A New Local-Main-Gradient-Orientation HOG and Contour Differences based Algorithm for Object Classification

Xiaoqiong Su[1], Weiyao Lin[1*], Xiaozhen Zheng[2], Xintong Han[1], Hang Chu[1], Xiaoyun Zhang[1]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University, China ([*] Corresponding author)
[2]Research Department of Hisilicon Semiconductor and Component Business Department, Huawei Technologies, China

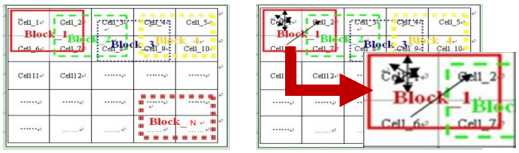
***Abstract*—This paper presents a new algorithm to better classify objects in videos. In our case, the objects are cars, vans, and people on the roads. First, in order to extract the moving objects more precisely, we have proposed a method for foreground extraction based on the contour differences between the video frame and the background image. Second, after we got the integrated moving object, we have proposed a new algorithm to extract better features from the object. The new algorithm is based on two extended Histogram of Oriented Gradient (HOG) descriptor. We have improved HOG in two aspects: (a) selecting the gradient information from the moving objects and discarding the background gradient; (b) weighting every bin of gradient orientation histogram according to their significance within predefined area, in order to emphasize the important gradient information. We obtained Contour-Difference HOG (CD-HOG) from the first extension and Local-Main-Gradient-Orientation HOG (LMGO-HOG) from the second extended HOG. These extensions can cope with the cluttered background and make the features more distinguishable. Each of the extended HOG descriptors can produce a satisfying performance separately and an even better one if they are applied in cascade. From extensive evaluations, we showed the wonderful performance of our algorithm, and the accuracy rate of 94.04% can be achieved in some cases.**

I. INTRODUCTION AND RELATED WORKS

Automobile classification aims at automatically telling the types of the cars in videos. This subject is important for the smart traffic systems, as it helps to solve the charging problems and increases the utility of roads. However, the task is challenging for the following two main reasons: (a) the poor performances of foreground extraction methods; (b) the difficulty in finding a robust feature to describe moving objects better under the circumstance of cluttered background. Some previous works have been done to solve these problems, but the challenges are still apparent:

(I) Many approaches[1][2] for foreground extraction have been developed. However, till now the up-to-date methods are still prone to degenerate their performances considerably due to different variation factors, such as camouflage, shadow effect and lighting variation. Therefore, the extracted foreground images are often incomplete or incorrect. For instance, if the frame difference method is being used, the moving object contour is larger than the real one in its driving direction. When car stops to wait for the traffic light, the foreground is faded away. If the background modeling method is used, the illumination changes can lead to huge difference between the gray scales of the same object at different frames. Mixture-of-Gaussian model (GMM)[3][4], one of the widely used methods, however, has the problem of introducing much noise when a large moving object shows up.

(II) The most challenging process for object classification is the feature extraction. Dalal[5] came up with a descriptor



(a) (b)

Fig. 1 (a) The common concept for HOG descriptor, (b) The Local-Main-Gradient-Orientation HOG descriptor.

HOG, and it soon became an effective method for describing objects, especially people. HOG descriptor is based on the gradient information provided by small scale cells. The cells are the fundamental elements which are included in relatively large scale blocks, as shown in Fig. 1 (a). However only the gradient information within cells is included, thus the larger scale gradient is omitted, which is essential information for rigid objects like cars. In addition, the cluttered background of training and testing images can bring much noise while describing the central objects with HOG. In order to solve the problem of cluttered background, some other methods[6] have been proposed, however, these methods either have poor performances or are difficult to implement. For example, the paper [6] proposed a method which is hard to implement due to the difficulty in cutting the moving object out manually. Other descriptor like Scale-invariant feature transform(Sift) is only good at recognizing the same object at different observing angles. As for multiple-scaled-Harr-like(MSHL) descriptor[7], which is recently presented, it discards all the detailed gradient information, and this is not appreciated.

In order to solve the problems of previous works, we proposed a new foreground extraction method and a new algorithm based on the extensions of HOG. The major contributions of this paper are listed below:

(a) We have extracted moving objects with more precise location, and the method has lower sensitivity to illumination changes. We have also extracted the still objects, and solved the problem of large amount of noise introduced by large moving objects;

(b) We have proposed the Contour-Difference HOG (CD-HOG) to describe the moving objects purely by getting rid of the background gradient information;

(c) We have proposed the Local-Main-Gradient-Orientation HOG (LMGO-HOG) as a better descriptor for classification by emphasizing the important features.

