Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **1**, January **2018**

A Pedestrian Detection Method Using SVM and CNN Multistage Classification

Yong-qiang He, Qin Qin

School of Computer Henan University of Engineering
W Henan Zhengzhou 451191 China
heyongqiangxz@163.com

Vychodil Josef

Department of Radio Electronics, Brno University of Technology,
Technicka 3082/12,616 00 Brno, Czech Republic.
vychod@phd.feec.vutbr.cz

Received January, 2017; revised July, 2017

Abstract. *Facing pedestrian detection applications for monitoring scenarios, a pedestrian detection method using support vector machines and convolutional neural networks is proposed. First, it uses motion detection method to locate interested area of suspicious targets; then calculates gray level co-occurrence matrix of image patches of these areas, uses principal component analysis method to extract textural feature vector, and uses support vector machines to classify textures for filtering out interference regions; finally, constructs multi-scale image blocks on the remaining area, and uses LeNet5 architecture of convolutional neural network to execute pedestrian classification. Experimental results on Caltech dataset show that, this method has high true positive rate, low false positive rate, and little average detection time.*

Keywords: Pedestrian detection; Motion detection; Gray level co-occurrence matrix; Support vector machines; Convolutional neural networks.

1. **Introduction.** With the development of video surveillance technology, the intelligent analysis demand of surveillance video is more and more urgent. Pedestrian is an important object of video surveillance. So, the application of pedestrian detection technology is very important in the field of video surveillance needs [1]. Pedestrian detection can automatically detect the pedestrian in the video or image based on video image analysis and processing technology. It has been a research hotspot in the field of computer vision [2]. There are many academic researches in recent years in the related fields. In the literature [3], the improved Haar-like feature was used to describe the pedestrian and the Adaboost classifier was used to describe the pedestrian classification. In literature [4], the fusion of the underlying characteristics and multi-level guide edge energy feature was completed based on the linear weight fusion principle and the histogram intersection kernel support vector machine classification, which was used to improve the robustness of the pedestrian detection system. The literature [5] proposed a simulation of the human eye to observe things divergent and significant features of the texture structure operator, which can effectively improve the robustness of the pedestrian recognition of noise and illumination changes. The literature [6-7] extracted the Histogram of Oriented Gradients people (HOG) features and used the support vector machine (SVM) classifier for pedestrian classification. Based on the HOG and SVM, the pedestrian detection efficiency was improved

by using the background area with the motion detection algorithm filter. The method deep learning pedestrian was used by the authors of literature [8-12] to improve the accuracy of pedestrian detection. In above all, the existing pedestrian detection methods have high detection accuracy on INRIA and other pedestrian image data sets. However, for the application of video surveillance, the operation efficiency of the pedestrian detection method is very low. The detection accuracy of some efficient pedestrian detection method is lower. Especially, the false positive rate index is very high. It is difficult to satisfy the requirements of the practical application of video surveillance technology. According to the pedestrian detection applications requirements in monitoring scene, a method combining support vector machine and convolutional neural network (Convolutional Neural Networks, CNN) method was proposed in this paper for pedestrian detection. The basic idea of the proposed algorithm is the combination of motion detection, texture classification, gray level co-occurrence matrix and SVM, as well as the multi-level classification of image block pedestrian classification for realizing the efficient and reliable pedestrian detection.

2. The proposed algorithm.

The implementation process is shown in Figure 1. The motion detection would be completed for each video frame at first. The target suspicious regions would be located preliminary. Then, gray co-occurrence matrix of the suspicious target image would be calculated. The principal component analysis (PCA) method is used for feature reduction and to get the suspicious target image block texture. The SVM classifier is used to filter out the obvious interference targets. Then, the multiscale image block would be constructed in the suspicious area of remaining construction. The CNN method was used for pedestrian classification and to record the pedestrian target classification window. The detailed process description is as follows.

