Improved Post-Processing for Human Detection in Railroad Surveillance

Xueying Xin, Zhenhua Guo, Bo Yuan
Graduate School at Shenzhen, Tsinghua University,
Shenzhen, P.R. China
xinxuey@163.com, cszguo@gmail.com, yuanb@sz.tsinghua.edu.cn

Jie Zhou
Department of Automation, Tsinghua University,
Beijing, P.R. China
jzhou@tsinghua.edu.cn

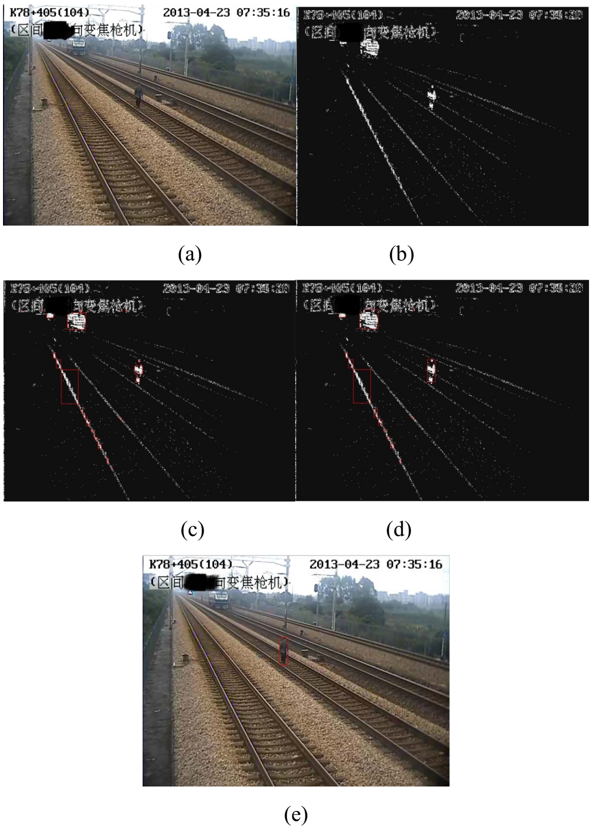
***Abstract*—Foreground detection (FD) has been widely used for moving object detection such as human detection. The practical performance of FD methods varies significantly in different real-world environments. Post-processing methods can improve the effectiveness of FD algorithms by providing high quality foreground masks for further detection. The accuracy of human detection in railroad surveillance is often limited due to occlusion and noise, causing incomplete objects in foreground masks. In this work, we introduce a novel integration step into the post-processing module following FD. We take *priori* knowledge into account and apply specific rules when combining the labeled components, resulting in significant improvement in the accuracy of human detection.**

***Keywords*—foreground detection; post-processing; railroad surveillance**

I. INTRODUCTION

Video surveillance has been gaining more and more popularity in our daily life and providing people with essential assurance of personal and property safety. In transportation industry such as airport or railroad systems, human or animals are prevented from approaching certain regions[1]. Generally, there are three modules included in a classic intelligent video surveillance system[2]: Object Detection, Object Tracking and Recognition, and Motion Analysis. Object Detection module detects moving objects of interest whose trajectories are then tracked by the Objection Tracking module. After that, the recognition process determines the nature of the object (e.g., a car or a human being). Finally, the motion analysis is implemented based on the specific application.

Foreground detection (FD) is an effective tool used in the Object Detection module. It aims to detect foreground objects from the static background. Many foreground detection methods, such as Background Subtraction (BS) have been proposed recently[3][4][5][6]. BS is a popular method for foreground detection and an important component in many vision-based applications, especially visual surveillance[7]. A foreground mask is obtained after the foreground detection module with 1 for the foreground and 0 for the background. Then the specific position for a connected area in a foreground mask is needed to give the initial position for tracking each segmented object. The accuracy and integrity can directly affect the effectiveness of tracking. In real world, harsh environments and low-quality cameras can make the foreground masks noisy, as shown in Fig. 1a and Fig. 1b. Noises include swaying trees, reflection rails and uninteresting moving objects such as trains. These noises can make the foreground masks objects incomplete, as shown in Fig. 1c. Few existing studies have put efforts in solving this problem. In this work, we propose a novel integration step into the post-processing module following FD. We take *priori* knowledge into account and apply specific rules when combining the labeled components, making a significant improvement in the accuracy of human detection.



