Pedestrian detection method and Research of Attention Mechanism in Traffic Scene

Fanshen Xu

University of Pittsburgh, Pittsburgh 15260, USA

Abstract. Aiming at the occlusion problem when pedestrian detection algorithms are applied in traffic scenes, an occlusion-aware pedestrian 2D saliency texture operator TS-LBP detection algorithm combined with dual attention mechanism is proposed. The algorithm first uses the differential feature-aware fusion method to mine the complementary information between multi-modal features to optimize the channel features; then uses the detection framework with an efficient anchor-free mechanism to greatly reduce the computational complexity of the model and improve the detection speed. The experimental results show that the method in this paper is fast and accurate, and has a good detection effect.

Keywords: Pedestrian detection; Region of interest; Color feature; Differential perception.

1. Introduction

Pedestrian detection has a wide range of application scenarios, such as security, autonomous driving, mobile robots, etc. Therefore, pedestrian detection has always been a very popular research direction in the field of computer vision. Before FasterR-CNN was proposed, all CNN-based detection algorithms were first the traditional region proposal algorithm is used to generate the proposed box, and then CNN is used to classify and regress the proposed box. Due to the huge amount of calculation of the traditional region proposal algorithm, it takes tens of milliseconds or even hundreds of milliseconds to process an image, which has always been the bottleneck of real-time detection [1]. To reduce proposal box generation time. The FasterR-CNN method proposes an anchor mechanism, uses RPN to slide the window on the feature map output by the backbone, and uses the anchor as the initial frame to directly generate the proposed frame, which can not only achieve a higher recall rate than the traditional region suggestion algorithm, but also the computational cost is almost zero, thus solving the problem of proposal box generation. Attention mechanism originated from the research of human vision, and has been widely used in various tasks of computer vision (such as image classification, detection and segmentation, etc.). There are two types of common attention mechanisms: one is the spatial attention mechanism, which adaptively adjusts the weight of each element in the feature map through network learning; the other is the channel attention mechanism, which uses the network to adjust the difference in the feature map. The weight of the channel. Using the attention mechanism can strengthen the network's attention to the features of the pedestrian's visible area, thereby improving the algorithm's occlusion processing ability. Some scholars use the component heatmap generated by the pre-trained pedestrian pose estimation model as the supervision information to guide the learning of the channel attention mechanism, which effectively improves the detection effect of occluded pedestrians, but it only uses a single channel attention mechanism and requires additional network to generate supervision information, and the detection framework is complex [2]. This paper proposes a two-dimensional saliency texture operator TS-LBP, which is combined with color features to describe the target. First, obtain the area of interest that pedestrians may exist in the image, and focus on the pedestrian detection in the area of interest. Secondly, the fusion features of the target in the region of interest are extracted, and the fused features are used to describe the pedestrian features. Finally, a pedestrian classifier is used to identify pedestrians.

2. Detection method of this paper

The whole framework consists of two parts: a feature extraction module and an anchor-free detection head module. The algorithm uses VGG16 as a feature extraction network for visible and infrared images, and embeds a DMAF module and a DFA module between each layer of VGG16. The former is used to solve the problem of blind fusion of multimodal features, while the latter is an attention mechanism whose purpose is to improve the expressiveness of features and filter redundant information [3]. The anchor-free detection head module adopts the anchor-free detection framework, which avoids the hyperparameter setting and post-processing flow of anchor boxes, greatly reduces the amount of computation, and enables end-to-end training. The algorithm mainly includes a training part and a detection part. The algorithm framework is shown in Figure 1 (the picture is quoted from An ROIs based pedestrian detection system for single images).

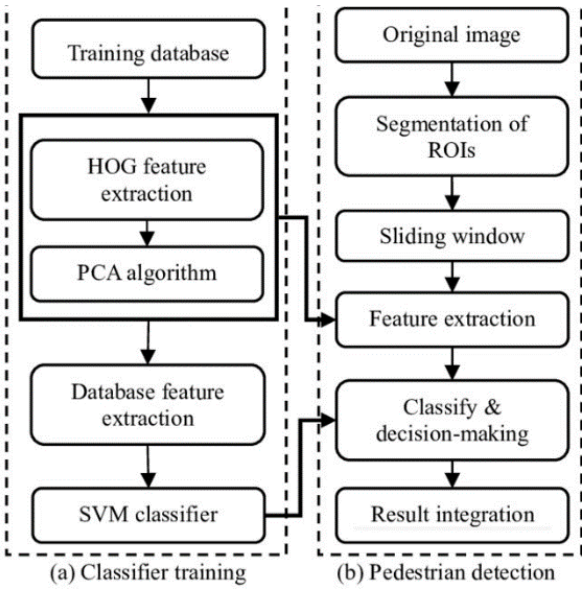

Figure 1. Pedestrian detection system framework

2.1 Obtaining the region of interest

There are a large number of non-pedestrian areas in the vehicle video image, and the traditional global image scanning will take a lot of time. Therefore, it is possible to first determine the area of interest that pedestrians may exist, and then focus on detecting the target in the area of interest to avoid global scanning. image and consume too much time [4]. Pedestrians have strong vertical symmetry whether they are standing or walking, especially their legs, which have obvious symmetry. This feature can be used to roughly extract areas of interest where pedestrians may exist. Equation (1) is used to calculate the edge symmetry of each column of the image.