2.1. Motion detection.

Motion detection can be used to locate the region of interest in video image and improve the efficiency of video analysis, as well as filter the interference. The traditional motion detection methods include background subtraction method, frame difference method and optical flow method. In this paper, the background subtraction method is used to locate the suspicious pedestrian target area, considering the low efficiency of the optical flow method and the frame difference method. In particular, the VIBE algorithm [13-14] is the basis. Because the algorithm is efficient, and the detection of the target integrity is good. However, when the pedestrian is close to the background color, there is still a phenomenon that the pedestrian target is incomplete after the motion detection[15-17]. In order to avoid subsequent undetected targets of pedestrians, the target area after motion detection is extended to ensure pedestrian areas as complete as possible after motion detection. In detail, for any suspicious target after motion detection, the external rectangle is denoted as (x_r, y_r, z_r). Here, x_r, y_r denoted the top left corner coordinates of the external rectangle. w_r and h_r denote the width and height of the outer rectangle respectively. The extended outer rectangle (x'_r, y'_r, w'_r, h'_r) can be obtained by the formula (1).

Table 1 shows the PSNR and GMG of the infrared simulated blurry image in the blurred image restoration experiment. The GMG value is calculated in the simulation experiment so as to make a reference to the real fuzzy image restoration experiment. The number of RL iterations is 15 times.

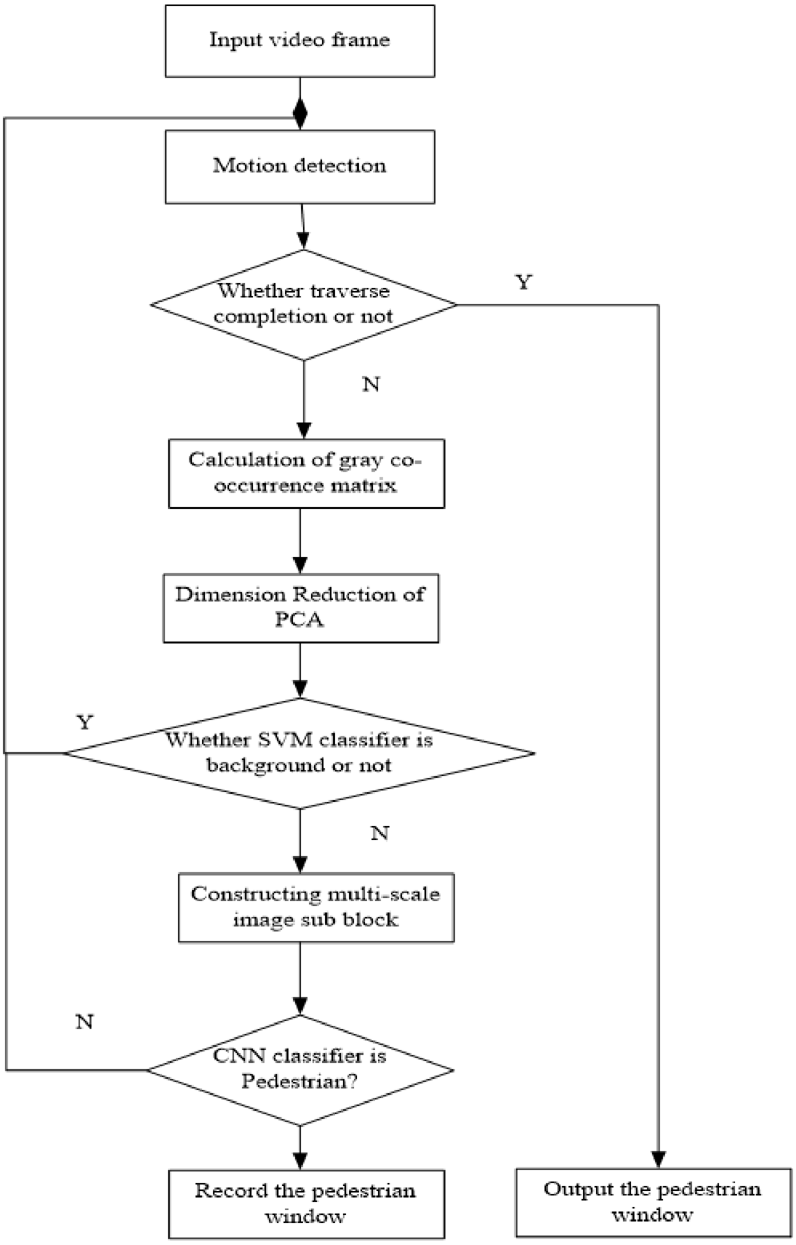
Input video frame
Motion detection
Whether traverse completion or not
Y
N
Calculation of gray co-occurrence matrix
Dimension Reduction of PCA
Y
Whether SVM classifier is background or not
N
Constructing multi-scale image sub block
N
CNN classifier is Pedestrian?
Record the pedestrian window
Output the pedestrian window