The rest of the paper is organized as follows: Section II discusses the overview of our algorithm. The details of the proposed algorithm are presented in Section III. Section IV shows the experimental results.

II. OVERVIEW OF THE PROPOSED ALGORITHM

For object classification in videos, the following steps

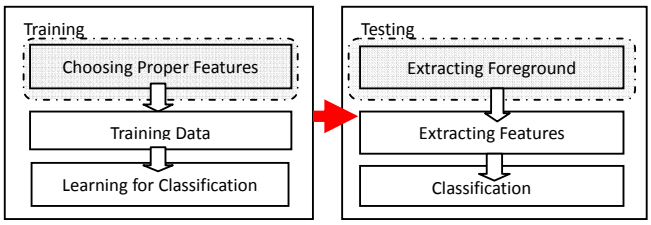


Figure 2 The framework instruction

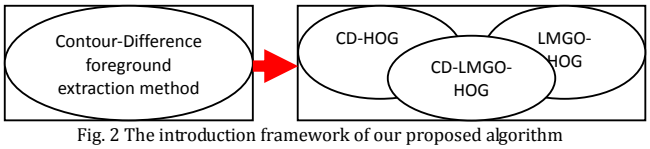


Fig. 2 The introduction framework of our proposed algorithm

are needed to be fulfilled: (a) choosing proper features; (b) training samples; (c) learning for classification; (d) extracting moving objects from the videos; and (e) classifying the subject. Fig. 2 shows the flow of the framework, in which the shaded areas are the principal parts. We will introduce our proposed algorithm based on Fig. 3.

First, as for extracting foreground, we observed that the gradient differences between the original video scene and the background scene could be used to describe the contours of the moving objects better.

Second, as for feature extraction, we developed a new algorithm based on two extensions of HOG descriptor for obtaining a better descriptor.

(I) We proposed the Contour-Difference HOG (CD-HOG) by selecting the gradient information from central objects only and discarding the gradient of background.

(II) We noticed that every object, especially cars, has local main gradient orientation, like the linear outline of the car; these orientations can be obtained in relatively large scale portion of an image. We incorporated the local main gradient information by weighting every bin, which is used to collect the information of gradient direction and value, in every cell within this portion; the basic principle is shown in Fig.1 (b) (the portion here is chosen to be a block). Those bins that share the similar direction with the local main gradient direction own higher weight. We named our method as Local-Main-Gradient-Orientation HOG (LMGO-HOG).

(III) We combined these two features together to form a complete algorithm for a better descriptor.

In the following section, we will provide detailed information of the proposed algorithm.

III. THE DETAILED ALGORITHM DESCRIPTION

As mentioned above, we will first focus on the new method for foreground extraction, and then introduce our new algorithm for better descriptor at the second section.

A. The Contour-Difference foreground extraction

The Contour-Difference foreground extraction (CDFE) method procedures are shown in Fig. 3.

First, we tried to recover a clean background image. Background extraction algorithms typically use techniques like image inpainting[8] on a single image or object-tracking techniques[9][10] on a sequence of images or a combination of both[11]. Here we used the Running Average background modeling method to set a clean street model shown in Fig. 3 (b), which had no moving object intrudes, by selecting a proper speed to update the background image.

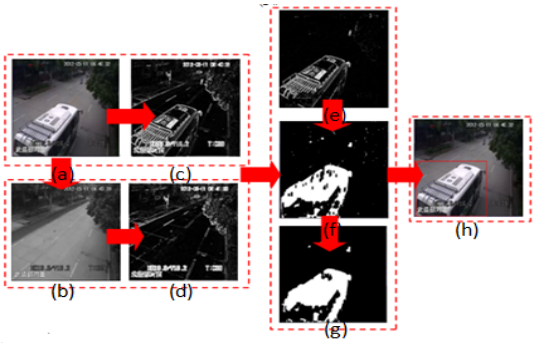


Fig. 3 The CDFE procedure. (a) The original video image. (b) The background image. (c), (d) The contour information for (a) and (b) respectively. (e) The subtraction result. (f), (g) The opening and closing operations. (h) The final result.

