Convolution Pedestrian Detection Based on Random Fusion of Color and Gradient

Gui Xiangquan[1], Jiang Jiajun[1], Li Li[1],Dongmei Chen[2], Lei Gao[2]

[1] School of computer and communication

LanZhou University Of Technology

No. 287 Lan Gong Ping Road, Qilihe District, Lanzhou, Gansu, China

[2] GanSu Province Lanzhou Three Dimension Big Data Standardization Research Institute Co.,Ltd., Lanzhou Gansu 730050

Keywords: double color; channel switching; extended operator; random fusion;

Abstract. The complex prospects in pedestrian detection, such as backpacks and other obstacles, are likely to cause interference to pedestrians. Since previous pedestrian detection can only use separate gradient information, the color information is neglected, and the gradient direction information is not accurate because of noise. In this paper, we propose a convolution network based on the combination of double color and improved Sobel extended gradient information to detect pedestrians and other prospects. The model combines convolution of RGB and HSI color channels and improved Sobel extended gradient fusion channels respectively. Then the stochastic fusion feature vector method is proposed to fuse the color and gradient information randomly, and the final result of pedestrian detection is obtained. Experimental results show that the proposed method improves the detection accuracy.

Introduction

Pedestrian detection is widely used in pedestrian behavior analysis, security system, intelligent transportation and other fields [1,2]. It is also one of the important research fields in computer image processing, pattern recognition and other fields. Pedestrian environment, attitude, angle of view and illumination are very different, especially when pedestrian foreground is complex, the irregular foreground shape and occlusion are the difficult points of pedestrian detection problems [3]. How to detect pedestrians and their belongings quickly and accurately from complex foreground is still one of the hot issues that need to be solved urgently [4]. The current mainstream method of pedestrian detection can be divided into three categories: the first category is based on HOG(Histograms of Oriented Gradients) and SVM(Support Vector Machine) combined with pedestrian detection method, and based on the improved HOG features derived from other methods. The method has good results in detecting images with uniform background and uniform scale [5]. The second kind is the pedestrian detection method [6] based on Adboost cascade classifier. This method combines many weak classifier detection, and this method is better to detect pedestrian in complex scenes. The third is the use of neural networks to detect pedestrians. The method trained the model in advance, and the training parameters were set in the weight of each level. The algorithm has good robustness to illumination and shadow, and is easy to identify different shapes of pedestrians [7], but it still needs to be improved in complex foreground pedestrian detection problems. In recent years, deep convolution network has been widely used in target detection fields, which greatly improve the accuracy and efficiency of pedestrian detection [8,9].

Based on the feature extraction of deep convolution network, a convolutional neural network based on color fusion and gradient direction stochastic fusion is proposed to detect pedestrians in complex foreground. Compared with the previous only gradient feature extraction or single channel color features, the image is converted into two color channels and a gradient of channel, then many features were extracted and integrated. Finally, we make fusion detection and classification with randomization. Experimental results show that this method can better detect pedestrians in complex foreground.

Improved convolution neural network

In this paper, an improved convolutional neural network is proposed. The original image is transformed into HSI channel image, and the original image and transformed image are extracted at the same time. The improved Sobel extended convolution kernel is used to extract contour map as the boundary detection information. Two parallel convolution layers and two layers of full layer were added to the pre-training model, each convolution layer is used to extract features from different channels. In this paper, the object is pedestrian, and the size of convolution kernel is modified so that the original convolution kernel is more suitable for detecting objects. After extracting the features of different image channels by convolution, a stochastic fusion method is proposed. The feature vectors are multiplied by random parameters to obtain the final feature after fusion, and the model identification is improved. The model modifies the activation function at the same time, which is more suitable for the detection of the people in the image. Combining the original model, the whole model is shown in FIGURE 1:

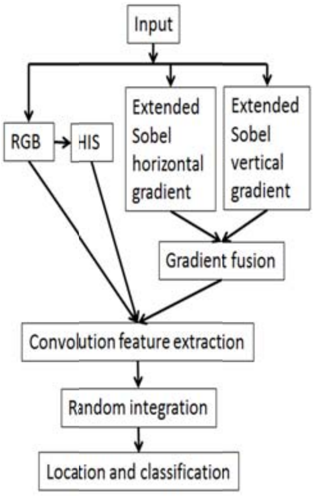

Input
RGB
HIS
Extended Sobel horizontal gradient
Extended Sobel vertical gradient
Gradient fusion
Convolution feature extraction
Random integration
Location and classification

FIGURE 1. Flow chart of convolution detection.

