Improved Pedestrian Detection Based on Extended Histogram of Oriented Gradients

ZHANG Li-hong
School of Physics and Electronic Engineering
Shanxi University
Tai Yuan, China
e-mail:lhzhang@sxu.edu.cn

LI Lin
School of Physics and Electronic Engineering
Shanxi University
Tai Yuan, China
e-mail:lin.ahch@gmail.com

Abstract—**In order to further improve pedestrian detection accuracy and avoid the disadvantage of original histogram of oriented gradients (HOG), differential template, overlap ratio and normalization method and so on are improved when HOG features are extracted, then more gradient information are extracted and feature description operators can be obtained which describe human detail features better in lager image regions or detection windows. Considering speed, we select support vector machine (SVM) using linear function kernel as a classifier. Multi-scale detection technique and non maxima suppression method are employed for precisely locating the pedestrians in the image. Experiments show that the human detection system improves detection accuracy and still maintains a relatively satisfactory speed.**

Keywords- histogram of oriented gradients; pedestrian detection；support vector machine; multi-scale detection

I. Introduction

Pedestrian detection is not only an important branch of target detection, but also a hot and difficult point of studies in the field of computer vision. Its applications can touch upon intelligent transportation, safety monitoring, autonomous driving, human computer interaction and so on [1-4]. In view of the nonrigid factors such as the light, posture and color of clothes, pedestrian detection is still a difficult problem.

At present, there exist many technical methods for pedestrian detection, most of which are based on machine learning. The detection technique possesses two important aspects. The first one is feature descriptor which can represent detection targets. The other one is the chosen learning algorithm. Features commonly used include edgelet [5], Harr-like wavelet feature set [6], local binary pattern [7], histogram of oriented gradient [8] and so on. All these features are used to extract information about marginal changes and contour shape. The learning algorithms mainly include cascade connection AdaBoost and SVM [9][10]. AdaBoost is a kind of iterated algorithm. It trains repeatedly on the same train dataset and then combines different weak classifiers according to the weight of the last samples to obtain ultimate strong classifiers. The main advantage of cascade classifier is its excessively high detection speed. SVM maps the sample set to the linearly separable higher dimensional space by kernel function. In the higher dimensional space, dot product operation is only needed to get the discriminant results. Its main advantage is the robustness if the target mode changes.

Based on the previous studies, further improvement is made in this paper. Firstly, in consideration of the fact that Dalal'HOG cannot extract the body local features in comparatively large image region[11], this paper extracts more gradient information and thus get more description operators which can indicate more information about pedestrian features.

II. The Basic Algorithm Theory

A. Histogram of oriented gradients

Put forward by Dalal, HOG is a kind of descriptor which can extract local information. It extracts distribution information of the target edges in the local regions and can represent the target shape better. The original HOG adopts one dimensional differential template [–1, 0, 1] to do the gradient computation. In a 2×2 cell, or a block, a 36 dimensional row vector is generated. In the process of the block moving, 0.5 overlap ratio is adopted. Thus, in a 64×128 detection window, 105 blocks generates a 3780 dimensional row vector, i.e. a sample eventually.

We make the following improvement on the HOG. At first, [–0.5, –0.5, 0, 0.5, 0.5] differential template is used and 81 dimensional row vector is produced in the larger 3×3 cell. In the process of the block moving, 1/3 overlap ratio is adopted. In a 64×128 detection window, there are 84 blocks, which can generate 6804 dimensional row vector at last. Dalal uses the L2 normalization feature in every block while the method of L1 normalization is adopted in this paper to accelerate the speed.

I is used to stand for a picture and I(x, y） stands for the grey level of the image in the pixel point (x, y). The specific calculating process of the extended HOG is showed in the following part.

1) Gradient calculation :

$$G_x(x,y) = \frac{1}{2}[I(x+2,y)+ I(x+1,y) - I(x-2,y) - I(x-1,y)], \quad (1)$$

$$G_y(x,y) = \frac{1}{2}[I(x,y+2)+ I(x,y+1) - I(x,y-2) - I(x,y-1)]. \quad (2)$$

Published by Atlantis Press, Paris, France.

x, y respectively represents the horizontal and vertical coordinates of a certain pixel point in the picture. $G_x(x,y)$ and $G_y(x,y)$ are for representing respective gradient value of this point in the direction of x and y.

2) The calculation of gradient strength M (x, y) and gradient direction θ(x,y）:

$$M(x,y)=\sqrt{G_x^2(x,y)+G_y^2(x,y)} \quad , \tag{3}$$

$$\theta(x,y)=\arctan(G_x(x,y) / G_y(x,y))+\frac{\pi}{2} \ . \tag{4}$$

$\theta(x,y)$ is restricted in the range of $[0, \pi]$.