Next, we used Eq. (1) to obtain the contour for moving object, where C_{diff} is the contour for central moving object only. C_1 and C_2 are the contours for the original frame and the background image respectively. $p_1(i,j)$ and $p_2(i,j)$ are the points on C_1 and C_2 at the same spot. T denotes the threshold, which is set to tolerate the grayscale difference caused by changing of illumination. The result is shown in Fig.3 (e).

$$C_{diff}(p(i,j)) = \begin{cases} C_1(p_1(i,j)), & (\| C_1(p_1(i,j)) - C_2(p_2(i,j)) | > T) \\ 0, & (\| C_1(p_1(i,j)) - C_2(p_2(i,j)) | < T) \end{cases} \quad (1)$$

We tested many operators to check their performances in procedure from Fig. 3 (a) to Fig. 3 (c), including Sobel, Prewitt, and Canny operators. We observed that Sobel operator with adjustable kernel achieves the best result in obtaining enough information for a contour. In some videos with low resolution, we applied histogram equalization to improve the quality of the images. When a video with large cars, like buses and big vans, were concerned, fragmented contour problem could be solved by increasing the kernel's size of Sobel operator or decreasing T to preserve more information. Moreover, we used LK optical flow to determine the direction of the moving objects, and adjusted opening and closing operations corresponding to it.

Through the aforementioned processes, we could extract a perfect moving object, as shown in Fig. 3 (h).

B. CD-HOG AND LMGO-HOG

After we extracted the moving objects with our CDFE method, we proposed a new algorithm to obtain a better descriptor.

(i) CD-HOG

Most Surveillance cameras for traffic are installed on the crossroads where, in general, the backgrounds, like zebra crossings, have more gradient information which we don't want. So they introduce much noise when we describe a moving object with HOG descriptor. And another problem the cluttered background brings us is that when we are dealing with different videos, the training samples from one video are not able to be used to identify the moving objects' type in other videos due to differences in background.

In order to solve the cluttered background problem, we proposed the Contour-Difference HOG to eliminate the background interference.

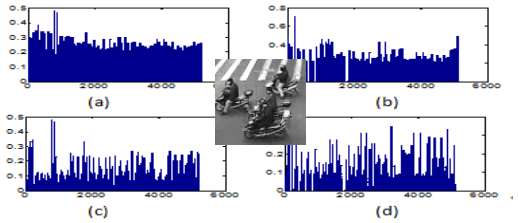


Fig. 4 Four features for the same moving object. (a) The original HOG feature. (b) The CD-HOG feature. (c) The LMGO-HOG feature. (d) The CD-LMGO-HOG feature.

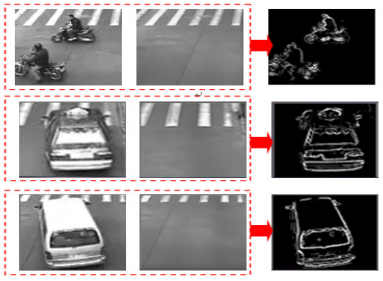


Fig. 5 The result of background elimination

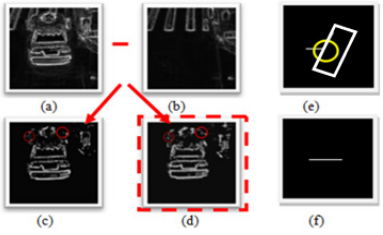


Fig. 6 (a) The contour of video image. (b) The contour of background image. (c) The result of CD-HOG descriptor without preserving good points. (d) The result of improved CD-HOG. (e), (f) The concept of good points.

The process is similar to that of CDFE method in previous section. After we got the portion containing moving objects from the original video, we calculated the gradients for both this portion and the corresponding portion on the background image. And we compared the differences between these two portions in cells. Applying comparing rule in Eq. (1), we got CD-HOG. The Fig. 4 (a) and (b) demonstrate the change in histogram of orientation gradient features before and after eliminating the background. The performance gain brought by eliminating background redundancy is obvious. Fig. 5 shows some results of the objects' contours after eliminating the background, and these contours are the information that CD-HOG descriptor describes.

However this single method may bring a disadvantage: some good points are removed. "Good points" refer to some points on the moving object, whose gray scale is similar to the corresponding point on background image, so these good points are easily to be removed. Fig. 6 (e) and (f) demonstrate the concept of good points. Fig. 6 (f) is the background scene and Fig. 6 (e) demonstrates the video image; the white rectangle is the moving object, the horizontal line belongs to the cluttered background. The crossing point of the horizontal line on the background and the oblique line on the moving object is the good point.

