International Journal of Control and Automation
Vol.9, No.2 (2016), pp.239-248
http://dx.doi.org/10.14257/ijca.2016.9.2.23

An Ensemble Framework for Object Detection in Intelligent Video Surveillance System

Ning Sun*, Yuze Shan, Feng Jiang, Guang Han and Xiaofei Li

Engineering Research Center of Wideband Wireless Communication Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
**E-mail: sunning@njupt.edu.cn*

Abstract

In this paper, we present an ensemble framework with hierarchical and feedback mechanism for object detection. The proposed method is mainly composed of three phases: coarse detection, fine detection and tracking filter. In coarse detection, moving foreground can be rapidly extracted by improved ViBe background subtraction algorithm. FPDW as a fine detector scans the foreground image area, not entire the image, to determine the precise location and the number of targets. In the tracking filter, the detection results are processed to generate trajectories by the Kalman filter. And the current and the next status of the pixel is fed back from the former phases. For assessment of the effectiveness, we implement the proposed framework into pedestrian counting method. Several experiments are carried out based on the benchmark datasets. Experiment results show that the ensemble framework can achieve better detection results and real-time execution in comparison with other the state-of-art methods.

Keywords*: object detection; ensemble framework; background subtraction.*

1. Introduction

Intelligent Video Surveillance System (IVSS) [1] is able to actively notice and alarm the abnormal situations in the observing scenes. The smart ability of IVSS remarkably decreases the work intensity of monitor watchers and improves the effectiveness of traditional video surveillance system. The video analysis algorithms are the key to make the video surveillance system become so-called 'smart'. Generally, video analysis can be divided into three levels that is object detection, object recognition and object behavior understanding. Therefore, object detection is the primary step of most intelligent video analysis functions.

There are mainly two kinds of object detection application scenarios: a) object detection in stationary background, b) object detection in dynamic scene. In the first application scenarios, the implied constraint is that the interested target is moving in the video and the static parts belong to the background. Motion scene is usually due to camera movement, such as the video captured from the moving vehicle or aircraft. Since the almost cameras are fixed installed in the video surveillance system, detecting objects in stationary background is the main purpose of this paper.

In recent year, much research has been devoted to the field of object detection in stationary background. Background subtraction is most popular method to extract moving foreground when the observing scene of the camera is fixed. The earlier method is based on probability density function (PDF) like Gaussian mixture model (GMM) [2, 3], which is a parametric technique to model each pixel as Gaussian distribution, and non-parametric Kernel Density Estimation. Recent efforts include ViBe [4] and PBAS [5]. ViBe maintains a fixed number of samples for each pixel and classifies a new observation as background when it matches with a predefined number of samples.

ISSN: 2005-4297 IJCA

Generally, the background subtraction algorithm has the advantages of simple structure and a small amount of calculation, and it is easy to implement for real-time application. However, the problems of "ghost", shadow and sensitivity shake limit the object detection based on background subtraction to achieve higher performance.

Another effective framework for object detection is the sliding window paradigm [6-11], which is an exhaustive scanning mechanism to every layer of image pyramid. Some priori rules or trainable templates are performed to compute a similarity value on the sub image pixel by pixel through the entire image pyramid. Viola's [7] presents a robust and real-time face detector composed of a set of haar like feature trained by adaboost algorithm. From 2005, the HOG [8] based detector is plentifully proposed, and achieves the state of the art performance results on pedestrian detection. In addition, deformable parts model (DPM) [9] combined with several feature operator can perform better detection result in the crowd and occluded scenes. The sliding window methods provide the best detection accuracy, but it is not fast enough for real time application (the reported fastest method FPDW [10] is able to detect pedestrian at a speed of 6.5 fps).

In this paper, we present an ensemble framework (EF) to detect objects in the stationery background. There are three phases in the EF method, which are combined into one whole solution by hierarchical and feedback mechanism. The first phase is coarse detection, and improved ViBe is applied to rapidly extract the moving objects as foreground area and discard most image areas as background. We improve the update policy of ViBe algorithm based on the feedback information from the next two processing steps to reduce the reliability problems of foreground detection suffering from some shortcomings of classical ViBe. Secondly, fine detection, further identification is utilized to the foreground sub images by the more sophisticated trainable detector. This operation also can be accelerated up by feedback information. In the last phase, object trajectories are generated to smooth the current result of object detection and predict the object position in the next frame by the Kalman tracking filter. This predict information feeds back to the first two steps to improve the performance of detection.