The principle of this model is to detect pedestrian with complex foreground using more accurate double color information and Sobel extended gradient information. The specific model flow includes channel conversion and gradient extraction, convolutional network feature extraction, and feature random fusion.

Image conversion and feature extraction

HSI color space is based on the human visual system, with hue, saturation and brightness to describe color. The HSI color space is described by a cone space model, which is quite complex, but does make clear the changes in hue, brightness, and saturation. Hue and saturation are usually referred to as chromaticity, which is used to indicate the extent of color. Because people's visual sensitivity to brightness is much stronger than the sensitivity to color shades, in order to facilitate color processing and recognition, the human visual system often uses HSI color space, it is more than RGB color space visual characteristics. The algorithm can be processed separately in HSI color space.

Traditional CNN only uses gray information of the image but loses color information when training or learning. The RGB and HSI channel information are input into the convolution layer to detect the color features, and the color information is effectively utilized. The convolution network is similar to the human optic nerve mechanism. This method uses HSI and RGB channel for detection, which is more likely to enhance the detection effect. The conversion formula for RGB channels to HSI channels is shown as follows:

$$\theta=\cos^{-1}\{\frac{[(R-G)+(R-B)]/2}{\sqrt{(R-G)^2+(R-B)(G-B)}}\} \tag{1}$$

$$H=\begin{cases}\theta, B\leq G\\ 360-\theta, B>G\end{cases} \tag{2}$$

$$S=1-\frac{3\times\min(R,G,B)}{R+G+B} \tag{3}$$

$$I=\frac{(R+G+B)}{3} \tag{4}$$

Finally, the HSI channel conversion effect is shown, as shown in FIGURE 2:

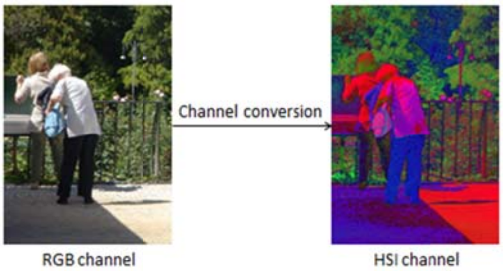
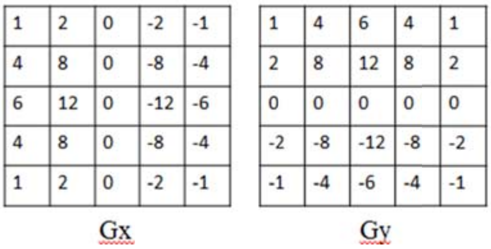


FIGURE 2. Channel conversion

In order to make better use of the gradient information in the image to detect pedestrians and reduce the impact of noise on detection, this paper proposes an improved Sobel extended kernel to extract feature information as edge information. The Sobel extended operator is shown in FIGURE 3:



1	2	0	-2	-1
4	8	0	-8	-4
6	12	0	-12	-6
4	8	0	-8	-4
1	2	0	-2	-1

Gx

1	4	6	4	1
2	8	12	8	2
0	0	0	0	0
-2	-8	-12	-8	-2
-1	-4	-6	-4	-1

Gy

FIGURE 3. Improved Sobel extended operator

Among them, Gx is a convolution extended operator in the horizontal direction, and Gy is a convolution extended operator in the vertical direction. The fusion effect diagram is shown in FIGURE 4:

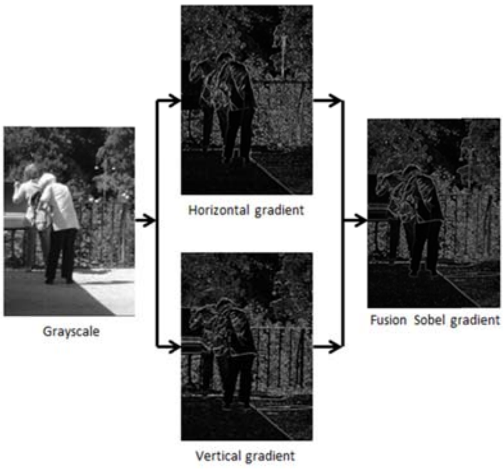


FIGURE 4. Sobel extended feature fusion

In FIGURE 4, it can be seen that the pedestrians in the gray scale are largely upright, so the gradient in the horizontal direction and the vertical direction is the most obvious when the pedestrians are detected, and the gradient profile is obtained after the fusion. Due to the use of extended Sobel operator, the extraction of features will filter out a lot of useless noise points. Compared with 3*3 size Sobel operator, it improves the pedestrian detection accuracy.

Convolutional neural network is a kind of effective model in deep learning. It extracts feature directly from the image, and memory in each weight of convolution. Convolutional neural networks use multilayer networks of different types, including input layer, convolutional layer, pooling layer, full connectivity layer, and output layer. The error of the calibration results and the predicted results is minimized by the error back propagation algorithm, so as to optimize the model to deal with the accuracy of a certain kind of problem.

At the convolution layer, each neuron of the convolution nucleus is connected to the local receptive field of the previous layer, and the convolution operation is performed to extract the local feature of the position.

When extracting features, neurons of the same convolution kernel share a set of weights, and different convolution kernel weights are different to extract different features. In the training process, we continuously adjust the parameters of each weight so that the feature extraction is carried out in the direction that is conducive to solving certain problems. The improved convolution kernel size is shown in FIGURE 5. In general, the convolution layer is calculated as equation (5):

$$x_j^l = f\left(\sum\nolimits_{i\in M_j} x_i^{l-1} \times k_{ij}^l + b_j^l\right) \qquad (5)$$

Among them, l represents layer， k represents the convolution kernel， M_j on behalf of the input layer of the field, b on behalf of the bias.

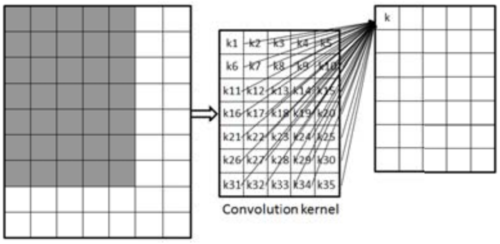


FIGURE 5. Improved convolution kernel size

In the lower sampling layer, the input feature graph has the same number as the pool layer, and the size becomes the original $\frac{1}{n}$ times (Suppose that the length of the pool layer is n)。There are many methods to pool the layers, and there are two kinds of methods: maximum down sampling and average down sampling. The lower sampling layer is shown in equation (6):

$$x_j^l = f(\omega_j^l \mathrm{down}(x_j^{l-1}) + b_j^l) \qquad (6)$$

$\mathrm{down}(.)$ is the pool layer function and ω is the weight coefficient.

Image feature fusion

In the current network weights, the output of three multilayer convolutional networks is calculated. The three groups of one-dimensional feature vectors are obtained by using the convolution kernel to extract images from different channels.

The $R_i(i=1,2,3)$ for each one-dimensional feature vectors, then the final fusion feature vector is R, namely $R=[\alpha\times R_1, \beta\times R_2, \gamma\times R_3]$. Among them, R_1 is RGB channel feature, R_2 is HSI channel feature, and R_3 is a fusion Sobel gradient feature. The characteristic parameters satisfy the condition: $\alpha+\beta+\gamma=1$, where $\gamma=\mathrm{random}(0,1)$, $\alpha=\beta=\frac{(1-\mathrm{random}(0,1))}{2}$.

The model is based on YOLO9000 to improve [10]. If there is no target, then the confidence of the pedestrian detection is zero. In addition, the confidence region of the detected pedestrian and the intersection region of the real position are multiplied to obtain a confidence value for detecting the target as a pedestrian in the region.