In a video with a lot of good points, the moving object is intermittent after calculating its CD-HOG. So we came up with an easy solution to preserve these good points. We compared every pixel in video frame and background image not only by their gray scales but also by their four adjacent pixels' gray scales, and if all of them satisfy the threshold standard (the difference of each pair is below T), this point is removed. If not, this point is defined as a good point, and is preserved. Fig. 6 (a), (b), (c), and (d) show the effect of preserving good points. The red circles in two images clearly show the differences before and after preserving the good points.

(ii) LMGO-HOG

In this section, we will introduce the LMGO-HOG descriptor. The HOG is a descriptor counts occurrences of gradient orientation in localized portion of an image: cell. The gradient information of the whole image is based on the pile of cells, so the local main gradient in large portion is omitted. In our case, car is a rigid body that consists of obvious lines, and the gradient in large portion is important as the Fig.8 declares. Another kind of descriptor, the MSHL descriptor uses a varying-size window to scan the whole image, and calculates the local main gradient direction, in order to obtain the gradient information more precisely. However, small gradient information is completely discarded. So we chose to combine these two descriptors together to throw away the dross, select the essential, and form a better descriptor: Local-Main-Gradient-Orientation HOG.

The cell collects the detailed gradient information, so we tried to preserve this information. And we used a predefined window, whose size is suitable to describe the partial basic structure of the moving object, to scan the entire image, as Fig.8 shows. This window could be adjusted according to different circumstances. A portion of the image could be obtained during every move of the window; we calculated the local main gradient direction of the portion. Every cell's gradient bins within this portion were weighted, and the weights were determined by the difference between the local main gradient orientation and the direction that the bin represented. The principle scheme is shown in Fig. 1 (b). Eq. (2) and Eq. (3) show the way to obtain LMGO-HOG.

$$v_{lmgohog}(i) = v_{hog}(i) \times w(i) \tag{2}$$

$$w = \frac{1}{|O_{main} - o_{bin}| + 1} \tag{3}$$

where, $v_{lmgohog}(i)$ denotes the i-th element of LMGO-HOG descriptor vector; $v_{hog}(i)$ denotes the i-th element of HOG descriptor vector; $w(i)$ is the weight of $v_{hog}(i)$. w *is* obtained through Eq. (3). O_{mian} denotes the local main gradient orientation. o_{bin} is the direction of the individual bin. The smaller the difference of O_{mian} and o_{bin} is, the larger w is.

Comparing to the MSHL descriptor, LMGO-HOG preserves more detailed information. In the meantime, it removes the noise brought by small irrelevant gradient information comparing to HOG descriptor.

To check the effectiveness of our method, we chose the predefined window to be a block, shown in Fig. 1 (a). The comparison between the original HOG feature and the LMGO-HOG feature can be seen in Fig. 4. The LMGO–HOG histogram is shown in Fig. 4 (c), in which the important features are emphasized comparing to Fig. 4 (a).

(iii) CD-HOG and LMGO- HOG in cascade

After we got *CD*-HOG according to the previous section, we tried to obtain the local main gradient orientation in every block of original video image, and calculated the LMGO-HOG by applying Eq. (2), only after

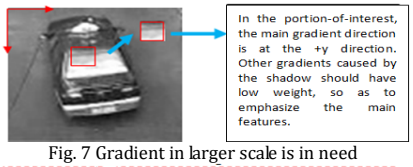


Fig. 7 Gradient in larger scale is in need

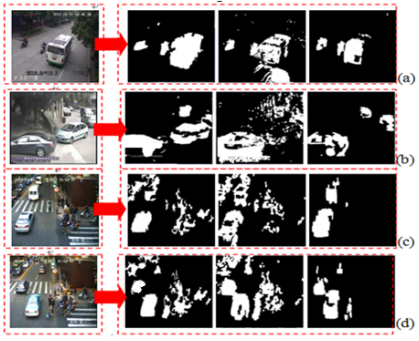


Fig. 8 Four comparisons among our proposed method, GMM and frame difference foreground extraction method. In every row, the leftmost is the original image, and the second one is the obtained with our method and the third is the result of GMM; the rightmost one is the result of the frame difference method. (a) Our method can eliminate the shadow caused by sunlight and extract integrated contours. (b) Our method can avoid large noise brought by large moving objects and detect still cars. (c) Our method can extract precise contours for moving objects. (d) Our method has less chance to join two separate moving objects together.

replacing $v_{hog}(i)$ with CD-HOG. The final feature of CD-LMGO-HOG is shown in Fig. 4 (d). The redundancy is eliminated and the main features are salient. The result of this method is better than the previous ones which are shown in Fig. 4 (b) and Fig. 4 (c).

IV. EXPERIMENTAL RESULTS

In this section, we will show experimental results of our proposed algorithms.