2. Method

The ensemble framework consists of three phases: coarse detection, fine detection and tracking filter. As shown in Fig 1, the proposed method is a typical coarse-to-fine hierarchical structure with a feedback mechanism. The main idea of EF is that the foreground area of input video is quickly extracted by a simple approach at first. Then, fewer sub images are scanned carefully by more sophisticated and more complex method. Finally, The target prediction information produced by tracking filter feeds back to improve the detection reliability and execution speed of previous phases. The flow chart are described in the following.

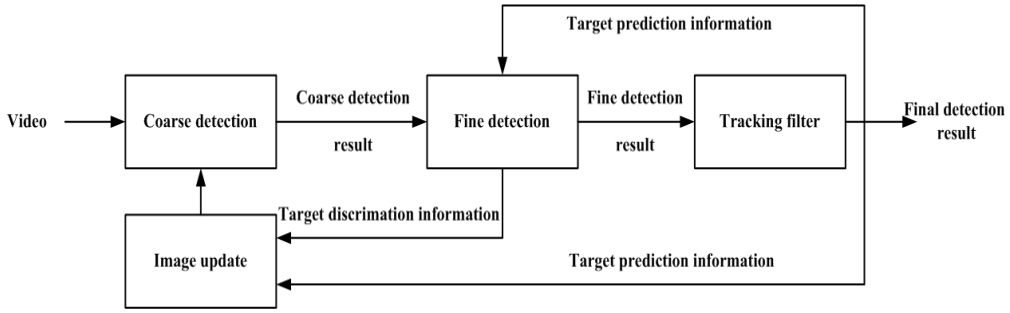

Figure 1. The Ensemble Framework Flow Chart

2.1 Coarse Detection

In this phase, we apply the improved ViBe based background subtraction as a coarse detection to quickly extract the moving foreground area. ViBe is a powerful and convenient pixel-based background subtraction technique. It has been proven to have much successful implementation in many real-time applications. However, ViBe algorithm often fails to obtain good results in certain circumstances, such as a) Stop target, that is the moving target stops in the field of view for a long time. b) Broken target, which means an entire target is split into several pieces because of some parts of target very similar to the background.

We make some improvement to ViBe algorithm according to the above-mentioned disadvantages. The two stages of the classical ViBe algorithm, pixel classification and pixel update, are processed separately in the proposed framework. The stage of pixel update is executed after the phase of fine detection and tracking filter rather than following the stage of pixel classification. As a result, the pixel update stage use the semantic information provided by fine detection and tracking filter to make the update processing faster and more reliable. Specifically, we add two variables U_c and U_p to every pixel in the image, which record the current and next status information of pixel from the tracking filter, respectively. To deal with the problem of stop target, we add a constrain to the neighboring pixel updating model of ViBe to stop the propagation of background pixel when the neighboring pixel is proved to be foreground base on U_c. Secondly, the value of the threshold $\#_{\min}$ is increased where the pixel is predicted to be foon inund pixel based on U_p in order to reduce the probability of occurrence of broken target.

2.2 Fine Detection

In this paper, the main purpose of fine detector is to determine the precise location and the number of targets in the foreground sub image area. We choose FPDW as a fine detector base on two facts. Firstly, FPDW is quite fast because of approximating multi-resolution image features via extrapolation from nearby scales. Secondly, FPDW can achieve top-class level detection rate in comparison with others the state-of-art pedestrian algorithms, even on crowed scene. With the help of hierarchical and feedback mechanism, FPDW only process the foreground sub image instead of exhausting scanning full-sized image task in every frame. In addition, with the support of feedback information, those targets with high quality trajectories are not necessary to be verified every frame.

