International Journal of Smart Home
Vol. 10, No. 11, (2016), pp. 197-206
http://dx.doi.org/10.14257/ijsh.2016.10.11.17

Pedestrian Detection Method Based on Convolution Nerve Network

Jiang Yingjun[1], Wang Jianxin[1] and Guo Kehua[1]

[1]*School of Information Science and Engineering, Central South University, Changsha, Hunan 410083*
381959813@qq.com

Abstract

One new pedestrian detection method integrating static high-level features and movement features based on convolution nerve network is proposed in this paper. During the phase of unsupervised deep learning of pedestrian features, the hierarchical static features of pedestrians are extracted from the low to the high with convolution nerve network; the pedestrian movement features are obtained through mean value approach of rectangular block pixel difference. During the logic regression recognition phase, static features and movement features are integrated. The results show that the pedestrian detection algorithm of convolution nerve network integrating movement features greatly improve the pedestrian detection performance under complicated background.

Keywords: *Pedestrian detection; Convolution nerve network; Unsupervised learning; Hierarchical features; Movement features*

1. Introduction

Pedestrian detection plays very crucial role in areas like security monitoring, safe traveling of vehicles and robots. With advanced pedestrian detection technology, auxiliary automobile driving system is able to reduce effectively the collision between automobiles and pedestrians, the security monitoring system is able to notice timely the security people of the suspicious people in the monitoring area and help the security people locate and track rapidly the criminal suspects [1-3]. Remote care system make timely warning to the elderly and children by analyzing the occurrence law and abnormal action of the walking people in the monitoring area based on the pedestrian detection technology. Besides, the pedestrian detection is also the basis for technologies like gait recognition of pedestrians [4].

In recent years, the pedestrian detection technology has been improved greatly centering on two objectives which are improvement of detection precision and detection speed. Dalal *et. al.*, based on INRIA [5] pedestrian database, adopt the histogram of oriented gradient technology and get the detection rate of 80% when the number of each figure fault positive instances (FPPI) is 1; Sun *et. al.*, get the detection rate of 86% based on the significance test and in combination with histogram of oriented gradient. However, when these algorithms are applied to special occasions which are different from training scenarios, the accurate detection rate decreases rapidly and the wrong detection rate increase rapidly. One main reason for the wrong detection is: the integral channel features extracted from pedestrian detection technology, oriented gradient features and color features are the hand-tailored low-level image features; while in images or videos with complicated background, it is easy for the local pixels to produce the distribution similar to the low-level features of the pedestrians and be wrongly detected as pedestrians. The object recognition based on deep learning imitates the hierarchical recognition process of human brains for the objects, takes the low-level features as the basis through multi-layer learning and without knowing the specific low-level features (not specified) and make extraction from the low to the high so as to realize the automatic recognition of the multiple kinds of target objects.

Currently, the models used for pedestrian detection based on deep learning theory are mainly the convolution neural networks (CNNs) and the expansion based on convolution neural network [7-10]. Pierre, *et. al.,* [10] get the better detection result with average miss rate being 10.55% on the INRA database based on deep learning theory, with the CNNs and through the unsupervised monitoring and training. Wanli [8] *et. al.,* use the hierarchical CNNs to integrate the four key steps usually treated individually which are feature extraction, deformation handling, occlusion handling and classification into one whole learning frame so as to firstly extract the overall features and then perform part detection and part feature mapping for the pedestrians. With this approach, the average miss rate on the Caltech database with foreground and background changing dynamically is 30%. Zsolt [9] uses the CNNs to extract separately the apparent characteristics and parallax characteristics of the pedestrians for the video sequence under different environment conditions. The data source of apparent characteristics is the videos acquired by monocular camera while the data source of parallax characteristics is obtained by stereoscopic camera. These two kinds of characteristics are extracted by different channels and integrated in the categorization phase. Since CNNs keep sampling during the process of forming high level features from low level ones, this algorithm is not prominent for foreground and it is difficult to recognize the pedestrians of medium and long distance when the height is lower than 50 pixels and the wrong detection rate is high.

