Fast Human Detection Using Motion Detection and Histogram of Oriented Gradients

Hou Beiping
School of Automation and Electricity, Zhejiang University of Science and Technology, Hangzhou, China
Email: houbeiping@126.com

Zhu Wen
School of Automation and Electricity, Zhejiang University of Science and Technology, Hangzhou, China
Email: joywenzhu@126.com

Abstract—**This paper presents a real-time Human detection algorithm based on HOG (Histograms of Oriented Gradients) features and SVM (Support Vector Machine) architecture. Motion detection is used to extract moving regions, which can be scanned by sliding windows; detecting moving region can subtract unnecessary sliding windows of static background regions under the surveillance conditions, then detection efficiency can be improved. Every sliding window is regarded as an individual image region and HOG features are calculated as classified eigenvectors. At last, the detected video objects can be categorized into pre-defined groups of humans and other objects by using SVM classifier. Experimental results from real-time video are provided which demonstrate the effectiveness of the method.**

Index Terms—**Human Detection, HOG, Motion Detection, SVM**

I. INTRODUCTION

Finding people in images has attracted much attention in recent years for practical applications such as visual surveillance [1], Vehicle auxiliary driving [2] and image understanding. It is a next step after the development of successful face detection algorithms. The detection of humans has become an own research field.

However, unlike other object detection, human detection has some of its own characteristics. Humans usually have many different appearances in pose and style, and the background of the images or videos is often cluttered and has on general describable structure. So, human detection in image/videos is a challenging task for the variable appearance and various poses, which can influence the algorithm of choice. The articulated pose, style and color of clothes, illumination conditions in outdoor scene will affect the detection results.

The work of finding people from images or video can be divided into three stages; the first is ROI (Region of Interest) selection; the second is the selection of effective features; the third is objects classification.

For the ROI selection method, the recent research [3, 4, 5] indicate that sliding window is the predominant method being used in object classification, face recognition, human detection, due to its good flexibility and effectiveness. For the sliding window approach, each frame image is shifted from the top left to the bottom right with rectangular sliding windows in different scales, it is shown in Figure1. In each sliding window, some certain features such as texture, shape information, and gradient directions are extracted and fed to a classifier, which is trained offline by sample training image data. The trained classifier can estimate the sliding windows region is a person or not. For sliding window, the computational costs are often too high to real-time applications [6]. Some significant speed-up methods are used to reduce computation time [7, 8].

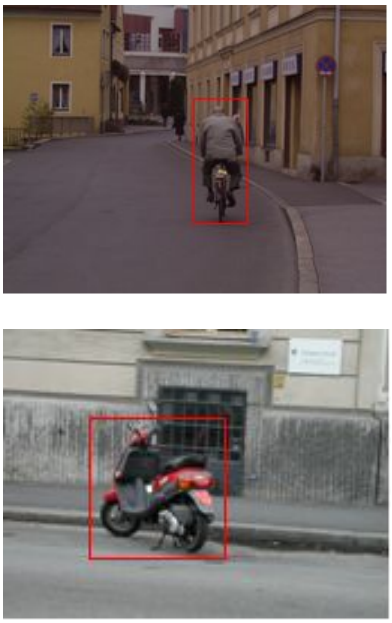


Figure 1. Sliding windows on images

The feature selection is a key problem in the process of human detection. The selected features must embody the objective characteristics, proper features can improve classification accuracy, while the improper features may lead to misjudge. So, how to select human features is very important. Some researchers [9] use global features like body shape or silhouettes to express the difference

doi:10.4304/jcp.6.8.1597-1604

between human and other objects. However, these features are not flexible since human can have many kinds of poses and shapes. So it's hard to model human for a trained classifier.

Compared to global features, local features are more suitable. SIFT (Scale Invariant Feature Transform) [10] features are useful local features to object classification; they are invariant to scale and rotations.

Gradient orientations are very robust local features [11].Multi-scale features can be found by calculating orientation histograms. Dalal and Triggs [3] present a human detection algorithm with excellent detect precision. The method uses a dense grid of histograms of oriented gradients (HOG), and which is proved to be powerful enough to classify humans and other objects. However, the process of HOG features extraction is time-consumed. To speed up Dalal's method, some improvements are achieved; Qiang Zhu and Shai Avidan [12] proposed a fast human detection with variable-size blocks. Xiaoyu Wang[13] use histograms of oriented gradients(HOG) and Local Binary Pattern(LBP) as the feature set, the method include global detector for whole scanning windows and part detector for local regions, the detection result is good.