Figure 1. The implementation process

$$\begin{cases} {x'}_r = \max\left(x_r - \lambda, 0\right) \\ {y'}_r = \max\left(y_r - \lambda, 0\right) \\ {w'}_r = \min\left(w_r + 2\lambda, W_{ori} - 1 - {x'}_r\right) \\ {h'}_r = \min\left(h_r + 2\lambda, H_{ori} - 1 - {y'}_r\right) \end{cases} \tag{1}$$

Here, W_{ori} and H_{ori} represent the original video image width and height. The max and min represent the maximum and minimum calculation value respectively, which are used to prevent the outside rectangle boundary. λ is the expansion scale factor. There is a big difference in the imaging scale due to the different target distance from the camera at the same time. Therefore, the scale factor should be adaptively changed with the different scale of the target image. In this paper, according to the detected target width and height,

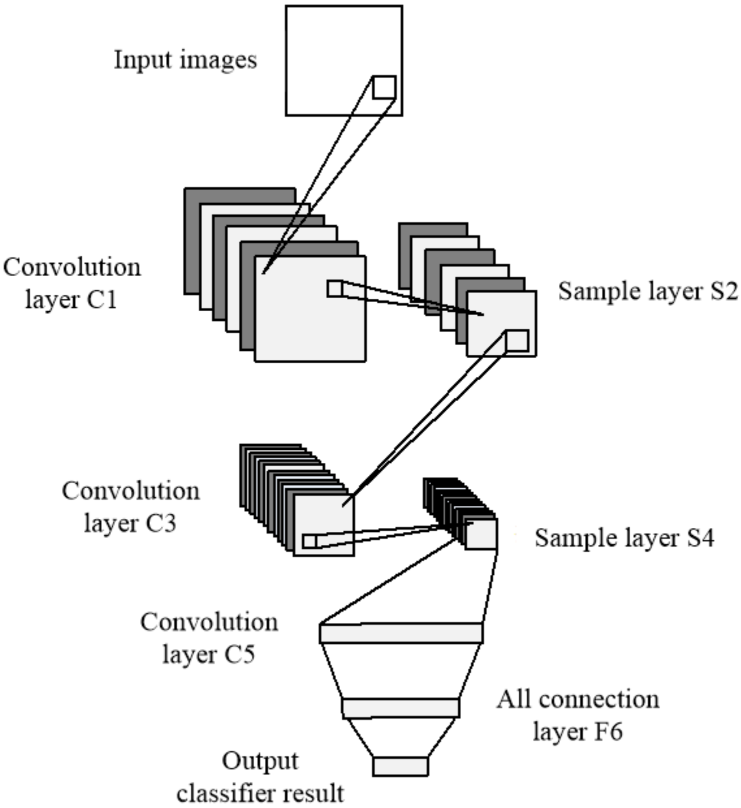

FIGURE 2. Diagram of imaging system

the scale factor is derived as following:

$$\lambda = \mathrm{Int}\left(k * \sqrt{{w_r}^2 + {h_r}^2}\right) \tag{2}$$

Here, is the rounding operation. Symbol k is the ratio coefficient, which is 0.1. After the external rectangle box, the image block is cut from the original video image to the corresponding position of the external rectangle frame. Then each image block is traversed for subsequent pedestrian detection.