2.3 Tracking Filter

Kalman filter [12] is chosen as a tracking filter for generating the target trajectories, which filters the current detection results and predicts the target parameters (position and size) in the next frame. A tracking quality variable Q_n is set to express the confidence level of every trajectory in EF method. The Q_n is obtained based on the correlation between the results of objection detection and target trajectories. The information of position prediction and tracking quality is respectively fed back to the previous phases to improve the performance of detection.

2.4 The Procedure of the Ensemble Framework

We outline the processing procedure of ensemble framework as below.

Input: Video sequence $I_m, m \in [t_i : t_j]$

Coarse detection:

If I_m is the first frame

Then build a sample set of every pixel $S_m(x, y), (x, y) \in I_m$ with size of $R + 2$ based on ViBe algorithm, where store R neighbor pixel value, U_c and U_p.

Else classify each pixel by threshold $\#_{\min}$, and segment the foreground image I_m^f after the morphology and segmentation processing.

Fine detection:

Obtain target number and location in the foreground area of I_m^f every P frames by FPDW. The P value is controlled by the tracking quality variable Q_m^l from tracking filter. Write the value U_c to sample set of each pixel.

Tracking filter:

Use Kalman fliter to generate and maintain the target trajectory and predict the target position in the next frame. Write the U_p to sample set of each pixel and send Q_m^l to fine detection step.

Image update:

If feedback information U_c U_p of each pixel is ready

Then update the each pixel sample set according to the update policy of ViBe algorithm

Except stop to propagate pixel values to neighbor proven foreground pixel. Increase the threshold $\#_{\min}$ of the pixel predicted belongs to the foreground area in the next frame.

Output: object detection result $R_m^q, m \in [t_i : t_j], q \in [0 : N]$

3. Experiments and Results

In this section, we exploit the experiment of pedestrian counting to evaluate the effectiveness of the proposed EF method. For comparison, the other three the state-of-art person detection algorithms are respectively integrated into the same processing procedure for pedestrian counting. The four pedestrian counting methods are identical except the stage of people detection.

3.1 Pedestrian Counting

We respectively integrate four object detection algorithms into the same processing procedure to build four pedestrian counting method. In Figure 2, these methods mostly composed of three stages: pedestrian detection, tracking filter and trajectories analysis. It is notable that we use kalman filter, the same algorithm as that in the ensemble framework, to generate target trajectories in tracking filter. In practice, the output of the EF method is already trajectory, which directly is send to the stage of trajectories analysis. And the output of the rest of three methods is sent to the stage of tracking filter for generating trajectories. For consistency, four pedestrian counting methods in Figure 2 all have three stages. In the trajectories analysis, the spatiotemporal information of pedestrian target recorded in the trajectories is analyzed to determine the direction of walking and calculate the number of passing.

The implementation of pedestrian counting methods is based on C++ and OpenCV library. In the main program, we call the Matlab function of DPM and FPDW detector by hybrid programming with C++ and Matlab. And all experiments run on the image processing server with Intel Xeon E7 4820 CPU and 32GB memory.