Colour information is used to detect humans. Sebastian ntabone and Alvaro Soto [14] present a human detection system based on visual saliency mechanism and color features.

Michael Oren and Constantine Papageorgiou [15] present trainable human detection architecture on the basis of wavelet template that defines the shape of an object in terms of a subset of the wavelet coefficients of the image. Papageorgiou and Poggio [24] use absolute values of Harr wavelet coefficients at different orientations and scales as their local descriptors. Horizontal, vertical, and diagonal wavelets are used. The descriptor vectors are used in a kernels SVM framework.

Component-based human detection system is proposed by Mohan and Poggio [16].This system is structured by four distinct detectors to find four components of the human body: head, legs, left arm, and right arm. Although this method is robust, it is a time-consuming method. Wu and Nevatia [17] use Bayesian combination to combine the part detectors to get robust detection.

Viola [18] proposed a human detection method which integrated image intensity information and motion features, this algorithm can detect humans with front and back views.

Apart from features selection, another key point is the design of classifier architecture. After features are extracted from each image, some classifiers for supervised learning such as neural network [22, 23], support vector machine are then used to classify objects based on sample data. Discriminative classification techniques aim at determining an optimal decision boundary between pattern classes in a feature space. Neural network [26] is applied to many research fields; it is an effective tool to image classification and recognition. In the process of pedestrian detection, multi-layer neural networks have been utilized in conjunction with adaptive local feature in the hidden network layer. This method unifies feature extraction a neural network classification within a single model.

Support vector machine (SVM) [19] have evolved as a powerful tool to solve pattern classification problems. In contrast to neural network, SVM do not minimize some artificial error metric but maximize the margin of linear decision boundary to achieve maximum separation between the object classes. For human detection, SVM classifiers have been used in combination with various features.

Except for neural network and SVM, Adaboost [20] has been used successfully in face detection and face recognition. It has been used to construct strong classifiers as weighted linear combinations of the selected weak classifiers, each involving a threshold on a single feature. The cascade architecture is tuned to detect almost all pedestrians while rejecting no-pedestrians as early as possible.

This paper describes a method for classifying moving objects as either a pedestrian or not. The work includes three steps. Firstly, motion regions are detected. Secondly, the HOG features are extracted from motion region. Thirdly, pedestrians can be detected by SVM classifier. Our method builds on the above described earlier work; compared to the sate-of-art detectors, our method achieves some improvements for real-time applications.

The paper is organized as follows: Section I describes the research situation of human detection. Section II outlines our human detection approach. Section Ⅲ introduces the motion detection method. Section Ⅳ presents the feature extraction algorithm. The structure and algorithm of SVM are shown in section V. Experimental verification of the proposed method is shown in section VI. Finally, the conclusion is given in section VII.

II. THE ARCHITECTURE OF OUR APPROACH

The human detection algorithm is presented under the condition of visual surveillance, and the camera position is fixed, so the view is constant. The main purpose of motion detection is to decrease image regions which will be scanned by sliding windows, the detection of moving objects is valuable under the surveillance condition. Some static regions are usually backgrounds which are not necessary to be detected by sliding windows. So, just motion regions need human detection after motion detection, it will reduce computation time, and this method can be used to real-time human detection application.

As shown in Figure 2, moving regions can be found after motion detection of frame image from real-time video. Because moving region may contain multi-person, person with other objects, vehicles, etc. Sometimes, the change of illumination conditions can produce motion region. So, the detected motion region should be detected

by sliding window method, while not classified straightforwardly by classifier.

Motion regions will be scanned by sliding windows with varied size and sliding step, every window means a small frame image region. HOG features can be calculated from window region, then, the trained classifier can determine the corresponding window region is a person or not. Classified results rely on two key points, one is the selection of feature vector, and it is a critical factor of right classification. The second key point is the design of classifier.

In this paper, HOG is used to reflect the human body features. After motion detection, sliding windows can be used to shift on detected regions. Each window means an individual image region, and corresponding HOG feature vector will be calculated, which reflect the edge and gradient information of image region.

SVM is selected to classification; it should be trained by sample data before classifications. The sample data is the HOG feature vectors of general human and background images. When the support vectors are trained sufficiently, the SVM classification can be used to recognize humans from static images or real-time video. The architecture of classification system is shown as Figure 2.

