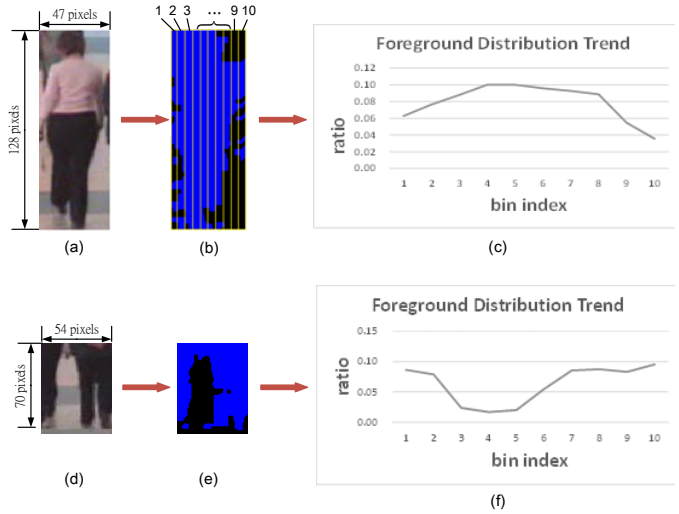

Fig. 4. The concept of foreground distribution trend. (a) and (d) show the content of a pedestrian and a non-pedestrian region, respectively. (b) and (e) show the foreground pixels in (a) and (d), respectively. (c) and (f) show the foreground distribution trend of (b) and (e), respectively.

pedestrians in adjacent frames are close and have similar appearances, we put them into the same tracking list. The appearance similarity was defined based on the chi-squared distance of the LAB color histogram. As shown in Fig. 5 the pedestrian in tracking list (b) is lost in some frames due to occlusions. In this case, we interpolate missing detections if the number of consecutive null nodes is less than three. Also, shadows may cause false detections and result in rather short tracking lists such as the list show in Fig. 5 (c). In this case, we simply delete lists shorter than 10 (frames).

III. EXPERIMENT RESULTS

The CAVIAR database [10] is a benchmark for human detection evaluation. We utilized a video named as 'OneStopMoveEnter1cor' to perform our detection algorithm. The specification of this video and our experimental environment are listed in Table 1.

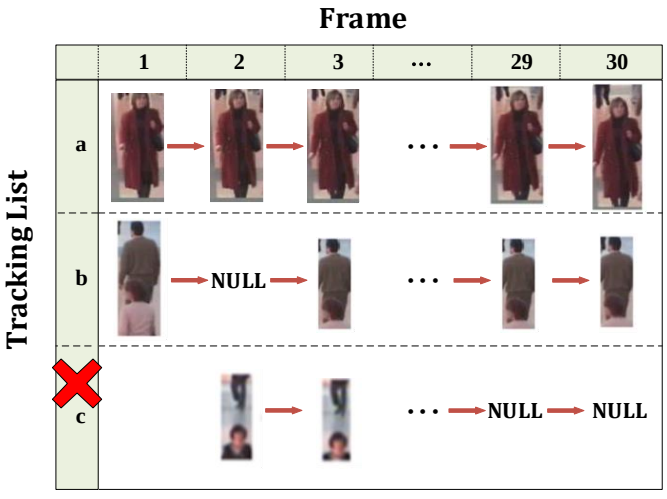

Fig. 5. The process of establishing tracking lists.

TABLE I. SPECIFICATION OF VIDEO AND EXPERIMENT ENVIRONMENT

Specification of Video	
Frame Rate	25 fps
Frame Resolution	384 x 288 pixels
Specification of Experiment Environment	
CPU	Core-i5 (use only dual core)
Graphic Card	NVIDIA GeForce GT 630M

Some example results of the proposed pedestrian detection method is shown in Fig. 6 below. In the experiment, the background subtraction step spends a few seconds in learning background information. Hence, the video sequence from frame 201 to 1,588 are used for accuracy evaluation. We exclude the pedestrian object visibility less than 50% in the evaluation process. In total there are 11,149 pedestrians in 1,588 frames. The detected pedestrian was considered as a true positive if it matches the ground truth by the following criterion:

$$\max\left(\frac{A_{AND}}{A_{GT}}, \frac{A_{AND}}{A_{DR}}\right) > 0.5 \qquad (1)$$

where A_{GT} and A_{DR} are the area of ground truth and the detected pedestrian, respectively. A_{AND} denotes the overlapping area between A_{GT} and A_{DR}. We first investigated into the effect of applying different definitions of center-body region, including method [7], the middle $2w/3$ area of the detecting window, the middle $w/2$ area of the detecting window, the middle $w/3$ area of the detecting window, and the middle $w/4$ area of the detecting window. As shown in Figure 7, using the middle $w/3$ area of the detecting window to construct the center-body region resulted in better performances. The recall increased approx. 3% and the precision increased approx. 0.7% compared to method [7].