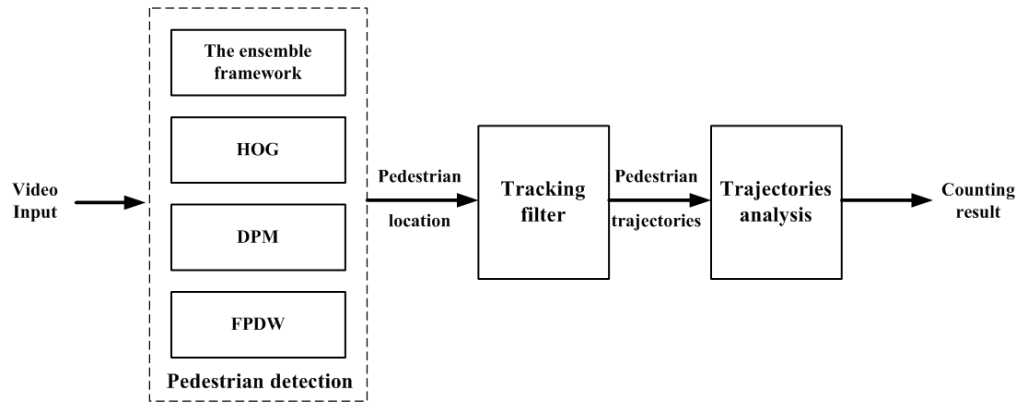

Figure 2. Pedestrian Counting Methods Flow Chart

3.2 Dataset

Three videos are used to assess these pedestrian counting methods in this section, as shown in Figure 3. The first video is 2371 frames image sequence named S3-T7-A View4 in PETS2006 [13]. This video contains low density crowd passerby and two people temporarily stay on the platform. The second one is a 241 frames image sequence containing medium density crowd scenario named S1-L1 View1in PETS2009 [14]. The last video is captured from the 720p HD camera installed to the campus gate of NJUPT, which is 4 minutes with low density crowd scenario containing pedestrian and vehicles.

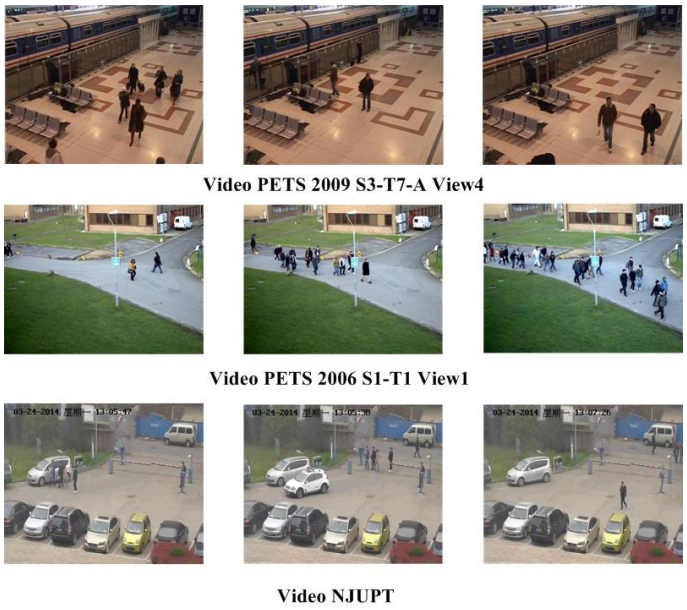

Figure 3. Dataset Used in the Experiments

3.2 Experimental Results and Discuss

In these experiments, we mainly compare the detection accuracy and execution speed of four methods. The indices used to express accuracy are the Mean Absolute Error (MAE) and the Mean Relative Error (MRE) [15] defined as:

$$MAE = \frac{1}{N}\sum_{I=1}^{N}\left|G(i) - T(i)\right| \qquad (1)$$

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|G(i)-T(i)|}{T(i)} \quad (2)$$

where N is the frame number of video, $G(i), T(i)$ are respectively the detected and true number of pedestrian in the i frame.

3.2.1 Experiment on Video Pets2006

In the Table 1, the detection performance of the EF method on the video PETS2006 is compared with that of the others three methods. Obviously, the MAE and MRE indices of the EF method are better than those of the rest. It is shown that the ensemble framework is able to effectively detect a specific kind of object in the low density crowd scene. The fact of the EF method is superior to FPDW in term of the detection accuracy, which is demonstrated that the hierarchical and feedback mechanism not only can keep the power of FPDW but also reduce the number of false positive. And the EF method is able to achieve real-time processing capacity on D1 resolution video. This speed outperformance is also owed to the hierarchical and feedback mechanism architecture. Most background areas are discarded in the phase of coarse detection, and the time-consuming fine detection is only performed in the foreground area. Meanwhile, the trajectories feedback information help speed up the procedure of fine detection.

Table 1. Experimental Result on Video PETS2006

	MAE	MRE	Speed(fps)
EF	0.179	7.5%	32.31
HOG	1.346	56.1%	4.83
DPM	0.24	10.3%	0.06
FPDW	0.22	9.1%	11.9

In video PETS2006, there are two people entering the scene and staying for a while. This is a typical situation of stop target. Figure 4 is the foreground image computed by classical ViBe and our improved ViBe in EF method. According to the update policy of classical ViBe, the static foreground is absorbed into the background, as the second column in Figure 4. In our ensemble framework, the update policy of ViBe algorithm receives the current and predicted status of pixels fed back from phase of fine detection and tracking filter. This information makes the improved ViBe to be updated based on the semantic content of pixels, not just rely on the similarity of pixels. The third column of Figure 4 clearly shows the advantage of improved ViBe in EF method.