In our research work, the moving objects in outdoor scenes can be classified to two categories, pedestrians and others. Every window region can be seen as an independent image, then its feature vector is calculated and its category will be distinguished by priori knowledge on the base of SVM. The feature vector will be introduced in section Ⅳ.

Ⅲ. Motion Detection

In this stage, a temporal differencing detection algorithm is used to extract moving region. There are many variants on the motion detection method, but the simplest is to take consecutive video frames and determine the absolute difference. A threshold value is used to determine the results. If I_n is the intensity of the nth frame image from real-time video, I_{n-1} is the previous one. Then the pixelwise difference function Δn is defined as

$$\Delta n = \left| I_n - I_{n-1} \right| \tag{1}$$

Accordingly, the motion image M_n can be calculated by threshold $T_{threshold}$

$$M_n(i,j) = \begin{cases} I_n(i,j), & \Delta n(i,j) \geq T_{threshold} \\ 0, & \Delta n(i,j) \geq T_{threshold} \end{cases} \tag{2}$$

Where, *i* and *j* are pixel positions along horizontal and vertical directions.

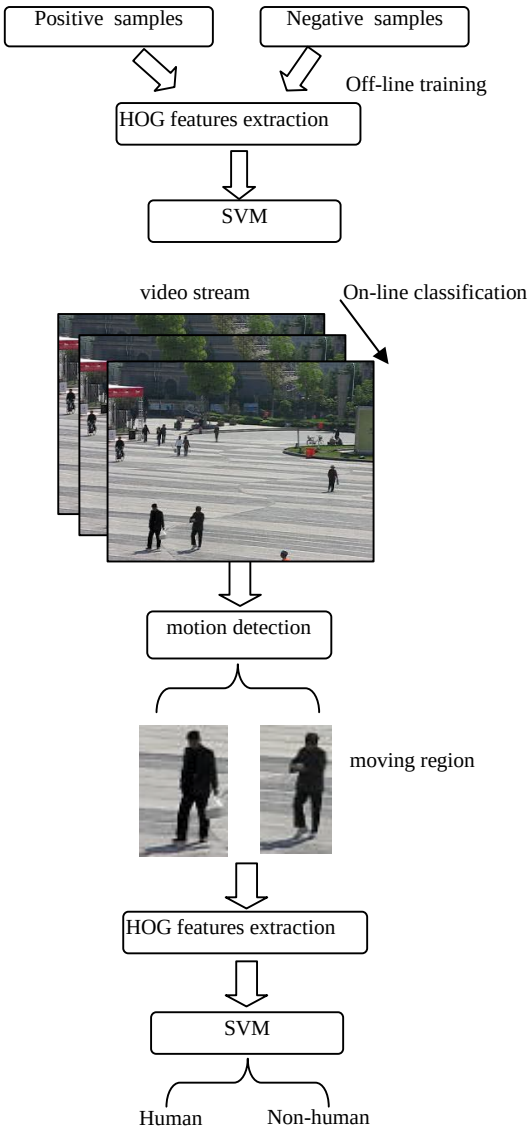
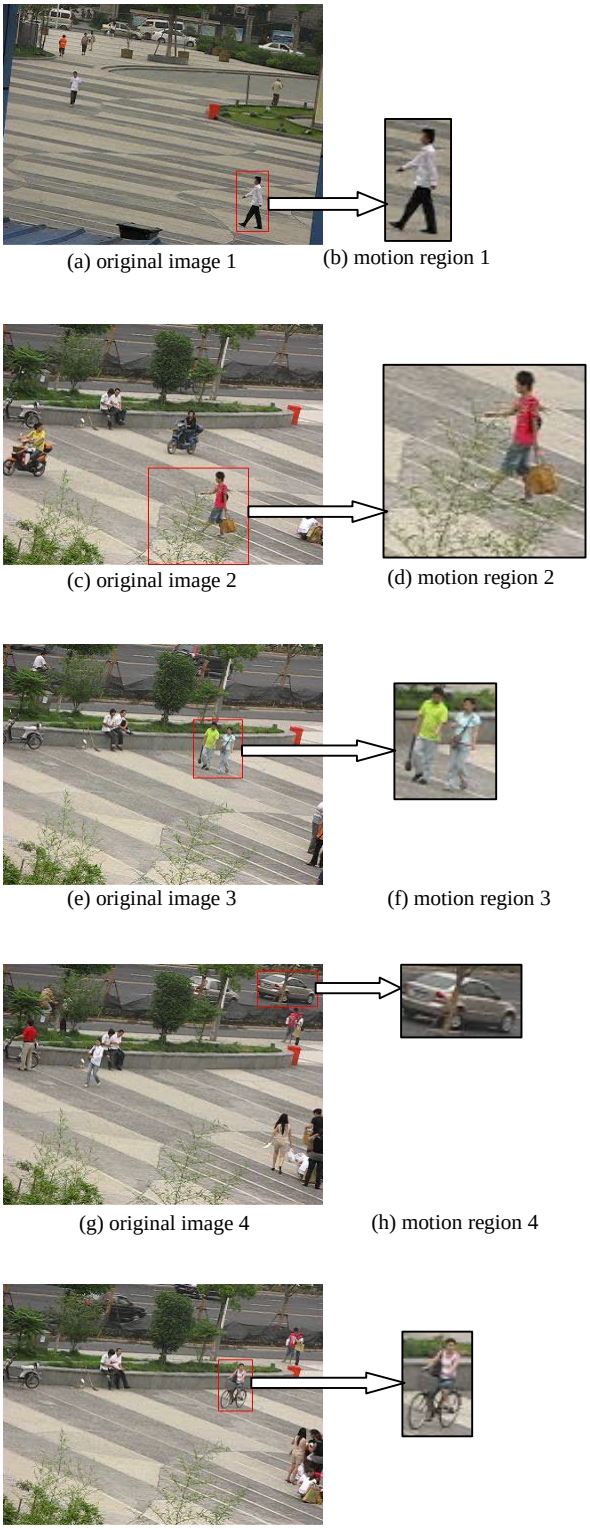