3) The tri-linear interpolation calculation: $[0, \pi]$ is evenly divided into nine bins and the tri-linear interpolation method is utilized to distribute every angle of gradient direction to adjacent bins according to certain proportion.

4) L1 norm normalization: In every block, L1 norm normalization is adopted to eliminate noise and a feature $B_i=(f_1,f_2,\ldots,f_{81})$ is collected.

5)To obtain ultimate sample: Move the block at 1/3 overlap ratio and repeat the procedure from 1) to 4) until every pixel point in the detection window is calculated. The ultimate sample can be represented as $F=(B_1,B_2,\ldots,B_{84})=(f_1,f_2,\ldots,f_{6804})$.

B. SVM

SVM is a kind of pattern recognition technique based on statistical learning theory [12]. Kernel function is adopted to map the data in the input spaces to a high-dimensional feature space. Then, in this high-dimensional space, the generalized optimal classification face is calculated. Thus, the linearly inseparable data in the original space can be separated linearly in the high-dimensional space [13]. The general expression is:

$$g(x) = w \cdot x + b. \tag{5}$$

The optimal classification face demands that the distance of samples which are the nearest from the classification face should be as large as possible. The sample point is (x_i, y_i) and the sample class label is $y_i = \{-1, +1\}$. Thus, the possibility of the miss-separation of the detection samples is comparatively small, that means, it is needed that

$$y_i[(w \cdot x) + b] -1 \geq 0. \tag{6}$$

When $\| w \|$ is set to the minimum value, the class interval $d=2/\| w \|$ is the maximum.

The frequently-used kernel functions are as follows.

- The linear kernel function:

$$K(x,y) = x \cdot y . \tag{7}$$

- Polynomial kernel function:

$$K(x,y) = (x \cdot y + 1)^d,\ d=1 ,2, \ldots \tag{8}$$

- Radial basis function:

$$K(x,y)= \exp(-\gamma \| x-y \|^2). \tag{9}$$

- The Sigmoid kernel function:

$$K(x,y)= \tanh(b(x \cdot y) -c). \tag{10}$$

In the experiment conducted in this paper, the linear kernel function is chosen because the input data should be mapped to the high-dimension spaces to choose the optimal classification faces if the other three kernel functions are used. However, the linear kernel function chooses the optimal classification faces directly in the original spaces. Thus, the linear kernel function is much faster than other kernel functions in the respect of speed.

III. Pedestrian detection system

When the trained classifiers are employed to detect, two problems will occur. The first one is that the size of detection window we have chosen is 64×128. If the size of the target in the detect-waiting pictures is excessively large or small comparing with the detection window, the miss-detection will occur easily. The second problem is that when the same target is detected, the sliding of the detection window in the detect-waiting images will result in the multi-results at the same target, so amalgamation of these detection results is required.

Considering the above problems, multi-scale detection is adopted in this paper. The main idea is to set the detect-waiting images in zoom mode according to certain proportion firstly. Moreover, the windows whose overlap ratio exceeds certain proportion in the detection results are regarded as the multi-windows for the same target. Then, non-maxima suppression technique is adopted to choose the window whose score is the highest as the position of the target.

Combining with HOG extraction and SVM training, the flow chart in this paper is showed in Fig. 1. There are three parts in this flow chart: features extraction, training and detection.

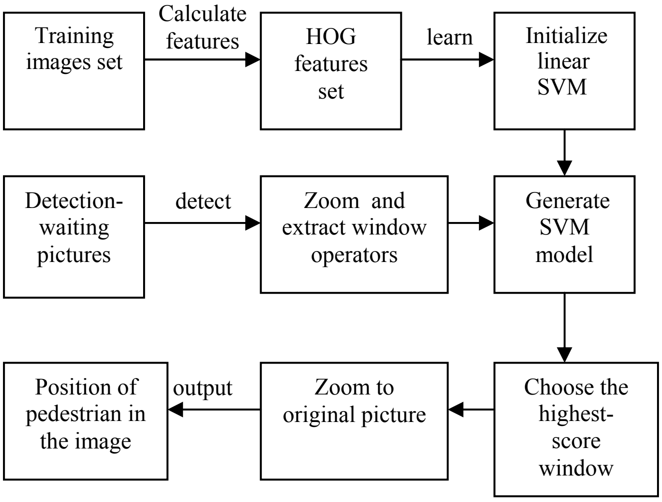


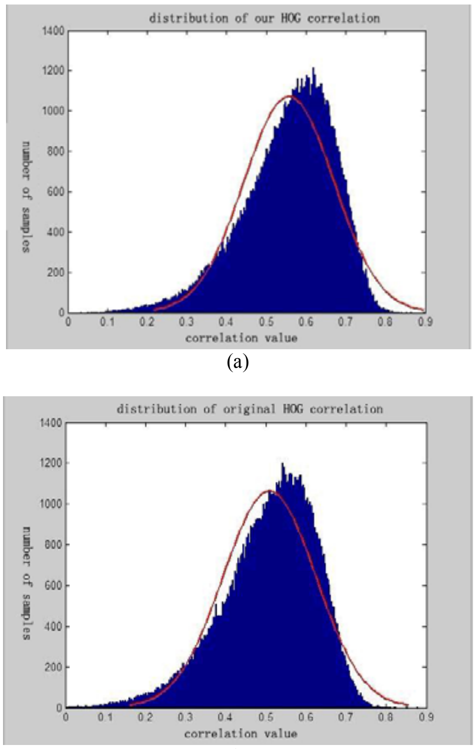
Figure 1. Flow chart of detection system