We further evaluated the effectiveness of the proposed candidate filtering method, which included a center-body-based foreground ratio filter and a foreground distribution trend filter. On average, adding the foreground distribution trend filter increased the precision by 3% and decreased the recall by less than 1%. If we further apply the short-term tracking process, the deletion mechanism can minimize the misjudgment of detected pedestrians, and the interpolation mechanism can recover some

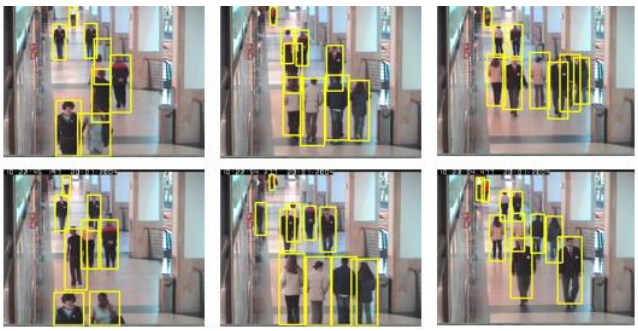

Fig. 6. The demo of our proposed pedestrian detection system.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

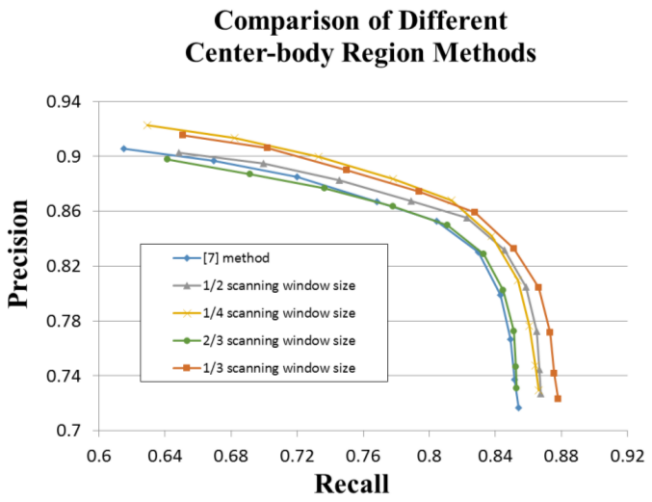

Fig. 7. The PR curve of the proposed pedestrian detection method using different center-body region definitions.

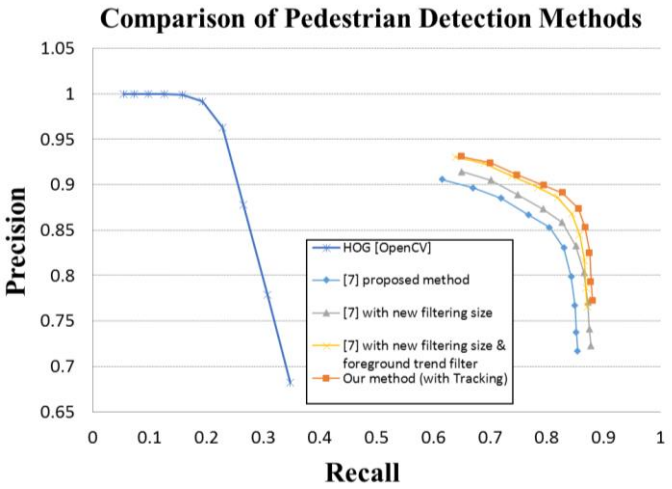

Fig. 8. The P-R curves of the proposed pedestrian detection method with using the best filtering size, foreground distribution trend filter and deletion/interpolation of our tracking algorithm. Results show that our method achieves higher precision and recall than method [7].

missing pedestrians. Fig. 8 shows that comparing to method [7], our proposed pedestrian detection method successfully increases the precision and recall by 5.1% and 3.6%, respectively. Moreover, comparing with the HOG method built in OpenCV, our method has a much higher recall with the same precision.