2.2. **Extraction of gray level co-occurrence matrix and SVM coarse screening.** In general, there is a background region in the target area of motion detection, which is caused by the change of light and shade. The texture complexity of these regions is obviously lower than that of the pedestrian area. Based on this phenomenon, this paper uses regional coarse screening method for texture classification of this suspicious target, the background area with low texture complexity would be removed quickly, to not only improve the finishing efficiency of the algorithm, but also reduce the false alarm phenomenon caused by these areas. Gray level co-occurrence matrix (GLCM) is a common feature to describe the image texture distribution, which can describe the gray distribution of the pixels on the image and can effectively distinguish the different texture images. Here, the pixel has to satisfy certain conditions. Such as, the pixel (x, y), horizontal distance a and vertical distance b consist the pixel $(x + a, y + b)$ .The gray level of (x, y) is g_1 . The gray level of$(x + a, y + b)$ is . The gray level of (x, y) and $(x + a, y + b)$ is . Each pixel corresponds to a gray level distribution . L is the image gray level. The value combination range of (g_1, g_2) is 0 . After the number statistics of each class (g_1, g_2) combinations in the statistical image, the total number of pixels in the image would be normalized and the probability of occurrence of the combination is shown as formula (3).

$$p(g_1, g_2 \,|a, b\,) = \frac{n(g_1, g_2 \,|a, b\,)}{N} \tag{3}$$

Here, is number of gray distribution(g_1, g_2) in image. N is the number of pixels. The gray level co-occurrence matrix is consisted of $p(g_1, g_2 \,|a, b)$ and $L \times L$ matrix. The gray level co-occurrence matrix is used to filter the background region with low texture complexity. Therefore, this paper focuses on the complexity of the image texture, but not the direction of the texture and the period distribution. Therefore, in this paper, the combination value of (a, b) commonly are $0^o, 45°90°$ and 135° in four directions. Here, the occurrence probability of each combination takes the average value of four directions, as shown in formula (4).

$$\begin{aligned} p'(g_1, g_2) = \frac{1}{4 \times N}\, & (n(g_1, g_2 \,|1, 0\,) + n(g_1, g_2 \,|0, 1\,) \\ & + n(g_1, g_2 \,|1, 1\,) + n(g_1, g_2 \,|{-1}, -1\,)) \end{aligned} \tag{4}$$

In this way, each suspicious image region can be used as the texture feature of the image. Usually, the gray level of the image L is 256. So, the dimension of the gray co-occurrence matrix of the suspicious image region is 65536. Obviously, the storage and classification of the large dimension will consume a lot of resources. Hence, this paper uses PCA method [10] to reduce the dimension of texture feature, and reduce the dimension of texture feature to 80. Then, the trained SVM classifier is used to classify the feature vectors after dimensionality reduction. If the classification result of the texture feature vector of the suspicious image region is the background class, this region would be removed. For the remaining suspicious image region, the final pedestrian detection results are obtained by CNN classification. Here, the training process of classifier is introduced in the experiment.

2.3. **Multi scale image sub block construction and CNN classification.** In this paper, the CNN method is used to classify pedestrians in the area of the remaining suspicious images. Details are as follows. First of all, considering that the distance between the target and the camera is different, the scale of the image is different. In this paper, we construct a multi-scale image sub block. The specific steps are shown as follows. Bilinear interpolation is used to normalize the size of image sub block $W' * H'$ Output: Image sub block set Here, the value of $W_0 * H_0$ and $W' * H'$ is $21 * 41$ Then, each image sub block is classified by CNN classifier. This paper uses the CNN architecture for the LeNet5 network architecture [11], as shown in figure 2.

The CNN classifier is used to classify the image sub blocks for pedestrians, and the external rectangle window on the original video image corresponding to the image sub block is recorded. Due to the overlapping phenomenon of the multi-scale image sub block, the same pedestrian target may be detected on the multi-scale image sub blocks. Therefore, it is necessary to filter the pedestrian window detected by the multi-scale image sub block and merge the pedestrian window belonging to the same target. Specifically, for the same video images of all pedestrian windows, the coincidence degree between 22 is calculated, that is, the ratio of the area of the cross section of the rectangular cross section of the two pedestrian windows to the total area. If the coincidence degree of the two outer rectangular frames is greater than 50%, the pedestrian window is considered to be from the same target, and the two pedestrian windows are merged at the same time, and the outer rectangular frame of the combined pedestrian window is represented as a

```
Input: Suspicious image area I Size: $W * H$
Progress:
Initialization: Initial scan window: $W_0 * H_0$
Scaling factor:$s$;
Target image sub block size: $W' * H'$ ;
for $(h = H_0; h < H; h = h * s\ )$ do
    for $(w = W_0; w < W; w = w * s\ )$ do
        for $(y = 0; y < H - h; y = y + 2\ )$ do
            for $(x = 0; x < W - w; x = x + 2\ )$ do
                Cropped image sub block$w * h$ at $(x, y)$point in I;
            end for
        end for
    end for
end for
```