IV. Analysis of experiment results

In this section, the detection system is tested on the INRIA data set and makes a comparison with the existing technique. It is realized in the VC and OpenCV developing environments.

A. *Comparison of the correlation coefficient of HOG*

Five hundred images are chosen among the 2416 pedestrian images and the correlation coefficient is calculated using original HOG and the improved HOG respectively. 124,750 correlation coefficients are obtained by 500 samples in all, as shown in Fig. 2.



(a)

(b)

Figure 2. Stability of our HOG and original HOG

These correlation coefficients are utilized for measuring the stability of two kinds of characteristics. It can be seen in the extraction Fig. 2 that most improved HOG correlation coefficients are concentrated in 0.55 to 0.65, the original HOG correlation coefficients are concentrated in 0.5 to 0.6. The correlation coefficient is larger the similarity among features is greater, and the extracted features are to be able to characterize more information of the pedestrian.

B. *Comparison of the classifier performance*

The method employed by Dalal & Triggs and that of this paper are to be compared. In Dalal's technique, hard samples are trained repeatedly. This procedure cannot have an effect on the comparison of these two features, so this procedure can be omitted. 7,308 negative samples and 2,416 positive samples are extracted from the 1218 negative image samples in the INRIA library. The linear SVM detectors are obtained by the training. Fig. 3 shows the comparison of the classifiers trained by two kinds of samples. It can be seen from the Fig. 3 that when the FPPW (False Positive Per Window) is comparatively small, the miss rate of this paper is much smaller than that of Dalal. Lower FPPW is usually required in the practical application, and this paper can keep comparatively high detection rate in the situation of low FPPW (1 - miss rate).

In Fig. 4, some examples about detection results are given. It is clear that most pedestrian targets can be detected utilizing the technique in this paper. For example, for the five pedestrian targets in 4(b), the pedestrians in the complex backgrounds where there is a juncture of strong light and soft light are not detected and three miss-detections also occur. However, the pedestrian targets whose postures are changed to a certain extent can be detected effectively.

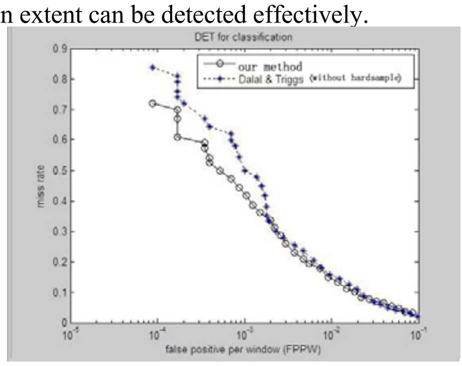
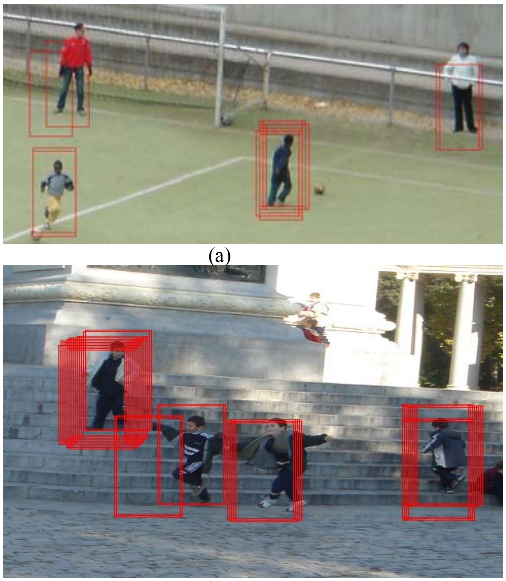


Figure 3. Performance and comparison



(a)