Figure 2. The architecture of detection system

After the motion image is determined, several motion regions can be extracted. Figure 3 shows the motion extraction results.

From Figure 3 (a) (b), we can see that walking pedestrian region can be extracted, most of the original region can be filtered by motion detection and the computational complexity can be reduced. The motion region 2 is bigger than a pedestrian area because the tree branch is swinging with wind, then walking pedestrian and branch form one moving region, the person can be found by the following detection algorithm. As is shown in Figure 3 (f), the motion region 3 corporate two people, we may call it "multi-person" region. Because the distance between people is small, so, they can be seen as one region after motion detection, sliding windows and human detection algorithm can divide them into two individuals. Motion region 4 is a running car; of course, it can be classified as a non-human object.



(a) original image 1 (b) motion region 1

(c) original image 2 (d) motion region 2

(e) original image 3 (f) motion region 3

(g) original image 4 (h) motion region 4

(i) original image 5 (j) motion region 5

Figure 3. Motion region examples

Ⅳ. HOG DESCRIPTOR

This section gives an introduction of HOG feature extraction algorithm [3]; the method is based on evaluation well-normalized local histograms of image gradient orientations in a dense grid. The basic idea is the local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions [25, 26].

HOG features are similar to SIFT [10] descriptor, gradient orientation information is used to embody characteristics of objects; the latter is calculated on the basis of key point detection and is used to image matching. While HOG features are descriptions of orientation histograms and can be calculated as follows.

The first stage applies a simple template $[-1,0,1]$ to calculate gradients of every point along two directions.

In the second stage, the Image detection window is divided into small spatial regions, called "*cells*". For each cell, a local 1-D histogram of gradient of gradient or edge orientation over all the pixels in the cell is accumulated. This is the representation of "*orientation histogram*" representation. The gradient magnitudes of the pixel in the cell are used to vote into the orientation of histogram. Several cells form a group, called "*block*".

The third stage is the calculation of normalisation, for better invariance to illumination, shadowing, it is useful to contrast-normalize the local responses. This can be done by accumulating a measure of local histogram "energy" over "*block*". The result is used to normalise each cell in one block. The cell thus appears several times in the final output vector with different normalisations. Then the normalised block descriptors are defined as Histogram of Oriented Gradient (HOG), as shown in Figure 4.

The last stage collects the HOG descriptors coving from all blocks the detection window into a feature vector for the classification use.

The HOG descriptors can capture local contour information, the edge and gradient structure, and reflect the characteristics of local shape. HOG features can be used to pedestrian detection.