The value obtained by each lattice represents the probability that the lattice prediction is pedestrian. This value fraction represents the probability that pedestrians will appear in the bounding box. The evaluation on the test set, using the modified convolution kernel, namely the grid number is 35, the number of the boundary of each grid box is 2, we pre-set 20 final target classification, the final prediction result is a eigenvector with 3150 dimensions.

Activation function selection

This model presents an improved tanh and RELU binding function as a new activation function for complex foreground pedestrian detection. The specific formulas and functions are shown in the following equation 7 and FIGURE 6:

$$\mathrm{f}(x)=\begin{cases} x, x>0 \\ \dfrac{2(1-e^{-2x})}{1+e^{-2x}}, x\leq 0 \end{cases} \tag{7}$$

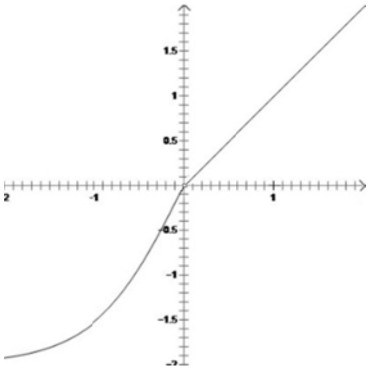


FIGURE 6. Improved activation function

There are two main reasons for using this improved activation function: First, the convergence of the tanh network is faster than that of sigmoid in the literature [11]. Since the output mean of tanh is closer to 0 than sigmoid, SGD will be closer to the natural gradient, thus reducing the number of iterations required and accelerating the training of convolution networks [12]. In a certain range, improving the left part of the activation function parameters by 2, so that the function is not easy to saturation, the model of the input can be more robust to changes or noise. Second, the right side uses RELU's straight line, makes the calculation simple, and iterative quickly. The function is unlikely to be saturated [13]. The convergence properties of the improved activation function are better than those of RELU and PRELU.

Sample set selection and training

The experimental sample set is mainly collected by hand, in Lanzhou streets and suburb wilderness. A total of 2200 pedestrian pictures were selected for pedestrian classifiers for training and verifying deep convolution networks, of which 1100 positive samples were selected and 1100 negative samples were selected. In the pre-training phase, the experiment uses the image flip, translation, region extraction of multiple images, to achieve the data expansion, and ultimately get 5000 images. Training and verification using the method of cross validation, all images will be disrupted by order. The ratio of training set and verification set is 8:2. The experiment was conducted five times, taking the average of the final training and testing as a result.

This experiment was developed on the Caffe platform, and in the case of CUDA and cuDNN accelerated library, GPU acceleration calculation was realized, which greatly improved the efficiency of detection.

Experimental results and discussion

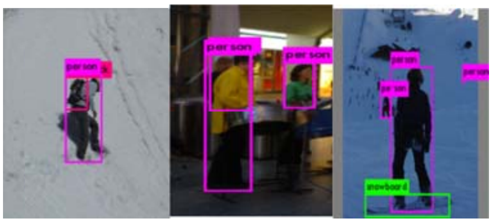
In order to evaluate the performance of the objective evaluation method, the paper selects two indexes, false detection-rate and missed detection-rate, and puts forward the comprehensive error detection index to analyze and evaluate the comprehensive results of detection. Under this index, the minimum value of relative synthesis is obtained, which makes the accuracy more reliable. Performance indicators are as follows:

$$false\ detection\ -\ rate = \frac{fp}{tp+fp} \times 100\% \tag{8}$$

$$missed\ detection\ -\ rate = \frac{fn}{tp+fn} \times 100\% \tag{9}$$

$$\begin{aligned} &synthetic\ error\ \text{-}\ rate = \\ &\alpha \times fd\text{-}rate + (1-\alpha) \times md\text{-}rate \end{aligned} \tag{10}$$

In the formula, tp is the correct number of samples for pedestrians; fp is the number of samples that have been wrongly checked for pedestrians; fn is the number of pedestrians that are not detected. When α is 0.6, there is the lowest synthetic error detection rate. At this time, the ratio of false detection and missed detection is appropriate. Later, it is necessary to judge that the model has the lowest error rate when α is 0.6.