(a) (b)

(c) (d)

(e)

Figure 1: (a) Original frame from video; (b) Foreground mask produced by VIBE[8]; (c) Output (marked with red rectangle) from original post-processing operation; (d) Output (marked with red

2165-3828/14 $31.00 © 2014 IEEE
DOI 10.1109/ICSS.2014.11

rectangle) from improved post-processing operation; (e) Ground truth.

This paper is organized as follows. In Section 2, we introduce related work on post-processing for foreground detection. Section 3 gives the details on the two-stage post-processing module. Experimental results are reported in Section 4 and this paper is concluded in Section 5.

II. RELATED WORK

The quality of foreground masks depends largely on the application scenarios. A perfect FD algorithm is supposed to work robustly under illumination change, slowing-moving background, slight camera shake and so on. However, most FD algorithms only manage to deal with some but not all of the problems. Post-processing aims to compensate for the deficiency and improve the binary foreground masks. Generally, post-processing is divided into two types of operations: pixel-level operation that aims at removing noise and object-level operation that aims at labeling connected component and providing their position for further tracking.

For noise removal, different noise filtering algorithms are applied, such as median filtering or average filtering. For object-level operation, connected components are labeled. In an ordinary case, a connected component is found under the condition of 8-adjacent image segmentation.

Previous studies did achieve certain improvements on post-processing. Brutzer[9] evaluated nine BS methods with post-processing, which included median filtering, morphological operations (opening, closing, and their combinations). It was shown that post-processing worked effectively to increase the performance of BS methods, especially in elimination the gap between good and bad BS algorithms. Parks and Fels[10] also considered several post-processing techniques for state-of-the-art BS algorithms. There were five components in their post-processing framework: Noise removal, Blob processing, Saliency test, Optical flow test and Object-level feedback.

Shadow removal is also included in the domain of post-processing. Shadows misclassified as foreground objects can cause errors in object recognition and tracking. Andrea Prati et al.[11] organized various shadow removal algorithms in a two-layer taxonomy, the layer of deterministic approaches and the layer of statistical approaches. A comparison of four representative algorithms in the two categories was conducted. Lei[12] proposed an effective two-layer shadow removal scheme based on brightness distortion in different color channels, blob-based object tracking and occlusion handling.

However, few studies focus on the improvement of the connected component labeling process and take into account the correlation of the segmented component. However, combination of the connected components, as we are to show in the following sections, is helpful for improving the quality of foreground masks and further improving the accuracy of human detection.

III. IMPROVED POST-PROCESSING

A. System Framework

Our improvement of post-processing for foreground detection is specially designed for a railroad surveillance system. A normal railroad surveillance system aims at detecting human beings who appear in the surveillance area and further recognizing their action.

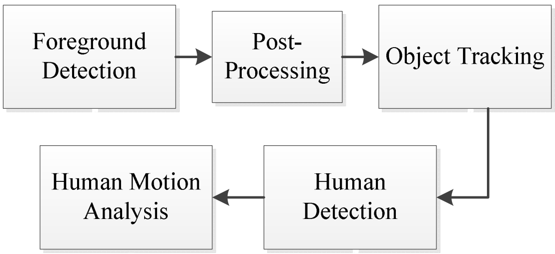


Figure 2: The framework of our human detection system

The general process of human detection is shown in Fig. 2. Foreground Detection module detects every suspected object basically based on their movements, providing foreground mask, as shown in Fig. 1b. Post-processing module labels the connected components in foreground masks, providing location, as shown in Fig. 1c and Fig. 1d. Object Tracking module aims to accomplishing a real-time tracking on the objects detected. Finally, it is Human Detection module that tells the system whether it tracks a person or not and decides to further trigger an alarm. Human Motion analysis is a necessary module for an intelligent surveillance system..