(I) Our proposed CDFE method is good in many ways:(a) It has low sensitivity to illumination changes; (b) It can preserve the foreground of the moving objects that stay still during a certain period in the video; (c) No additional noise will be introduced into the system when large cars pass by; (d) It has lower chance to join two separate moving objects into one contour. Our method has been tested in 10 different and complex videos, and all achieved good results. Some typical results have been shown in Fig.9 to demonstrate the advantages of our proposed method by comparing it with GMM and frame difference foreground extraction method.

(II) Dalal[5] mentioned in his paper that he used 9 unsigned bins for the human descriptor. In our case, we applied 16 signed bins for car classification. We tested our algorithm on 10 individual videos, which included 386 moving objects. Among these moving objects, there were 219 cars, and 93 vans. If several people were detected in one testing window, we took down as one event for detecting people. Regarding parameterization, for the extended and original HOG- descriptors, we used a cell size of 10×10, a detector block size of 20×20, and a window size of 40×40. The stride of block was 10×10, and 20×20 for the window. As for MSHL feature, we used windows with 11 different scales to scan the image. The statistics are listed in Tab.1.

In the table, GMM and CDFE refer to the foreground extraction methods. We apply Basic HOG with 16 bins so as to make impartial comparison. By comparing the first two statistics columns, we can see that our proposed foreground extraction method has a better performance than GMM. CDFE is low sensitive to light, and provides a precise location of moving object, therefore the result is better. By comparing the last five columns, we can see clearly that MSHL feature is not good at classifying cars in cluttered background, and we can also see that our proposed methods (the last three columns) all have better performance in classification than the original HOG. The cascading algorithm (CD-LMGO-HOG) is obviously the best of all, and it achieves better performance with high resolution videos. Fig. 9 shows some final classification results.

Table 1 The comparison results between our proposed algorithm and previous works

Detection result		GMM+Basic HOG	CDF+Basic HOG	CDFE+ MSHL	CDFE+LM GO-HOG	CDFE+ CD -HOG	CDFE+CD -LMGO-HOG
car	car	190	195	189	199	196	**202**
	van	18	18	21	17	18	**12**
	people	11	6	9	3	5	**5**
van	car	15	16	29	8	11	**7**
	van	74	77	64	83	82	**85**
	people	4	0	0	2	0	**1**
people	car	0	0	16	1	0	**0**
	van	2	2	7	1	0	**0**
	people	74	74	53	74	76	**76**
Accuracy		87.56%	89.61%	79.27%	92.1%	91.71%	**94.04%**

Fig. 9 The final classification results.

Acknowledgements

This work was supported in part by the following grants: National Science Foundation of China grants (61001146), Huawei Innovation Program grant (YB2012120150), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), the SMC grant of SJTU, Shanghai Pujiang Program (12PJ1404300), and the China National Key Technology R&D Program (2012BAH07B01).

Reference

[1] C. Stauffer and W. E. L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747-757, 2000.

[2] Z. Zivkovic and F. Heijden, "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction," *Pattern Recognition Letters*, vol. 27, pp 773-780, 2006.

[3] P. KaewTraKulPong, R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection", *2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[4] C. Stauffer, W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *IEEE Computer Vision and Pattern Recognition*, 1999.

[5] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[6] C. Thurau, V. Hlava, "Pose primitive based human action recognition in videos or still images," *IEEE Computer Vision and Pattern Recognition*, 2008.

[7] F. Kong, Q. Ye, N. Zhang, K. Lu, J. Jiao, "On-Road Vehicle Detection Using Histograms of Multi-Scale Orientations". *IEEE Int'l Youth Conference on Information, Computing and Telecommunications*, pp. 212-215, 2009.

[8] A. Criminisi, P. Perez, and K. Toyama, "object removal by exemplar-based inpainting," *IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

[9] A. Kokaram, B. Collis and S. Robinson, "A Bayesian framework for recursive object removal in movie post production," *IEEE Int'l Conf. Image Processing*, vol. 1, pp.937-40, 2003.

[10] O. Rostamianfar, F. Janabi-Sharifi, and I Hassanzadeh, "Visual tracking system for dense traffic intersections," *Canadian Conf. Electrical and Computer Engineering*, pp. 2000-2004, 2006.

[11] K.A. Patwardhan,G. Sapiro, and M. Bertalmio, "Video inpainting under constrained camera motion," *IEEE Trans. Image Processing*, vol. 16, no. 2, pp. 545-553, 2007.