The current detection algorithm in combination with movement characteristics of the pedestrians mainly performs the movement target detection and separates the movement foreground and background and then perform pedestrian detection in the movement foreground area. The movement target detection usually includes the following two categories: the first category is background deduction approach, which is suitable for movement target detection environment with fixed camera and little background change. Usually mixed Gauss model is adopted to realize background modeling and then movement area is obtained with background deduction. Shape, color, gradient and other features are extracted in the movement area to perform pedestrian detection [11-12]. However, the background deduction is very sensitive to brightness change and unordered scenario. Besides, when the cloth color of pedestrians is similar to the background color and the video resolution is low, the background deduction will cause lots of wrong separation, *i.e.,* the first pedestrian movement area is divided into several areas not connected, which leads to wrong detection and detection missing [13]. The second type of movement target detection method is light stream approach which is suitable for movement target detection environment with moving camera and continuously changing foreground and background. The light stream approach takes the light stream filed to represent the object movement filed, calculates the variation of gray (illumination) of image pixels of adjacent frames to determine the movement area in the image and then detects the pedestrian in the movement area [14-18]. Walk *et. al.,* [19] adopt improved light stream histogram approach and greatly improve the pedestrian detection precision with the pedestrian movement information and in combination with apparent features like gradient and self-symmetry of color. The pedestrian detection rate is 78% on the INRIA pedestrian database when FPPI is 0.1. However, the light stream approach uses iterative method to calculate the light stream filed which takes long time and it is difficult to realize the pedestrian detection. Besides, the light stream approach is sensitive to noise and light, so it is not suitable for pedestrian detection with low quality of images [20-21].

The study in this paper focuses on the realization of real time pedestrian detection with low miss rate under bad light and complicated background situation for monocular monitoring video of medium and low resolution. Figure 1, shows the learning training model adopted in this paper (CNNs+ Motion). During the learning period, we extract two kinds of features of pedestrians: as for single image, automatically coded sparse convolution nerve network is used to extract the static features of pedestrians in layers from the low level to high level; as for series images, mean value of rectangular block pixel difference (RBPVD) is adopted to summarize the law of change between current frame and preorder frame for the pedestrian rectangular block to be used as the movement features particular to pedestrians. During the training period, the static and dynamic characteristics of pedestrians are integrated and linear logic regression suitable for categorization of two kinds of objects is used to identify the pedestrians and non-pedestrians with high precision.

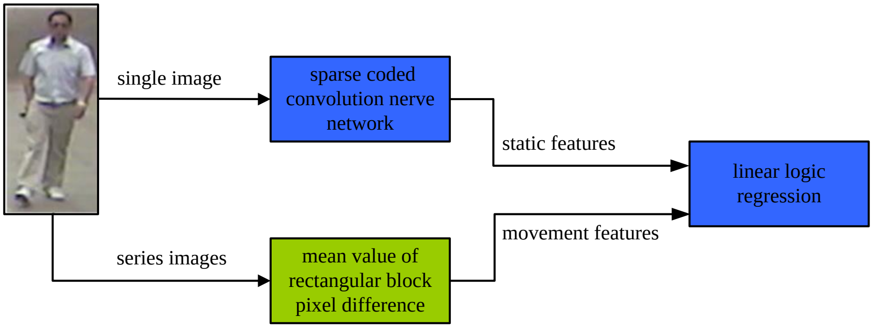

single image
sparse coded convolution nerve network
static features
linear logic regression
movement features
series images
mean value of rectangular block pixel difference

Figure 1. CNNs+ Motion Learning Training Model

2. Deep Learning of Pedestrian Characteristics

2.1. Convolution Nerve Network

Currently, great amount of pedestrian detection study focuses on the design of strong representative pedestrian features. In order to get such features effectively, we have generated successfully the high-level pedestrian features with deep learning algorithm of convolution nerve network with the precondition that the pedestrian target images are basically recognizable so as to get better pedestrian detection level. The learning model based on convolution nerve network adopts strict feedforward mode: when current input is only the output of previous layer and it is irrelevant to the output of other layers. With a series of convolution and sub-sampling (pooling), not only are the global features like pedestrian shape and structure are produced, but also the local features of pedestrians are extracted. The convolution nerve network forms the mechanism based on human visual sense to perform multi-layer feature extraction through local receptive filed, weight sharing and space sampling and keep the invariance of scale and displacement.