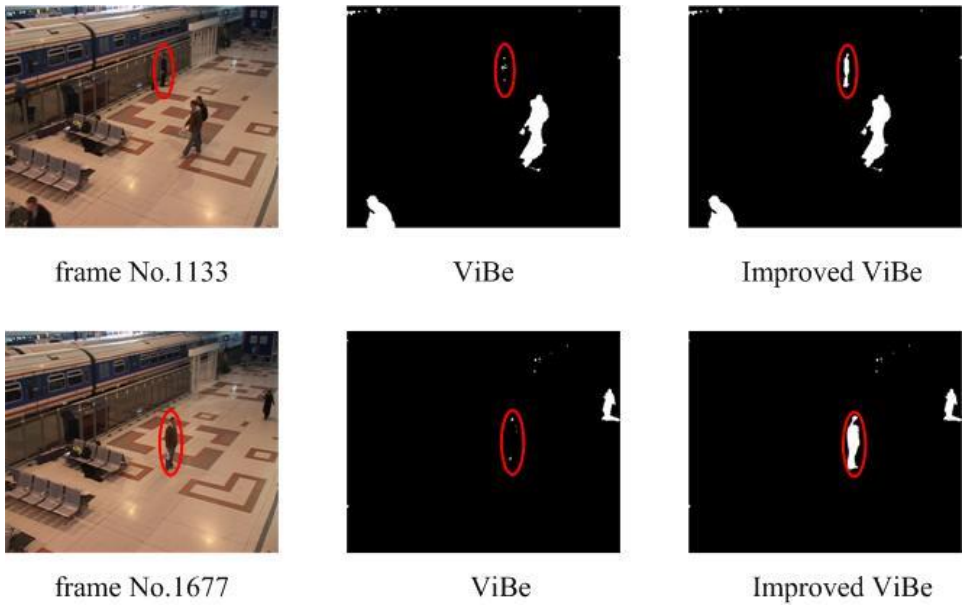

Figure 4. Comparison of Vibe and Improved Vibe on Ef Method

3.2.2 Experiment on Video Pets2009

Video PETS2009 is a 241 image sequence containing medium density crowd pedestrian. This video is used for assessing the detection performance on the crowd scene. From the experimental result shown in Table 2, we can draw the similar conclusions with those of experiments on video PETS2006. The proposed EF method still achieves the best detection accuracy and speed due to the hierarchical and feedback mechanism.

Table 2. Experimental Result on Video PETS2009

	MAE	MRE	Speed(fps)
EF	2.03	14.1%	22.21
HOG	4.99	36.7%	4.48
DPM	4.71	35.9%	0.06
FPDW	2.16	15.3%	9.35

In particular, Figure 5 displays the pedestrian number computed by the ensemble framework and other three methods in each frame together with the ground truth (GT). In Figure 5, it is found that FPDW is a powerful person detector. FPDW can obtain better detection rate in comparison with that of DPM based detector when more false positives are permitted. This result is consistent with the conclusions reported in the literature [16]. And the time cost of FPDW is far smaller than that of DPM. That is the reason why we choose FPDW algorithm to be as the fine detection in our EF method.