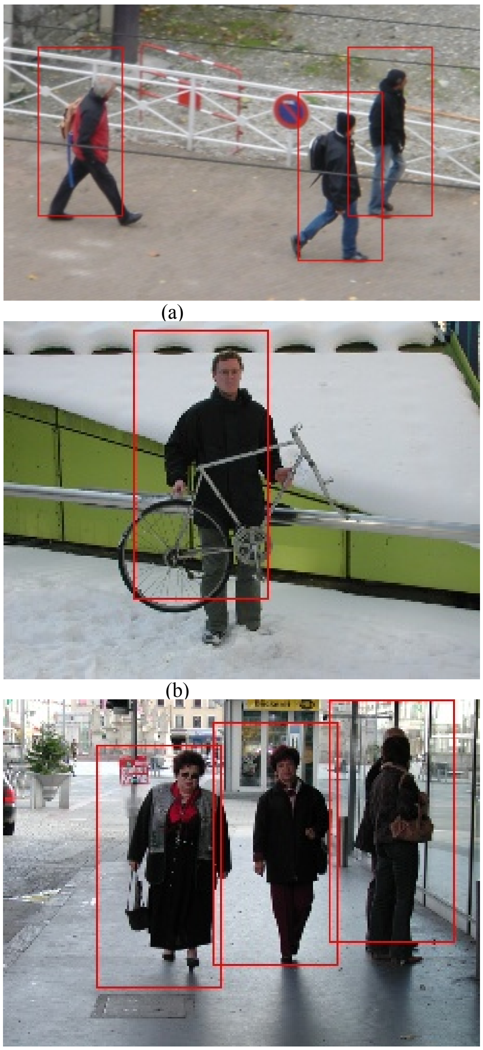
(b)

Figure 4. Detection examples

C. *Multi-scale detection*

The performance of the classifier is determined by the selected features and learning algorithm, so detection effect can be improved only by the zoom images before detection. Based on the multi scale images detection, non maxima suppression technique is employed to choose the optimal detection window in detection targets.

The final detection experiment results are shown in Fig. 5. It can be seen from Fig. 5 (a) to 5 (c) that the testing system can detect pedestrian target well.



(a)

(b)

(c)

Figure 5. Multi-scale detection examples

In practical application, the detection speed is also an important measure to an algorithm. The detection speed of Dalal algorithm is 1.2 times of that of ours. Although the algorithm in this paper is a bit slower than Dalal's , it greatly enhanced detection accuracy.

V. CONCLUSION

Feature and learning algorithm are most important in pedestrian detection based on machine learning. We presented refinements leading to a significant improvement on an existing pedestrian technique. The traditional HOG can not describe pedestrian body detail in larger image region. Considering this point, this paper extends original HOG features, but maintains the original computing system. Experiments shows the improved features can achieve satisfactory detection accuracy, higher than original HOG. But to solve the slower detection speed is an important point of our next work. We will endeavor other techniques (e.g. integral image) to solve this problem.

ACKNOWLEDGMENT

This work is supported by Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi Province (2010002).The authors also would like to express their gratitude to the other members of the algorithm team for valuable discussions.

REFERENCE

[1] Li L, Leung MKH. Unsupervised learning of human perspective context using ME-DT for efficient human detection in surveillance. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition.2008. pp.91-97.

[2] Wojek C, Walk S, Schiele B. Multi-Cue onboard pedestrian detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition.Miami:IEEE,2009. pp.794-801.

[3] Xu R, Zhang B, Ye Q, Jiao J. Cascade L1-norm classifier for pedestrian detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition. 2010. pp.89-96.

[4] LI H J, LIN SH X, ZHANG Y D. A survey of video based human motion capture [J]. Journal of Computer Aided Design & Computer Graphics, 2006, 18(11).pp. 1645-1651.

[5] Wu B, Nevatia R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of Edgelet part detectors. In: Proc. of the IEEE Int'l Conf. on Computer Vision.Beijing,China,2005. pp.90-97.

[6] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Conference on Computer Vision and Pattern Recognition (CVPR),Florida,USA:IEEE,2001. pp.511-518.

[7] Wu B, Nevatia R. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. (CVPR), 2008.pp.1-8.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Conference on Computer Vision and Pattern Recognition (CVPR), 2005. pp.886-893.

[9] Munder S, Gavrila D. An experimental study on pedestrian classification. IEEE Trans. on Pattern Analysis and Machine Computer Vision.vol.28,no.11 Nov.2007.pp.1863-1868.

[10] Wu B, Nevatia R. Cluster boosted tree classifier for multi-view, multi-pose object detection. In: Proc. of the IEEE Int'l Conf. On Computer Vision. 2007.pp.1-8

[11] P.Felzenszwalb, R.Girshick, D.McAllester, D.Ramanan. Object detection with discriminatively trained part based models, JAIR,29,2009.

[12] Cui Jianguo, Li Zhonghai. The application of support vector machine in pattern recognition [J]. IEEE Intrernational Confernce on Control and Automation, 2007, June.pp 3135-3138.

[13] Belousov A I, Verzakov S A, Von Frese J. A flexible classification approach with optimal generalization performance: support vector machines [J]. Chemometrics and Intelligent Laboratory Systems, 2002, 64.pp. 15-25.