window:

$$\begin{cases} x_{ij} = \dfrac{x_i + x_j}{2} \\ y_{ij} = \dfrac{y_i + y_j}{2} \\ w_{ij} = \dfrac{w_i + w_j}{2} \\ h_{ij} = \dfrac{h_i + h_j}{2} \end{cases} \tag{5}$$

Here,

$$(x_i, y_i, w_i, h_i)$$

and

$$(x_j, y_j, w_j, h_j)$$

denote the external rectangle of two pedestrian window before combination. The two pedestrian windows would be merged. The above window merge process would be repeated, until degree of overlap of all the windows is less than 50

3. Simulation.

3.1. Experimental description.

In this paper, the method is mainly used in the detection of pedestrian detection in the field of pedestrian detection in the field of experimental data set. The Caltech data set is constructed by the actual monitoring site video. The video resolution is 640 * 480. The frame rate is 30fps, containing about 250000 frames. All pedestrians are marked. In this paper, the data set is used to test the performance of this method. This paper deals with two classifiers. One is the SVM classifier which is needed by texture classification. The other is the CNN classifier which is required by the pedestrian classification. In order to train two classifiers, the INRIA data field of pedestrian detection sets is used as training samples, which contains 5264 images of the data set. This set contains 3548 images of pedestrians and 1716 images that do not contain pedestrians. In the training of the texture classifier, the positive sample selects the INRIA data set which contains the pedestrian image. The negative sample is obtained from the INRIA data set which contains no pedestrians. The rules of cutting are: to intercept the image sub block of each image in the area of the gray change, such as the sky area, the

Table 1. Parameter settings of computer

content	parameter
CPU	Intel I5 3.2GHZ
Memory	16G
OS	Windows 7 64
Software	Visual Studio 2013
Dependency Library	OpenCV 2.4.11

ground area, etc.. The image entropy can be calculated by the gray level co-occurrence matrix, which is used to describe the complexity of the image texture

$$S = -\sum_{g_2=0}^{255}\sum_{g_1=0}^{255} p'(g_1, g_2) lgp'(g_1, g_2) \tag{6}$$

Here,

$$p'(g_1, g_2)$$

is obtained from formula (4). The smaller the entropy of the image, the smaller the complexity of the image texture, that is, the change of the gray level of the image is gentler. In accordance with the above rules, this paper randomly intercepted 7176 smooth images from the INRIA data set which does not contain the pedestrian images. In training the pedestrian classifier, the pedestrian image contains positive samples from INRIA data, the sample is negative will not contain a pedestrian image of four equal parts to get the INRIA data set. This negative sample image set a total of 6864. Then, the size of positive and negative sample images is normalized to $21 * 41$. In order to evaluate the performance of the method in this paper, we analyzed the real rate (True Positive, Rate, TPR), false positive rate (False Positive Rate FPR (Average) time and average detection of Detection Time, ADT).

$$\text{TPR} = \frac{\text{Number of correct detection}}{\text{Total number of pedestrian markers}} \times 100\% \tag{7}$$

$$\text{FPR} = \frac{\text{Number of error detection line}}{\text{Number of error detection line} + \text{Number of correct detection line}} \times 100 \tag{8}$$

$$\text{ADT} = \frac{\text{Detection time}}{\text{Number of Video frames}} \tag{9}$$

(9) Here, the correct detection of pedestrian refers to the detection of pedestrian window and marked pedestrian window two external rectangular box coincidence degree greater than 50%. The detection time is the total time spent on all the video images for pedestrian detection, but it does not include the time taken for video decoding and classifier training. The performance parameters of the computational platform used in the simulation experiment are shown in Table 1.