B. Post-Processing Structure

We divide our post-processing process into four steps, as shown by Fig. 3.

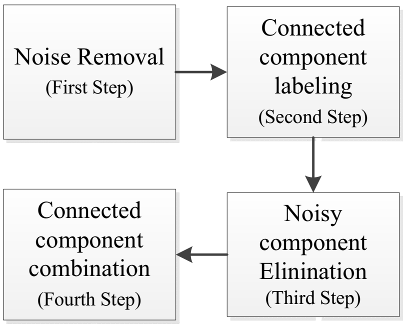

Noise Removal
(First Step)
Connected component labeling
(Second Step)
Noisy component Elinination
(Third Step)
Connected component combination
(Fourth Step)

Figure 3: The proposed framework of post-processing

C. Noise Removal

Preprocessing is always important when dealing with a noisy picture or video. Usually, median filter is used to reduce salt & pepper noise, and the mean filter is to reduce Gaussian noise. In our task, noise removal is done following the connected component labeling step. We choose mean filter with a rectangular template of size 5×5 because there are random noise mostly, close to Gaussian noise. Mean filter could eliminate certain random noise and maintain original information.

D. Noisy Component Elimination

Considering a very noisy foreground mask, Single method is obviously not enough. So after the connected component labeling process, the number of minimum bounding rectangles is far more than we have expected. We use certain means to cut the number of rectangles because it affects the speed of following steps.

A connected component, labeled by a minimum bounding rectangle, which is represented as:

(X, Y, width, height), “X, Y” is the position of the up-left point of a rectangle; “width” and “height” are the size of a rectangle.

A rectangle is removed when it meets at least one of three conditions:

$$width * height > 20000 \ or \ width * height < 16 \tag{1}$$

$$\begin{aligned} &width > frame\_width / 5 \ or \\ &height > frame\_height / 2.5 \end{aligned} \tag{2}$$

$$Y < 200 \ \& \ width * height > 16000 \tag{3}$$

Explanations are as follows:

(1) Rectangles that are too big or too small to be considered as a part or an object are removed;

(2) Rectangles whose size exceeds a real person of the world are removed;

(3) Cameras are put high above the rail road. View from them is somewhat parallel with the tracks in most cases. Under this circumstance, objects appear in the upper position in a frame are actually far from the viewpoint. Conversely, objects appear in lower position are much closer from the viewpoint. Accordingly, big rectangles in the upper position are deleted.

E. Connected Component Combination

According to the prior knowledge, the split parts which belong to the same object are physically close. So we list all the minimum bounding rectangles, and compare the horizontal and vertical distances of every two of rectangles. If a pair whose distance is under a certain threshold, these rectangles are chosen to be combined.

However, not all the rectangles which are physically close come from the same object because of large amount of noise. The combination of rectangles is canceled under three conditions:

$$\begin{aligned} &width > (frame\_width) / 5 \ or \\ &height > (frame\_height) / 2.5 \ or \\ &width * height > 20000 \end{aligned} \tag{4}$$

$$width / height > 6 \ or \ width / height < 1 \tag{5}$$

$$\sum_{\substack{x \in [X, X+width] \\ y \in [Y, Y+height]}} p(x, y) < 0.32 * width * height \tag{6}$$

Explanations are as follows:

(1) If the width, the height or the whole area of the combined rectangle reaches above a certain threshold, the combination is canceled;

(2) The aspect ratio should be in a certain range for a man would never to be too “fat” or too “thin”;

(3) The ratio of foreground points in the combined rectangles should be beyond a certain value. It can avoid a circumstance that a small part of an object is linked to a big part of another object.