V. SVM CLASSIFIER

There are discriminative classifiers which try to find characteristic differences between positive and negative examples. SVM is a widely used technique. SVM computes a high-dimensional hyperplane to separate the different object categories. To compute the plane, the chosen image feature space or a kernel of this feature space is used. SVM performs a two-class classification in two stages [21]:

First is the training stage, each window i is represented by vector x_i with label $y_i=\pm 1$,the classification function $C(x)$ is defined as

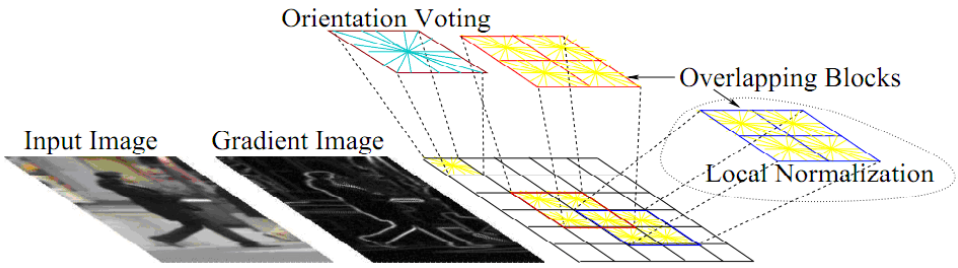


Figure 4. HOG feature extraction process

(a) Positive sample images

(b) Positive sample images

Figure 5. Positive and negative samples

$$C(x) = \sum_{i=1}^{\ell} \alpha_i y_j k(x_i, x) + b \quad (3)$$

where parameter b is estimated using Kuhn-Tucker conditions, parameter α_i is computed by minimizing the quadratic problem:

$$W(\alpha) = -\sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (4)$$

Under the constant $\sum_{i=1}^{\ell} y_i \alpha_i = 0$,with $K(x, x')$ a positive definite kernel. Training samples with non-zero α_i are the support vectors.

Second is the testing stage, the classifier function $C(x)$ is applied to image windows detected as road signs to estimate the confidence in class value.

According to the analysis above, $C(x)$ is the estimated classification value. We link the value with the human detection. If $C(x)$ >0, the sliding window can be regarded as a human region, the higher $C(x)$, the higher the confidence. On the contrary, if $C(x)$ <0, the detection window can be regarded as other objects.

VI. Experiment Results

According to the analysis of classification system, the first work is to train SVM classifier, we use static image as positive and negative examples from INRIAPerson training sets (available at http://pascal.inrialpes.fr/data/human). Each 1000 samples of human and non-human are prepared, and their size is 128×64,as show in Figure 5; HOG features of every sample image should be extracted.

Each 1000 test images of human and background from INRIAPerson training sets are prepared. Table I shows the right rates and error rates of test image samples. For instance, there are 96% human samples are classified as human, while 7% Non-human samples are recognized as human, so the right rate of Non-human sample is 93%, error rate is 4%. Figure 6 shows the classification result on static image.

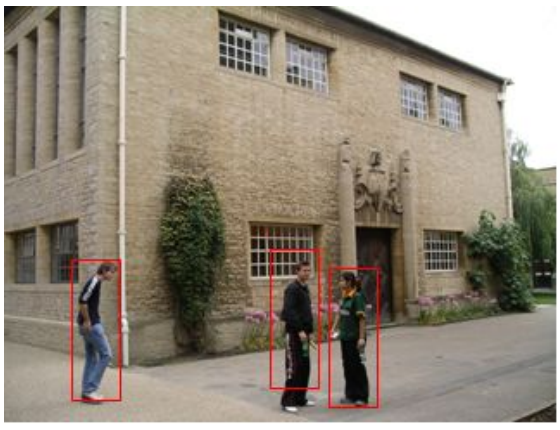


Figure 6. Detection results of static image

TABLE I.
SVM Classification Results for Static Images

Result / Test image	Human (%)	Non-human (%)
human	96	4
Non-human	7	93

To compare the performance of classifiers, BP neural network method is used to show its classification ability. The same with SVM, Each 1000 samples of human and non-human are prepared, the HOG features are input vector of neural network, and the teacher signal is defined as 1 and 0. New data is prepared to test the classification result after training. Table II shows the result.

TABLE II.
NN Classification Results for Static Images

Result / Test image	Human (%)	Non-human (%)
human	81	19
Non-human	28	72

By comparisons, SVM is preferable to human classification. Our classification method can be used to real time video, when the resolution of scene is 320×240, frame rate is 20fps, Figure 7 and table III show the classification results .