3.2. **Classifier training.** In this paper, the texture classifier is trained by SVM learning method. In particular, the gray level co-occurrence matrix of positive and negative sample images is calculated according to formula (4), and then the texture feature vector of dimension 80 is obtained by PCA method. Finally, the SVM method is used to train the texture feature vector of positive and negative samples, and the SVM classifier is obtained. Among them, the SVM kernel function is based on radial basis function. Detailed training process can refer to [7]. In this paper, the pedestrian classifier is trained by CNN learning method, using the network structure shown in Figure 2, detailed training process and parameter configuration can refer to [11].

3.3. **Performance evaluation.** In order to evaluate the performance of the method in this paper, the proposed pedestrian detection method with [6], [7] and [8] in the experiment would be evaluated based on the test data by using the four test set of pedestrian detection. But the training data set is slightly different, with the other three kinds of pedestrian detection method of training data set is the original INRIA data set, and the method in the training of two classifiers on INRIA data set was cut out and normalized. Figure 3 shows the results of the real and false positive rates of the four pedestrian detection methods. It is obvious that the real rate of this method is higher than the other three methods, and the false positive rate is more obvious. This is because this method adopts multi-stage filtering and classification strategy in motion detection and texture classification has two stages filtering many interference areas. These areas may reduce the false detection caused by pedestrian phenomenon, thereby significantly reducing the false positive rate index. At the same time, the image region pedestrian classification is far less than the original image, the multi-scale image multi-scale image sub block parameters are obtained under the same construction method in this paper block so closer to the pedestrian target, better classification performance than SVM classical classifier plus LeNet5 network architecture, so the real rate of this method there are also improved.

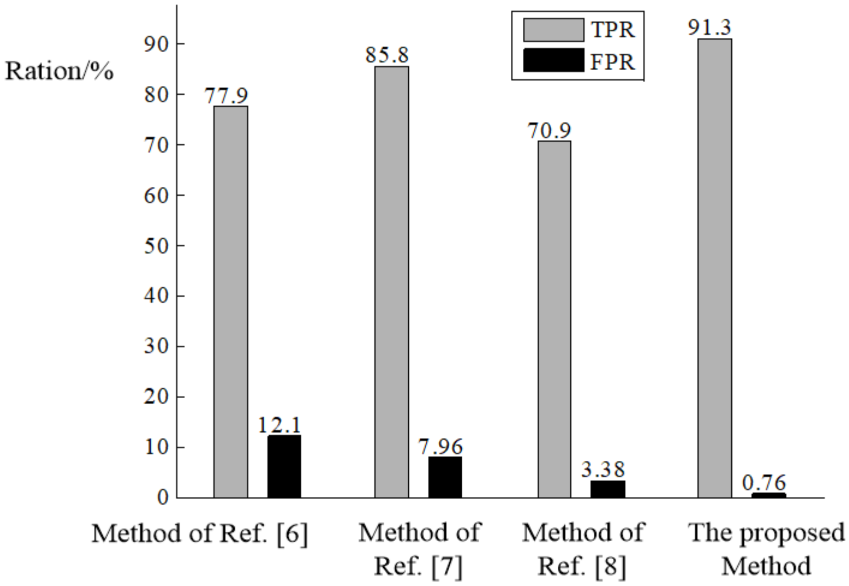

FIGURE 3. Comparison of true and false positive rates

Table 2 shows the average detection time of the four pedestrian detection methods. It can be seen that the average detection time of this method is the smallest, which is much less than that of [6] and [7]. This is because the method described in this paper and the [8] method in the literature before the pedestrian classification are filtering out a lot of interference areas in the image, reducing the detection time. Compared with the literature [8], this paper uses the motion detection and texture classification two levels of coarse screening, in most of the monitoring scenarios can be filtered more interference area, reducing the average detection time. Based on the above test results and analysis, the pedestrian detection performance of this method is better than the other three methods.