Using the three standards discussed above, most mis-combined rectangles can be avoided. Several instructions need to be clarified on how we set the standards.

In the third step before combination, we are careful on eliminating the small-size rectangle for it may be a small part split from a big object.

In the fourth step, to speed up computation speed, we compute horizontal and vertical distances by simple operation like plus or minus instead of Euclidean Distance that takes more time.

IV. EXPERIMENT RESULTS

To testify the rationality of our improvement, we present two experimental results, one gives an intuitive proof of the improvement and the other provides quantitative results.

A. Ground Truth Test

In the first method, we use ground truth (GT) as our reference standard. GT data here is the manly drawn minimum bounding rectangles, as shown in Fig. 1e. In most cases, generating GT data needs huge effort and is sometimes time-consuming. There exist several techniques that overcome the difficulties involved in manual GT annotation, such as works depend on GT data only[13][14] or automatically generate them[15].

In our experiment, we choose 30 to 40 samples from 3 test set randomly. Ground Truth rectangle of each frame is drawn manly and the overlap rate of the rectangle between the original/ improved method and the GT is computed and compared. Result of test 1 is shown in figure 3. Through the comparison, we can tell that the improved method gives a better location of the interested objects both in fitness and in integrity.

Overlap rate is computed through:

(“O” stands for overlap rate between two bounding rectangles; “S” stands for area of a rectangle.)

$$O_{original} = (S_{\text{original method}} \cap S_{GT}) / S_{GT} \tag{7}$$

$$O_{improved} = (S_{\text{improved method}} \cap S_{GT}) / S_{GT} \tag{8}$$

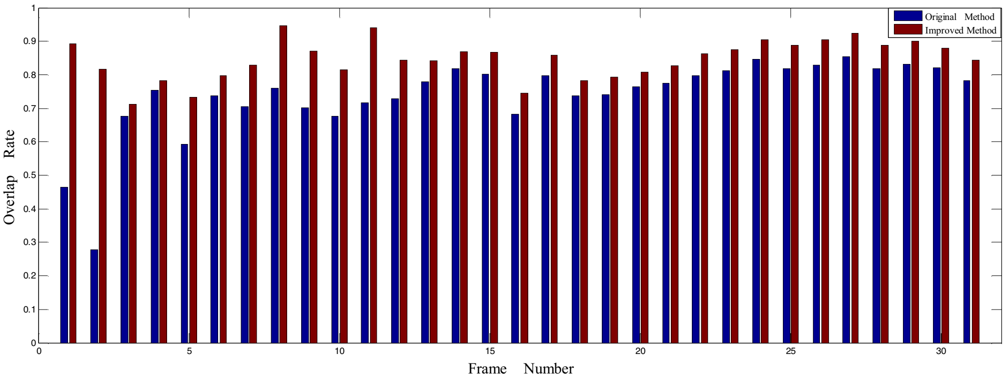
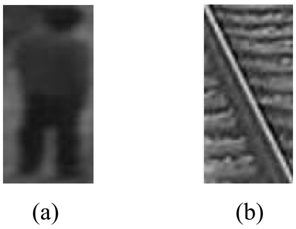


Figure 4: The horizontal axis shows the number of the sample in test set 1. The vertical axis shows the overlap rate, ranging from 0 to 1. Blue bar represents the overlap rate of minimum bounding rectangles between the original method and the GT. Red bar represents the overlap rate between the improved method and the GT.

B. *Classifier Accuracy Test*

Minimum bounding rectangles are input of human detection module. So more appropriate rectangles help promote the accuracy of human detection. In our experiments, human detection is accomplished by using feature HOG[16]\LBP[17] and classifier SVM[18]. We select positive and negative samples manually, and use part of the samples to train the SVM and the remaining samples to test the classifier. For the original method, we use 600 positive samples and 2500 negative samples to train a SVM classifier. Then test it on 5 positive sample sets and 5 negative sample sets respectively. For the improved method, we also select 600 positive samples and 2500 negative samples to train a SVM classifier. And test it on 5 positive sample sets and 5 negative sample sets respectively. Table 1 shows the result of comparison. The result shows that the improved method surpasses the original one significantly.