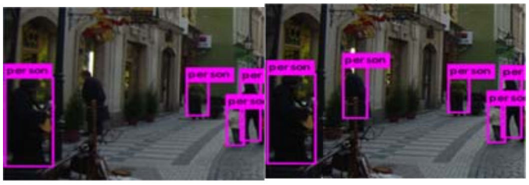
(a) Backpack (b) Drum Kit (c)Snowboard

FIGURE 7. Different foreground

After the experiment results are obtained, three representative pictures are selected, as shown in FIGURE 7. In FIGURE 7(a), the confidence level of pedestrian detection is 83%, and the knapsack on pedestrian is also detected. In FIGURE 7(b), the occluded areas of three people are different from each other, and the confidence of pedestrian detection in this image is 63%, 73% and 78% respectively. In FIGURE 7(c), two pedestrians were detected, with confidence of 89% and 70%, respectively. In addition, the skis are detected below, which shows that the model is effective for detecting pedestrians at different scales in the image.

In FIGURE 8(a), the setting threshold is 0.6. Because of the dark background of the image, the pedestrian and the background are fused, resulting in a pedestrian not being detected on the wall. When the threshold is set to 0.45, as shown in FIGURE 8(b), the algorithm detects all pedestrians in the image, with confidence levels of 70%, 65%, 71%, 80% and 57%, respectively.

By choosing different thresholds, we obtain the relation between the false detection-rate of f and the missed detection-rate of m with different thresholds, as shown in FIGURE 9 and FIGURE 10. Finally, the synthetic error-rate of s is obtained. As shown in FIGURE 11, it can be seen that when the threshold is 0.54, the improved model has the lowest synthetic error-rate.



(a) Missed pedestrian detection (b) Complete pedestrian detection

FIGURE 8. Different thresholds

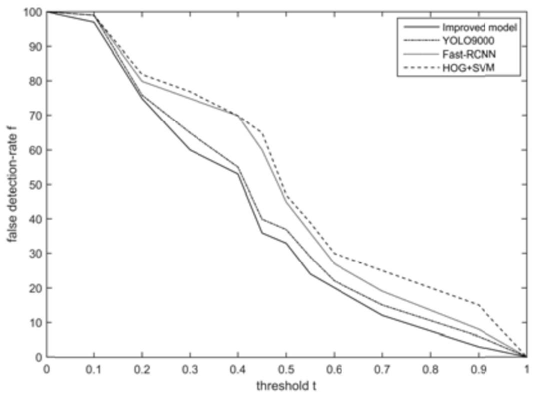


FIGURE 9. False detection-rate of f - threshold t

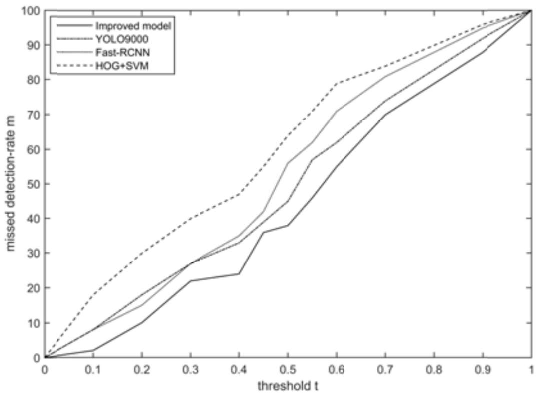


FIGURE 10. Missed detection-rate of m - threshold t

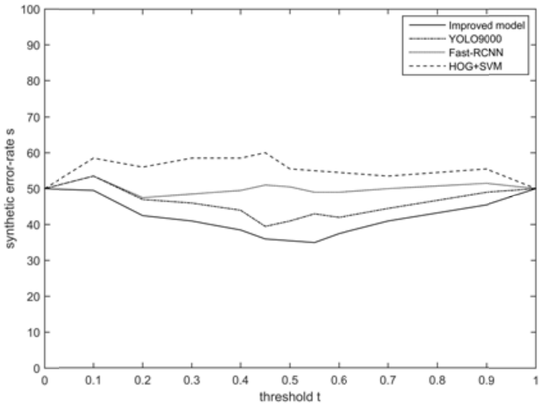


FIGURE 11. Synthetic error-rate of s - threshold t

By comparing different methods to detect pedestrians in complex environment, the following data in the table are obtained:

TABLE 1. Comparison of methods in complex foreground

Pedestrian detection method	Feature dimension	Comprehensive detection rate	Time (ms)
HOG+SVM[14]	3780	44%	56.28
Fast-RCNN[15]	4096	53%	19.24
YOLO9000	1470	62%	15.74
Improved model	3150	68%	24.38

By comparing the experimental results of TABLE 1, it shows that the total detection rate of this model is 15% higher than that of Fast-RCNN, 6% higher than the YOLO9000, and is obviously better than the artificial feature method. Although the execution time is slightly slower than the YOLO9000 and Fast-RCNN, the performance of GPU is better than the traditional feature detection algorithm because of its speedup.

Conclusion

In this paper, double color and extended Sobel gradient features are randomized for fusion in convolution neural network, which is used to carry out pedestrian detection in complex foreground environment. The RGB and HSI dual color channels and the improved Sobel extended operator are used to extract the features. Then we randomize the fusion, so that the model fully utilizes the color and the extended gradient information to detect pedestrians.

The validity of the model is verified on the sample library by using improved detection model, which improved activation function.

For the reasonable size of pedestrian detection, the effect is good. But for the obvious smaller pedestrians, detection effect still needs to be improved. In the follow-up work, it is necessary to combine small target pedestrians for more accurate detection of pedestrians with different foreground.

Acknowledgment

This work is partially supported by Guixiang Quan, Jiangjia Jun, Li Li,Cao Jie. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Zheng G, Chen Y. A review on vision-based pedestrian detection[C]// Global High Tech Congress on Electronics. IEEE, 2012:49-54.

[2] Schindler K, Ess A, Leibe B, et al. Automatic detection and tracking of pedestrians from a moving stereo rig[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2010, 65(6):523-537.

[3] Xu Y W, Cao X B, Qiao H. Survey on the latest development of pedestrian detection system and its key technologies expectation[J]. Acta Electronica Sinica, 2008, 36(5):962-968.

[4] Song-Zhi S U, Shao-Zi L I, Chen S Y, et al. A Survey on Pedestrian Detection[J]. Acta Electronica Sinica, 2012, 40(4):814-820.

[5] Xi H Y, Xiao Z T, Zhang F. Study on Pedestrian Detection Method Based on HOG Features and SVM[J]. Advanced Materials Research, 2011, 268-270:1786-1791.

[6] Cheng W C, Jhan D M. A cascade classifier using Adaboost algorithm and support vector machine for pedestrian detection[C]// IEEE International Conference on Systems, Man, and

Cybernetics. IEEE, 2011:1430-1435.

[7] Bertozzi M, Cerri P, Felisa M, et al. Pedestrian validation in infrared images by means of active contours and neural networks[J]. EURASIP Journal on Advances in Signal Processing, 2010, 2010(1):1-14.

[8] Li H, Wu Z, Zhang J. Pedestrian detection based on deep learning model[C]// International Congress on Image and Signal Processing, Biomedical Engineering and Informatics. IEEE, 2017.

[9] Chen X, Wei P, Ke W, et al. Pedestrian Detection with Deep Convolutional Neural Network[J]. 2014, 9008:354-365.

[10] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[J]. 2016.

[11] Gomar S, Mirhassani M, Ahmadi M. Precise digital implementations of hyperbolic tanh and sigmoid function[C]// Signals, Systems and Computers, 2016, Asilomar Conference on. IEEE, 2017.

[12] Amari S I. Natural Gradient Works Efficiently in Learning[M]. MIT Press, 1998.

[13] Nair V, Hinton G E. Rectified Linear Units Improve Restricted Boltzmann Machines[J]. Proc Icml, 2010:807-814.

[14] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2005:886-893.

[15] Girshick R. Fast R-CNN[J]. Computer Science, 2015.