TABLE III.
Classification Result for Real-time Videos

	Video Stream			
	1	2	3	Total
Total People	300	550	420	1270
Detected People	271 90.33%	498 90.54%	377 89.76%	1146 90.23%
False Positive	27	48	41	116

VII. Conclusion

In this paper, a human detection method based on motion detection is presented. Motion detection is used to extract moving regions, which can be scanned by sliding windows. Every sliding window is regarded as an individual image region and HOG features are calculated as classified eigenvectors. At last, the detected video objects can be categorized into pre-defined groups of humans and other objects by using SVM classifier. Experimental results from real-time video are provided the effectiveness of the method.

The experiments results prove the validity and effectiveness. The recognition accuracy was acquired under laboratory setting, so it had some limits. In future work, this classification method will be used in intelligent surveillance field.

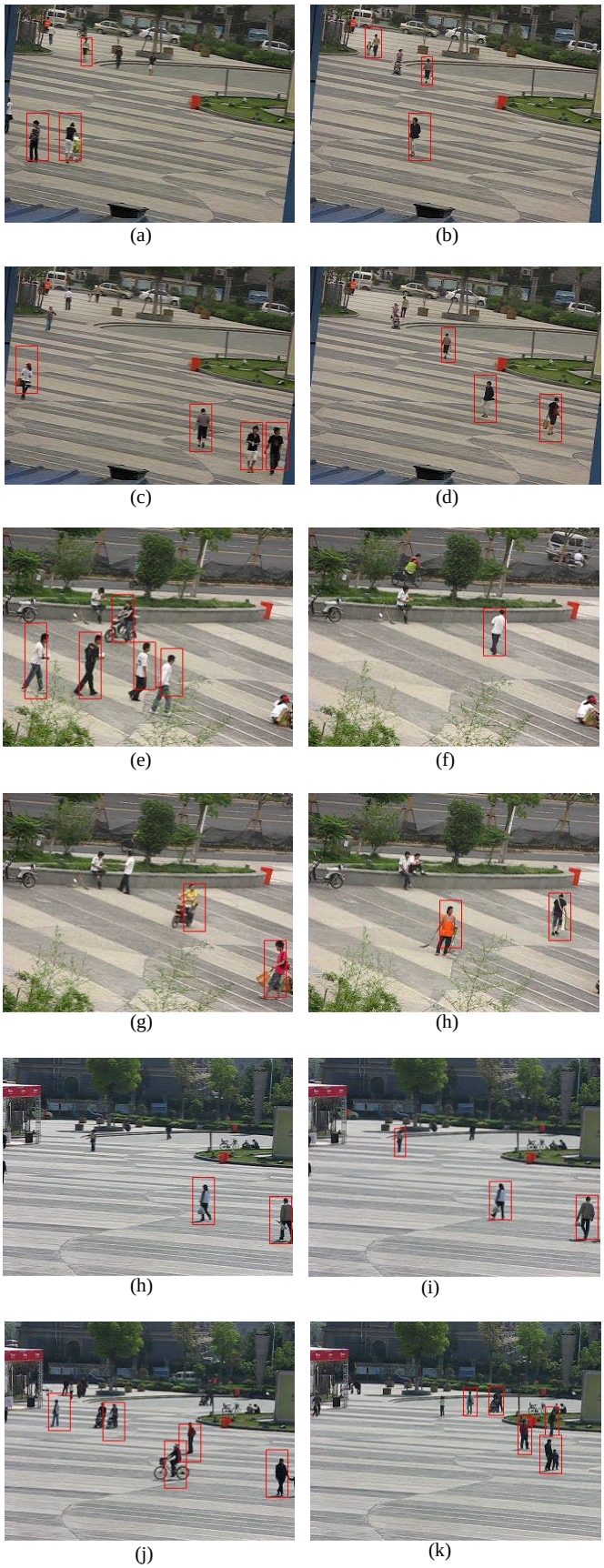


Figure 7. Human Detection Results

REFERENCES

[1] Roland Perko,Ales Leonardis, "A framework for visual-context-aware object detection in still images," *Computer Vision and Image Understanding*, vol.114,pp. 700-711,June 2010.

[2] Dashan Gao,Sunyoung Han, "Discriminate saliency, the detection of suspicious coincidences, and application to visual recognition," *IEEE Transaction on PAMI*,vol.31,pp.989-1003, June 2009.