Table 2. Comparison of average detection time

Method	ADT/ms
Ref. [6]	819
Ref. [7]	1197
Ref. [8]	81
The proposed one	77

4. **Conclusions.** Pedestrian detection has a wide range of applications in the field of video surveillance. Aiming at the application environment, this paper proposed a pedestrian detection method based on support vector machine and convolution neural network. The design idea is as follows: by using the method of motion detection and texture classification method of high operation efficiency, the background area of exist pedestrians would be removed. Then in the rest of the region,the fine detection result reliable would be obtain by using the the pedestrian classification method. Here, based on the VIBE method of motion detection, texture classification presents a symbiosis matrix, principal component analysis and support vector machine classification method with gray texture.The pedestrian classification is realized in multiscale image block using LeNet5 convolutional neural network architecture. The experimental results show that the proposed method can quickly and reliably detect the pedestrian targets on the Caltech data set, which is an effective pedestrian detection method for video surveillance applications

REFERENCES

[1] V. D. Hoang, M. H. Le, K. H. Jo, Hybrid cascade boosting machine using variant scale blocks based HOG features for pedestrian detection *Journal of Neurocomputing*, 135, no. 8, pp. 357-366 , 2014.

[2] P. Dollar, C. Wojek, B. Schiele, et al. Pedestrian detection: An evaluation of the state of the art *Journal of Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34, no. 4, pp. 743-761, 2012.

[3] S. Zhang, C. Bauckhage, A. B. Cremers, Informed Haar-Like Features Improve Pedestrian Detection , *Pro. of IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington, DC, USA*, 2014:947-954.

[4] R. Sun, N. Hou, J. Chen, Fast pedestrian detection method based on features fusion and intersection Kernel SVM *Journal of Opto-Electronic Engineering*, no. 2, pp. 53-62, 2014.

[5] D. G. Xiao, C. Xin, T. Zhang, H. Zhuan, X. L. Li, Saliency Texture Structure Descriptor and its Application in Pedestrian Detection *Journal of Ruan Jian Xue Bao / Journal of Software*, 25, no. 3, pp. 675-689, 2014.

[6] P. Yadav R, K. Kutty, S. P. Ugale Implementation of Robust HOG-SVM based Pedestrian Classification *Journal of International Journal of Computer Applications,* , 114, no. 19, pp. 10-16, 2015.

[7] C. Lu, G. Teng, X. Zou, et.al, Rapid pedestrian detection algorithm based on foreground *Journal of Video application and project*, 39, no. 1, pp. 113-116, 2015.

[8] Y. Tian, P. Luo, X. Wang, et al., Pedestrian detection aided by deep learning semantic tasks, *Pro. of Computer Vision and Pattern Recognition. IEEE*, pp. 5079-5087, 2014.

[9] T. Kryjak, M. Gorgon, Real-time implementation of the ViBe foreground object segmentation algorithm, *Pro. of Computer Science and Information Systems*, pp. 591-596, 2013.

[10] L. Kuncheva, Faithfull W J. PCA feature extraction for change detection in multidimensional unlabeled data., *Journal of IEEE Transactions on Neural Networks and Learning Systems,* 25, no. 1, pp. 69-80, 2014.

[11] Z. Yang, D. Tao, S. Zhang, et al., Similar handwritten Chinese character recognition based on deep neural networks with big data, *Journal of J. Commun,* 35, no. 9, pp. 184-189, 2014.

[12] R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, *Pro. of Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014.

[13] J. Hosang, M. Omran, R. Benenson, et al., Taking a deeper look at pedestrians, *Pro. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4073-4082, 2015.

[14] Y. Tian, P. Luo, X. Wang, et al., Pedestrian detection aided by deep learning semantic tasks, *Pro. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5079-5087, 2015.

[15] B. Hariharan, P. Arbelez, R. Girshick, et al., Hypercolumns for object segmentation and fine-grained localization, *Pro. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 447-456, 2015.

[16] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761-769, 2016.

[17] R. Girshick, J. Donahue, T. Darrell, et al. Region-based convolutional networks for accurate object detection and segmentation, *Journal of IEEE transactions on pattern analysis and machine intelligence*, 38, no. 1, pp. 142-158, 2016