The convolution nerve network in this paper utilizes the LeNet-5 [22-23] structure, including two convolution layers, two non-linear transformation layers and one fully-connected layer of convolution nerve network. As shown in Figure 2, after the pedestrian sample image is input to the convolution nerve network and with feature extraction of convolution layer and non-linear change of sampling layer, the bottom boundary features are learned from the original single pixel features, and middle-level pedestrian part features are learned from bottom features and high-level overall pedestrian features from middle-level pedestrian part features. The first convolution layer consists of six 5X5 nucleus. With the input sample image convolution of 64X32, six 60X28 feature mapping diagrams are produced and the total feature number is 150. What the convolution layer extracts is the boundary characteristics of sample images. The second convolution layer includes 60 5X5nucleus, and it performs the convolution for the previous six feature diagrams after sampling and pooling and produces 16 10X10 feature mapping diagrams. The total feature number is 1668. In this convolution layer, different connection mechanism is adopted to keep the connection quantity within reasonable scope and get richer characteristics by damaging the network symmetry. The third convolution layer includes 16 13X5 convolution nucleus and it performs convolution for the previous 16 feature diagrams after sampling and pooling and generates one 3796 dimension of feature vector as the input of supervised monitoring training. It is unnecessary for the convolution nerve network without supervised learning to label the learned positive sample and negative sample.

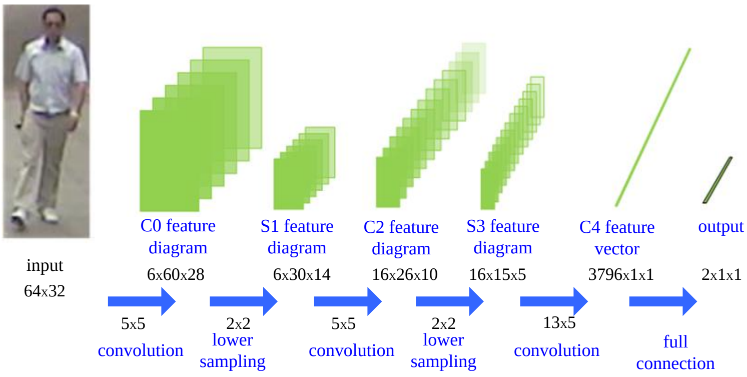


Figure 2. Hierarchical Feature Extraction

2.2. Non-Linear Transformation and Pooling

Non-linear transformation and pooling consist of absolute value correction, local contrast normalization and sampling with averaging method. The absolute value correction takes the absolute value of feature value obtained from previous convolution phase so as to avoid cancellation problem during the later two steps during non-linear transformation phase. Local contrast normalization is realized through subtraction and division of local features. The subtraction operation is deduct the feature value and neighborhood weight from certain local feature value, as shown in Formula (1), based on which the feature value after weighting is divided by the mean value of all weighted feature values, as shown in Formula (3). Local contrast normalization makes the active feature value more prominent while restricting the feature value of small amplitude.

$$v_i = x_i - x_i \otimes w \tag{1}$$

$$\sigma = \sqrt{\sum_i w \otimes v_i^2} \tag{2}$$

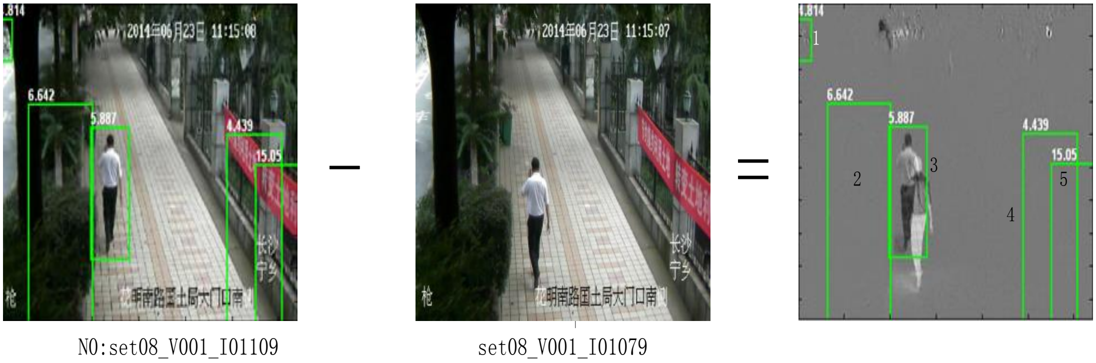
$$y_i = \frac{v_i}{\max(c, \partial)} \tag{3}$$

x_i is the input feature value of contract normalization, w is neighborhood weighting function and it is 9x9 Gauss distribution matrix. y_i is normalized output.