[3] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Conference on Computer Vision and Pattern Recognition, San Diego*,vol.1,pp,886-893,June 2005.

[4] P.Felzenszwalb,D.McAllester, "A discriminatively trained, multi-scale, deformable part model", *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage* ,pp.112-120,June 2008.

[5] C.H.Lampert,M.B.Blaschko,and T.Hofmann, "Beyond sliding windows: Object localization by efficient sub window search," *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage* ,pp.1-8,June 2008.

[6] A.Mohan,C.Papageorgiou, and T.Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol.23,pp.349-361,April 2001.

[7] K.Okuma,A.Taleghani,N.de Freitas,J,Little,and D.Lowe, "A boosted particle filter: Multi-target detection and tracking," *Pro.of the European Conference on Computer Vision(ECCV)*,pp.28-39,2004.

[8] V.D.Shet,J.Neumann,V.Ramesh, and L.S.Davis, "Bilattice-based logical reasoning for human detection," *Proc. of the International Conference on Computer Vision and Pattern Recognition*, pp. 1-8,June 2007.

[9] D.Gavrila, "Pedestrian detection from a moving vehicle," *European Conference on Computer Vision,Ireland*,pp.37-49,June 2000.

[10] D.Lowe, "Object recognition from local scale-invariant features," *Proceedings of the International Conference on Computer Vision*, vol.2, pp.1150-1157, 1999.

[11] K.Mikolajczyk C.Schmid and A.Zisserman, "Human detection based on a probabilistic assembly of robust part detectors, *"Proceeding of the European Conference on Computer Vision*, pp.69-82,2004.

[12] Qiang Zhu,Shai Avidan, "Fast human detection using cascade of histograms of oriented gradients," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,2006.

[13] Xiaoyu Wang, "An HOG-LBP human detector with partial occlusion handling," *IEEE International Conference on Computer Vision* ,2009.

[14] Sebastian Montabone and Alvaro soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image and Vision Computing*, vol.28, pp.391-402, 2010.

[15] Michael Oren and Constantine, "Papageorgiou Pedestrian detection using wavelet templates," *Computer Vision and Image Understanding*, pp.193-199,1997.

[16] Mohan A,PapageorgiouC, "Example-based object detection in images by components," *IEEE trans on PAMI* ,vol.23,pp.349-361,2001.

[17] B.Wu and R.Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.951-958, 2006.

[18] Viola P,Jones MJ, "Detection pedestrian using patterns of motion and appearance,"vol.63,pp.153-161,2005.
[19] V.N. Vapnik, " The nature of statistical learning theory," Springer, 1995.
[20] Y.Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Proc. of the European Conference on Computational Learning Theory*, pp.23-37, 1995.
[21] Simon, L.; Tarel , "Alerting the drivers about road signs with poor visual saliency," *IEEE Intelligent Vehicles Symposium* ,2009.
[22] Xin-Zheng Wang, Xiao-chen Duan, "Application of Neural Network in the Cost Estimation of Highway Engineering," *Journal of computers*,vol.5,pp.1755-1761,November,2010.
[23] Long Wang, Yanheng Liu, Xiaoguang Li, "Analog Circuit Fault Diagnosis Based on Distributed Neural Network," *Journal of computers*,vol.5,pp.1747-1754,November,2010.
[24] Papageorgiou and Poggio, " A trainable system for object detection," International Journal of computer vision,vol.38,pp.15-33,2000.
[25] K. Michalak, H. Kwasnicka, " Correlation based feature selection method," *International Journal of Bio-Inspired Computation*, vol.2,pp. 319-332, 2010.
[26] C. Henry, J.F. Peters, "Perceptual image analysis," *International Journal of Bio-Inspired Computation*, vol.2,pp. 271-281, 2010.
[26] Hou Beiping, Zhu Wen, "Moving target classification based on shape features from real-time video," *Chinese Journal of Scientific Instrument*, vol. 31, pp. 1819-1825, August, 2010.

Hou Beiping received his PhD degree from Zhejiang University, Hangzhou, China, in 2005. He is now an associate professor in Zhejiang University of Science and Technology, Hangzhou, China. His research interests include image processing, pattern recognition, and machine vision.

Zhu Wen received her Master degree from Tianjin University of Science and Technology, Tianjin, China, in 2003. She is now a full time lecturer in Zhejiang University of Science and Technology, Hangzhou, China. Her research interests include image processing and Intelligent Control.