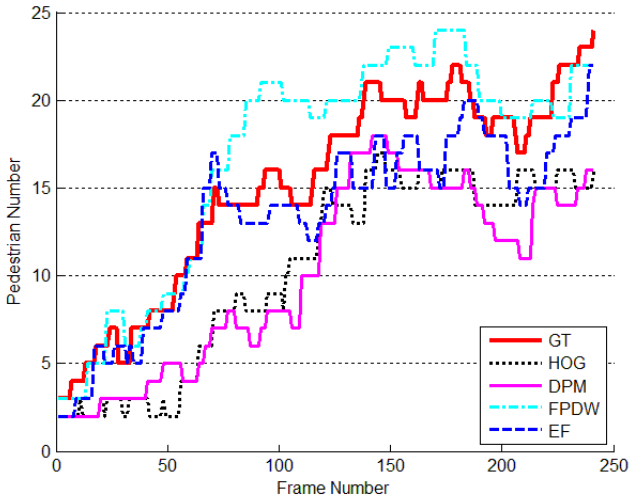

Figure 5. Comparison of the Pedestrian Number by Four Methods and Ground Truth

3.2.3 Experiments on Video Njupt

In this section, we use the four methods to count the pedestrian number of pass through one gate of our campus. This video named NJUPT is a about 6000 frames image sequence with 720*1280 high resolution (HD) recording the traffic of one of our campus gates. The ground truth of this video is 67 people coming in and 26 ones getting out.

Table 3. Experimental Result on Video NJUPT

	In	Out	Speed (fps)
EF	64	26	11.27
HOG	72	30	1.95
DPM	56	26	0.02
FPDW	73	31	4.51

In Table 3, It is indicated again that the EF method obtains the best result in comparison with the other three methods. The average processing speed of EF is 11.27 fps according to the image with size of 720*1280, which shows that of EF method is suitable for real-time HD video analysis applications.

4. Conclusions

In this paper, we propose an ensemble framework for object detection in stationary background. The ensemble framework composed of three phases, coarse detection, fine detection and tracking filter. These phases are combined into an entire object detection solution by mechanism of hierarchical and feedback. And several improvements are made for faster execution speed and more reliable detection result. We design the experiments of counting pedestrian to assess the performance of the proposed method. The experimental results compared with that of three methods based on HOG, DPM and FPDW show that the ensemble framework can efficiently detect pedestrian under crowed environment and be suitable for real-time applications.

Acknowledgments

This work was supported by the National Nature Science Foundation of China(61471206) and Natural Science Foundation of Jiangsu province(BK20141428).

References

[1] V. Gouaillier, A.E. Fleurant, "Intelligent video surveillance: Promises and challenges" Technological and commercial intelligence report, CRIM and Technopole Defense and Security, **(2009)**.
[2] C. Wren, A. Azarhayejani, T. Darrell and A.P. Pentland, "Pfinder: real-time tracking of the human body", IEEE Trans. on Pattern Analysis and Machine Intell., vol. 19, no. 7, **(1997),** pp. 780-785.
[3] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking", In CVPR 1999, vol. 2, **(1999)**, pp. 252.
[4] Barnich and M.V. Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences. Image Processing", IEEE Transactions, vol. 20, no. 6, **(2011)**, pp. 1709–1724.
[5] M. Hofmann, P. Tiefenbacher and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmentation", In CVPR Workshops 2012, **(2012),** pp. 38-43.
[6] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection", Int'l J. Computer Vision, vol. 38, no. 1, **(2000)**, pp. 15-33.
[7] P.A. Viola and M.J. Jones, "Robust Real-Time Face Detection", Int'l J. Computer Vision, vol. 57, no. 2, **(2004)**, pp. 137-154,.
[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in Proc. IEEE Conf. CVPR, **(2005)**.
[9] P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models", IEEE Trans. PAMI, vol. 32, no. 9, **(2010)**, pp. 1627-1645.
[10] P. Dolla´ r, S. Belongie and P. Perona, "The Fastest Pedestrian Detector in the West", Proc. British Machine Vision Conf., **(2010)**.
[11] Y. Liu, S. Shan, W. Zhang, X. Chen and W. Gao, "Granularity-Tunable Gradients Partition Descriptors for Human Detection", Proc. IEEE Conf. CVPR, **(2009)**.
[12] X. Wang, "Intelligent Multi-Camera Video Surveillance: A Review", Pattern Recognition Letters, vol. 34, **(2013)**, pp. 3-19.
[13] ftp://ftp.pets.rdg.ac.uk/pub/PETS2006.
[14] ftp://ftp.pets.rdg.ac.uk/pub/PETS2009.
[15] D. Conte, P. Foggia, G. Percannella, F. Tufano and M. Vento, "A method for counting people in crowded scenes", in IEEE International Conference on AVSS, **(2010)**.
[16] P. Doll_ar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 4, **(2012)**, pp. 743–761.