The sampling with mean value is to take the mean value of the feature value within specified neighborhood (the neighborhood size is determined by the sampling nucleus), then multiplied with trainable weight coefficient. The sum of product and the trainable bias is used as the input of non-linear activation function (hyperbolic tangent function tanh). The sampling with mean value takes the function of non-linear input of hyperbolic tangent function to reduce selectively the number of features produced during the convolution phase, realize the network sparseness of the implied layer of convolution nerve network and reduce the mutual dependence among static feature value of pedestrians.

2.3. Movement Feature Extraction

2.3.1. Pedestrian Movement Statistical Features: According to statistics, the walking speed of normal pedestrians is 1.5m/s. So intuitively, we can separate the pedestrians from the similar objects with pedestrian features in the static or dynamic background. In the actual outdoor environmental monitoring video, the position of the same pedestrian in the adjacent frame will change step by step, which is different from the objects similar to pedestrians in the static background which are static and different from the objects similar to pedestrians like automobiles in the dynamic background which are moving rapidly. As shown in Figure 3, five detection rectangular frames including pedestrians possibly are obtained, among which No.3 rectangular frame includes the real pedestrian (true positive) while the objects in other four rectangular frames are image background wrongly detected as pedestrians (false positive). In order to reduce the miss rate, we subtract the gray-scale map of current image frame from the gray-scale map corresponding to the preorder No. 30 frame and get the image on the right. As shown in the right image, the pixel gray value in rectangular frame 3 is relatively large while the difference of pixel gray value in the rectangular frame where the background objects are mistakenly detected as pedestrians is small. Therefore, we put forward the mean value of rectangular block pixel difference (RBPVD) and further judge whether the object in the rectangular frame is real pedestrian or object in background based on the mean value size of the pixels included in each rectangular frame in the right image.



(a) current detection frame (b) preorder No. 30 frame (c) frame difference after rapid pedestrian detection

Figure 3. Rectangular Block Pixel Difference Formation Diagram

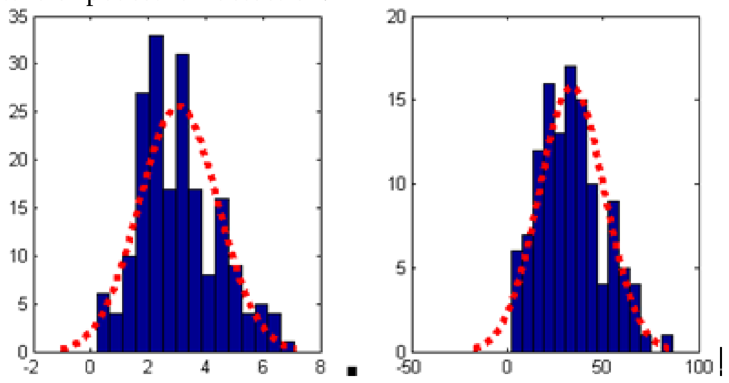
2.3.2. Mean Value of Rectangular Block Pixel Difference: As for the difference frame image listed on the right of Figure 3, the mean value μ_b and μ_p of all the pixel difference in background rectangular block inside the rectangular frame and real pedestrian rectangular block are calculated respectively, as shown in Formula (4). I_b represents the matrix set of pixel difference of background rectangular block where it is mistakenly detected as pedestrians, simply called sample set of wrong detection difference. I_p means the set of rectangular bock pixel difference including actual pedestrians in the foreground, simply called actual difference sample set. $I^i_{b(j,k)}(1 \le i \le L, 1 \le j \le m, 1 \le k \le n)$ represents the element of the pixel difference matrix I^i_p of No.*i* rectangular block in real difference sample set I_p.

μ_b and μ_p are taken as the basis to calculate separately the standard difference σ_b and σ_p of all pixel difference inside the background rectangular block inside the rectangular frame and inside the real pedestrian rectangular block, as shown in Formula (5).

$$\mu_b^i = \frac{1}{m\times n}\sum_{j=1,k=1}^{m,n} I_{b(j,k)}^i,\quad \mu_p^i = \frac{1}{s\times t}\sum_{j=1,k=1}^{s,t} I_{p(j,k)}^i \tag{4}$$

$$\sigma_b^i = \sqrt{\frac{1}{m\times n}\sum_{j=1,k=1}^{m,n}\left(I_{b(j,k)}^i - \mu_b^i\right)^2}, \sigma_p^i = \sqrt{\frac{1}{s\times t}\sum_{j=1,k=1}^{s,t}\left(I_{p(j,k)}^i - \mu_p^i\right)^2} \tag{5}$$