Fig. 9 shows the computation cost of applying different candidate filtering methods to detect pedestrians in one frame. For each method, the scanning step of the detecting window is 2 pixels, and a GPU is applied in the edge detection step. Since template matching is a procedure that can be independently conducted in each detecting window $I_{x,y}$, we can apply parallel programming techniques and utilize multi-core CPUs to achieve real-time detection. No matter we use [7] or our proposed definition of center-body region, adding the foreground distribution trend filter reduced the number of candidates for template matching, and consequently reduced the overall pedestrian detection time by about 2ms per frame. The overall

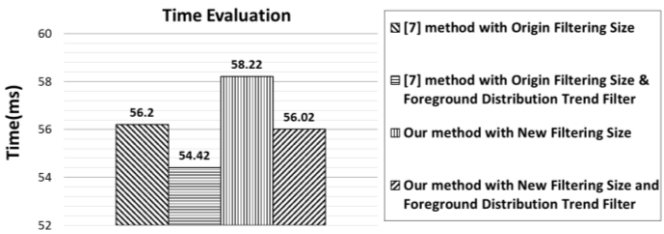

Fig. 9. Different filter size effect on execution time of detecting a frame.

execution time of our method is 56.02ms per frame, which is lower than that of the method proposed in [7].

IV. CONCLUSIONS

We propose a real-time pedestrian detection method for crowded environments based on contour and motion information. Sparse contour templates of human shapes were first generated on the basis of the PDM, and humans were detected by the template matching step. We took the ratio and distribution trend of foreground pixels inside each detecting window into consideration to reduce the detecting time complexity and improve the detection accuracy. A tracking method was further applied to deal with the short-term occlusions and noisy false alarms. In the experiment, we verified the proposed method in a crowded environment. The results show that comparing to method [7], the precision and recall of our method are improved (from 0.78% to 81%, and 83% to 87%, respectively). Moreover, the execution time of the proposed pedestrian detection method was almost the same as method [7] even though we had to additionally calculate the foreground distribution trend. In addition, we observed that the shadow of pedestrians would still be an issue that led to low precision rates even after applying the short-term tracking step. Hence, in the future we will apply shadow removal techniques to obtain more accurate foreground information. We will also design a more general method to determine the size of detecting window at each position in the frame.

ACKNOWLEDGEMENT

This work is partially supported in part by National Science Council of the Republic of China under Grant No. NSC101-2221-E-006-148-MY3, and Research Center for Energy Technology and Strategy, National Cheng Kung University.

REFERENCES

[1] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[2] R. Benenson, M. Omran, J. Hosang and B. Schiele, "Ten years of pedestrian detection, what have we learned?," in *Proceedings of the European Conference on Computer Vision, CVRSUAD Workshop*, Zurich, 2014, pp. 613–627.

[3] J. Yan, X. Zhang, Z. Lei, S. Liao and S. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, 2013, pp. 3033–3040.

[4] P.-T. Huang, Y.-M. Chan, L.-C. Fu, S.-S. Huang, P.-Y. Hsiao, W.-Y. Wu, C.-C. Lin, K.-C. Chang and P.-M. Hsu, "An image-range fusion pedestrian detection system in low illumination conditions," in

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Proceedings of the IPPR Conference on Computer Vision, Graphics, and Image Processing, Taiwan, 2014.

[5] W.-J. Liao, C.-W. Lin, C.-S. Chen and Y.-P. Hung, "Combining spatiotemporal background modeling and random forest classifier for foreground segmentation and shadow removal," in *Proceedings of the IPPR Conference on Computer Vision, Graphics, and Image Processing*, Taiwan, 2014.

[6] S. Tang, M. Andriluka, A. Milan, K. Schindler and S. R. a. B. Schiele, "Learning people detectors for tracking in crowded scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, NSW, 2013, pp. 1049–1056.

[7] M.-C. Hu, W.-H. Cheng, C.-S. Hu, J.-L. Wu and J.-W. Li, "Efficient human detection in crowded environment," *Multimedia Systems*, pp. 1–11, 2014.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models—their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[9] C. Beleznai and H. Bischof, "Fast human detection in crowded scenes by contour integration and local shape estimation," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 2246–2253.

[10] R. Fisher, "Caviar dataset," [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731