Fig. 5 gives examples of a positive sample and a negative sample (both are normalized to 128×64):



(a) (b)

Figure 5: (a) A positive with image size 128×64; (b) A negative sample with image size 128×64.

Table 1: Comparison of accuracy with SVM

	Positive Samples		**Negative Samples**	
	Mean	**Variance**	**Mean**	**Variance**
Original Method	94.80%	4.65E-5	90.90%	1.5×10E-3
Improved Method	99.80%	8.0E-7	96.80%	1.75×10E-5

V. CONCLUSION

In this paper, we proposed a series of post-processing operations for improving the quality of foreground masks and accuracy of human detection. We designed our post-processing module by considering the issue that one object can be split to several blobs because of poor foreground masks, which is common in the environment of railroad surveillance. A step was proposed after noise removal and connection component labeling operation. We also designed rules for combining the labeled components. Experiments results showed that our method can make a big difference in the localization of interested objects and improving the accuracy of human detection.

As to future work, we will make efforts in removing noises such as reflection rails, and make the foreground masks cleaner for human detection.

REFERENCE

[1] Zhong H, Shi J, Visontai M. Detecting unusual activity in video. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, 2: II-819-II-826 Vol. 2.

[2] Lee L, Romano R, Stein G. Introduction to the special section on video surveillance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 745.

[3] Piccardi M. Background subtraction techniques: a review. Systems, man and cybernetics, 2004 IEEE international conference on. IEEE, 2004, 4: 3099-3104.

[4] Bouwmans T. Subspace learning for background modeling: A survey. Recent Patent On Computer Science, 2009, 2(3): 223-234.

[5] Elhabian S Y, El-Sayed K M, Ahmed S H. Moving object detection in spatial domain using background removal techniques-state-of-art. Recent patents on computer science, 2008, 1(1): 32-54.

[6] Bouwmans T, El Baf F, Vachon B. Statistical background modeling for foreground detection: A survey. Handbook of Pattern Recognition and Computer Vision, 2010: 181-199.

[7] Yoo S, Kim C. Background subtraction using hybrid feature coding in the bag-of-features framework. Pattern Recognition Letters, 2013, 34(16): 2086-2093.

[8] Barnich O, Van Droogenbroeck M. ViBe: A universal background subtraction algorithm for video sequences. Image Processing, IEEE Transactions on, 2011, 20(6): 1709-1724.

[9] Brutzer S, Hoferlin B, Heidemann G. Evaluation of background subtraction techniques for video surveillance. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 1937-1944.

[10] Parks D H, Fels S S. Evaluation of background subtraction algorithms with post-processing. Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on. IEEE, 2008: 192-199.

[11] Prati A, Mikic I, Trivedi M M, et al. Detecting moving shadows: algorithms and evaluation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003, 25(7): 918-923.

[12] Lei B, Xu L Q. Real-time outdoor video surveillance with robust foreground extraction and object tracking via multi-state transition management. Pattern Recognition Letters, 2006, 27(15): 1816-1825.

[13] Chalidabhongse T H, Kim K, Harwood D, et al. A perturbation method for evaluating background subtraction algorithms. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Nice, France. 2003.

[14] Erdem C E, Murat Tekalp A, Sankur B. Metrics for performance evaluation of video object segmentation and tracking without ground-truth. Image Processing, 2001. Proceedings. 2001 International Conference on. IEEE, 2001, 2: 69-72.

[15] Grossmann E, Kale A A, Jaynes C O, et al. Offline Generation of High Quality Background Subtraction Data. BMVC. 2005.

[16] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance. Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006: 428-441.

[17] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(7): 971-987.

[18] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.