After calculation of 10760 background rectangular blocks which are mistakenly detected as pedestrians and 7290 real pedestrian rectangular blocks, the mean value of pixel difference of mistakenly detected rectangular blocks is 3.61 and the standard difference is 1.53; the mean value of pixel difference of real pedestrian rectangular blocks is 36.22 and the standard difference is 7.25. The histogram of distribution of mean value of pixel difference of two types of rectangular blocks is shown in Figure 4. As known from the statistical data, there is obvious difference between the mean value and standard difference of pixel difference of wrongly detected rectangular block and the corresponding statistical value of real pedestrian rectangular block, so it is reasonable to take the mean value and standard difference of pixel difference of rectangular block as the movement feature of pedestrian detection.



(a) Mistakenly detected pedestrians (b) Real pedestrians

Figure 4. Histogram of Distribution of Mean Value of Pixel Difference of Rectangular Block

2.4. Supervised Training in Combination with Movement Features

The supervised training is conduced after extraction of hierarchical features of pedestrians with convolution network. The samples used for supervised training must be labeled as pedestrians (positive samples) or non-pedestrians (negative samples). The input of supervised training is high-level pedestrian features obtained from hierarchical learning model and pedestrian movement features obtained from mean value of pixel difference of rectangular block, 3799 dimensions in total. We adopt the lineal logic regression to conduct the supervised training and the model is shown as follows:

$$h_{\theta(x)} = \frac{1}{1+\exp(-\theta^T x)}$$

x is the input pedestrian feature, θ is the training parameter, $h_{\theta(x)}$ is the input feature for certain group to observe the possibility of the samples to be pedestrians or non-pedestrians. The optimal value of θ can be reached with the maximum likelihood estimation.

3. Experiment Result and Analysis

3.1. Experiment Data and Environment

We have performed contrast test for three kinds of pedestrian detection algorithm like FPDW [25], MultiFtr+Motion [19] and CNNs+ Motion on CSUPD [24] database. The CSUPD database is created by computer science and engineering college of Central South University, consisting of video data created in the outdoor real monitoring environment. Its data source comes from the monitoring video of "Peaceful city", Changsha, China, including the monitoring data of streets, roads, parks, squares and bus stops. This data consists of pedestrian videos obtained under different angles and different lighting conditions of cameras with fixed positions and different parameters. The resolution of CSUPD video images is 640 x 480 and the total time length is 15 hours, including 1.60 million frames. The database consists of training data set and test data set and each data set consists of paired files of video segments and labeled files. The labeled files mark the pedestrian position, height, width, whether he is blocked and whether it is single pedestrian in the file of video segment. CSUPD database has labeled 3100 different pedestrians and the labeling time is 187 minutes.

3.2. Evaluation Standard

The approach provided by Caltech Pedestrian Website [26] is used in this paper to make detection precision evaluation with the measurement index being average miss rate. The average miss rate is converted from detection precision. The average miss rate considers comprehensively the detection precision corresponding to different regions where FPPI is located so as to reflect the merits and demerits of the detection algorithms from the overall aspect. The precision calculation method is shown as follows:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{6}$$

The $true\ positives$ represents the true positive number and $false\ positives$ is false positive number. We use the mutual overlapping degree between the pedestrian detection result rectangular block and real pedestrian rectangular block (already labeled before test) based on the PASCAL [27] about precision judgment method for visual target classification and predication to evaluate the detection result of each rectangular block in the detection phase:

$$a_o \doteq \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{7}$$

As for one image detected, BB_{dt} represents the rectangular frame obtained during the detection phase. When BB_{dt} could include or not include pedestrians. BB_{gt} is the rectangular block of pedestrians included in this image already labeled manually before test. When BB_{dt} includes pedestrians and the ratio between the intersection and union of the rectangular block area of BB_{dt} and certain BB_{gt} is larger than 0.5, then it means the BB_{dt} is true positive and the pedestrians are detected to be pedestrians; when BB_{dt} does not include pedestrians, while the aforementioned ratio is larger than 0.5, it means that the BB_{dt} is false positive and the background object is mistakenly detected as pedestrians.

3.3. Contrast of Detection Precision

The contrast test is performed in the experiment for pedestrian detection performance of three methods which are FPDW, MultiFtr+Motion and CNNs+ Motion based on the limitation conditions like height and width of the pedestrians in two different situations. The first situation is that the medium distance (including the near distance) pedestrians in the monitoring video are detected and the detection data is shown in Table 1. In this situation, the pedestrian height is larger than 50 pixel and the width is larger than 30 pixel, no pedestrian image is blocked, and the pedestrian profile can be recognized clearly with naked eyes. When FPPI is 1, the miss rate of CNNs+ Motion is 11% , better than the other two advanced pedestrian detection approaches. As shown in Figure 5, as for the same miss rate value, the FPPI value of FPDW algorithm is far larger than the corresponding value of CNNs+ Motion and MultiFtr+Motion. This is because the latter two kinds of algorithms consider the movement features of the pedestrians and avoids the problem that the static objects in the background are mistakenly detected as pedestrians. As known from Table 1, the average miss rate of CNNs+ Motion algorithm is reduced by 23% and 18% respectively compared with FPDW and MultiFtr+Motion.

Table 1. Contract of Test Results for Medium Distance Pedestrians on CSUPD Database with Three Kinds of Algorithms

	FPDW	MultiFtr+ Motion	CNNs+ Motion
total number of images of detection set (nImg)	5730	5730	5730
total number of pedestrians that shall be detected (np)	2755	2755	2755
total number of real pedestrians actually detected	2601	2545	2573
total number of pedestrians mistakenly detected	417385	10975	6447
average miss rate	64%	59%	41%

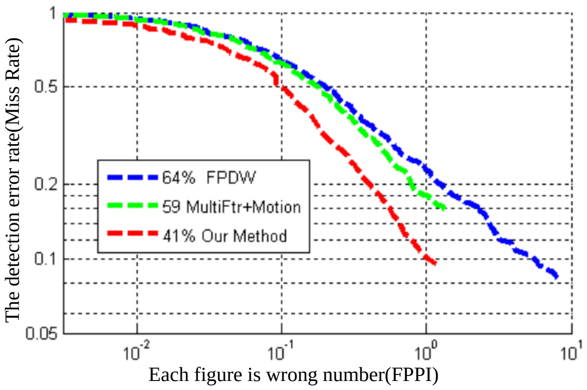


Figure 5. Detection Result of Medium-Distance Pedestrians

In the second situation, the far distance pedestrians in the monitoring video are detected. The detection data is shown in Table 2. In this situation, the pedestrian height is lower than 50 pixel, the width is lower than 30 pixel, no pedestrian image is blocked, and the pedestrian profile can be recognized basically with naked eyes. As known from Figure 6, and Table 2, compared with medium distance pedestrian detection, the miss rate of pedestrian detection with these three methods is obviously improved. This is because compared with medium distance pedestrians, the foreground profile of far distance pedestrians is blurred and the local feature is slightly week.

Table 2. Contract of Test Results for Far Distance Pedestrians on CSUPD Database with Three Kinds of Algorithms

	FPDW	MultiFtr+ Motion	CNNs+ Motion
total number of images of detection set (nImg)	5730	5730	5730
total number of pedestrians that shall be detected (np)	2478	2478	2478
total number of real pedestrians actually detected	2235	2097	2146
total number of pedestrians mistakenly detected	45738	8747	7133
average miss rate	77%	73%	61%

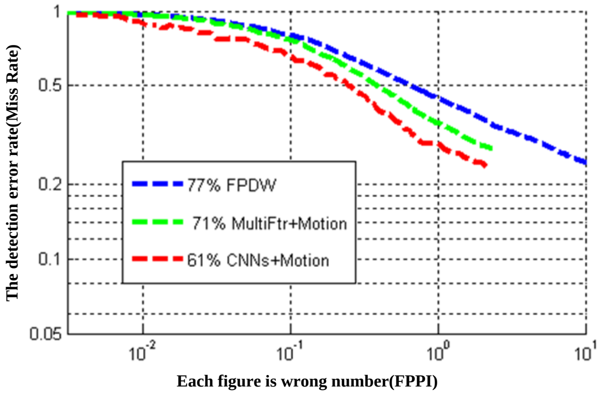


Figure 6. Detection Result of Far-Distance Pedestrians

4. Conclusions

Pedestrian detection as one widely applied biological recognition technology, has aroused continuous study by people. Complete and simple pedestrian feature extraction is the guarantee for detection precision and speed. Advanced convolution nerve network theory is adopted in this paper. As for static pedestrian images, high-level recognition features of pedestrians are extracted, mean value approach of rectangular block pixel difference is created based on the movement speed particular to the pedestrians so that the movement features of the pedestrians are extracted. The miss rate of the pedestrian detection is reduced by integration of movement features and static features of the pedestrians during the classification phase. The test shows that compared with other known research methods, this method in the paper boasts better detection results for monitoring video sequence of complicated background and low resolution.

The further study focuses on the pedestrian detection with height lower than 30 pixel. The pedestrian features on the time sequence can be extracted with 3D convolution nerve network so as to further improve the full pedestrian detection rate when the resolution is low.

Reference

[1] J. Miseikis and P. V. K. Borges, “Joint Human Detection From Static and Mobile Cameras”, IEEE Intelligent Transportation Systems, vol. 16, no. 2, **(2015)**, pp. 1018-1029.

[2] M. Pedersoli, J. Gonzalez and X. Hu, “Toward Real-Time Pedestrian Detection Based on a Deformable Template Model”, IEEE Intelligent Transportation Systems, vol. 15, no. 1, **(2014)**, pp. 355-364.

[3] W. Ge, R. T. Collins and R. B. Ruback, “Vision-Based Analysis of Small Groups in Pedestrian Crowds”, IEEE Pattern Analysis and Machine Intelligence, vol. 34, no. 5, **(2012)**, pp. 1003-1016.

[4] Y. Jiang, J. Wang and K. Guo, “A Novel Front-view Gait Recognition Approach in Real Surveillance Environments”, Journal of South China University of Technology, vol. 43, no. 1, **(2015)**, pp. 99-104.

[5] N. Dalal and B. Triggs, “Histogram of oriented gradient for human detection”, IEEE Computer Vision and Pattern Recognition ,San Diego, IEEE Press, **(2005)**, pp. 886-893.

[6] S. Rui, C. Jun and G. Jun, “Fast Pedestrian Detection Based on Saliency Detection and HOG-NMF Features”, Jounal of Electronics & Information Technology, vol. 35, no. 8, **(2013)**, pp. 1921-1926.

[7] W. Ouyang and X. Wang, “A discriminative deep model for pedestrian detection with occlusion handling”, IEEE Computer Vision and Pattern Recognition , IEEE Press, **(2012)**, pp. 3258-3265.

[8] W. Ouyang and X. Wang, “Joint Deep Learning for Pedestrian Detection”, IEEE Computer Vision, IEEE Press, **(2013)**, pp. 2056-2063.

[9] Z. Kira, R. Hadsell, G. Salgian and S. Samarasekera, “Pedestrian Detection using stereo and a cascade of convolutional network classifiers”, IEEE Intelligent Robots and Systems, IEEE Press, **(2012)**, pp. 2396-2403.

[10] P. Sermanet, K. Kavukcuoglu and S. Chintala, “Pedestrian Detection with Unsupervised Multi-stage Feature Learning”, IEEE Computer Vision and Pattern Recognition,Columbus, IEEE Press, **(2013)**, pp. 3626-3633.

[11] K. K. Htike and D. Hogg, “Efficient Non-iterative Domain Adaptation of Pedestrian Detectors to Video Scenes”, IEEE Pattern Recognition ,Stockholm,IEEE Press, **(2014)**, pp. 654-659.

[12] T. Zhao, R. Nevatia and B. Wu, “Segmentation and Tracking of Multiple Humans in Crowded Environments”, IEEE Pattern Analysis and Machine Intelligence, vol. 30, no. 7, **(2008)**, pp. 1198-1211.

[13] L. Liang, X. Yuanlu and X. Li, “Complex Background Subtraction by Pursuing Dynamic Spatio-Temporal Models”, IEEE Image Processing, vol. 23, no. 7, **(2014)**, pp. 3191-3202.

Authors

Yingjun Jiang, received his M.S. degree in electronic engineering from Guangxi Normal University in Guilin, China. He is currently a Ph.D. in College of information science and Engineering at Central South University. His research interest is mainly in the area of Computer Vision, Machine Learning. He has published several research papers in scholarly journals in the above research areas.

Jianxin Wang, Dr. Jianxin Wang received his M.Sc. (Computer Science) from Central South University of Technology, and his Ph.D. degree in computer science from Central South University, China. He is the Associate Dean and a full professor in School of Information Science and Engineering, Central South University, Changsha, Hunan, China. His research interests include algorithm analysis and optimization, computer network and bioinformatics. He has published more than 100 papers, which have been published in various journals.