$$\text{if } \begin{cases} w_{\min} < | x_i - x_j | < w_{\max} \\ y_i = y_j \end{cases} \quad Then\, S_v(\text{k})++, k = \frac{i+j}{2} \tag{1}$$

Where Wmin and Wmax are the search range of vertical edge points, and SV(k) is the symmetry measure corresponding to the kth column. At the same time, some scholars have calculated that the average pedestrian aspect ratio is 0.4. The ratio can be appropriately relaxed to extract more areas to be inspected to avoid missed inspections, and at the same time, a large number of non-pedestrian areas can be excluded. From this, the approximate location of the pedestrian in the image can be determined.

2.2 Color Feature Extraction

Color features are widely used in image processing due to their advantages of simple calculation and easy extraction. HSV (has a strong ability to adapt to changes in light. Therefore, first convert the input image from RGB color space to HSV color space according to formula (2). Non-uniform quantization of color is carried out according to the perception characteristics of human vision to color. The HSV color space is thus quantized into a 72-dimensional histogram feature vector.

$$h=\begin{cases}0^{\circ} & if\ \max=\min\\ 60^{\circ}\times\dfrac{g-b}{\max-\min}+0^{\circ}, & \text{if } \max=\mathrm{r}\ \text{and}\ \mathrm{g}\geq\mathrm{b}\\ 60^{\circ}\times\dfrac{g-b}{\max-\min}+360^{\circ}, & \text{if } \max=\mathrm{r}\ \text{and}\ \mathrm{g}<b\\ 60^{\circ}\times\dfrac{g-b}{\max-\min}+120^{\circ}, & \text{if } \max=g\\ 60^{\circ}\times\dfrac{g-b}{\max-\min}+240^{\circ}, & \text{if } \max=\mathrm{b}\end{cases} \tag{2}$$

$$s=\begin{cases}0, & if\ \max=0\\ \dfrac{\max-\min}{\max}=1-\dfrac{\min}{\max}, & otherwise\end{cases}$$

$$v=\max$$

2.3 Texture Feature Extraction

LBP is an operator proposed by Ojala et al. to extract local texture features of images [5]. It is widely used in the field of image processing because of its remarkable advantages such as grayscale invariance, rotation invariance, insensitivity to illumination changes, and fast and simple calculation. Taking the 3*3 template as an example, the LBP value of the center pixel of the window can be obtained according to formula (3). The schematic diagram of the LBP operator is shown in Figure 2.

$$LBP_{P,R}=\sum_{p=0}^{P-1}s(\mathrm{I}_p-I_c)2^p$$
$$s(\mathrm{x})=\begin{cases}1, x\geq 0\\ 0, x<0\end{cases} \tag{3}$$

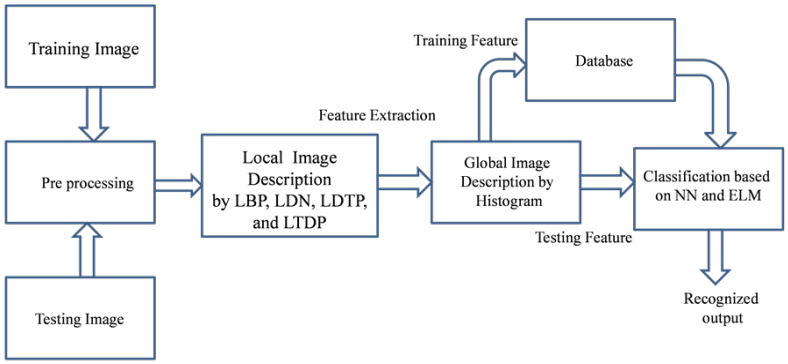

Figure 2. Schematic diagram of LBP operator

There are too many dimensions and more redundancy in the statistical LBP histogram. CS-LBP has centrosymmetric properties. Taking gc as the center, comparing the gray values of two symmetrical pixels clockwise, a histogram with only 16 dimensions can be obtained, which can reduce the dimension of LBP. The calculation formula of the CS-LBP operator is shown in Equation (4), and the schematic diagram of the operator is shown in Figure 3 (the picture is quoted from A Novel Improved Local Binary Pattern and Its Application to the Fault Diagnosis of Diesel Engine).

$$CS-LBP_{P,R}=\sum_{i=0}^{(P/2)-1}s(g_i-g_{i+(P/2)})2^i$$
$$s(x)=\begin{cases}1 & x\geq 0\\ 0 & otherwise\end{cases} \tag{4}$$

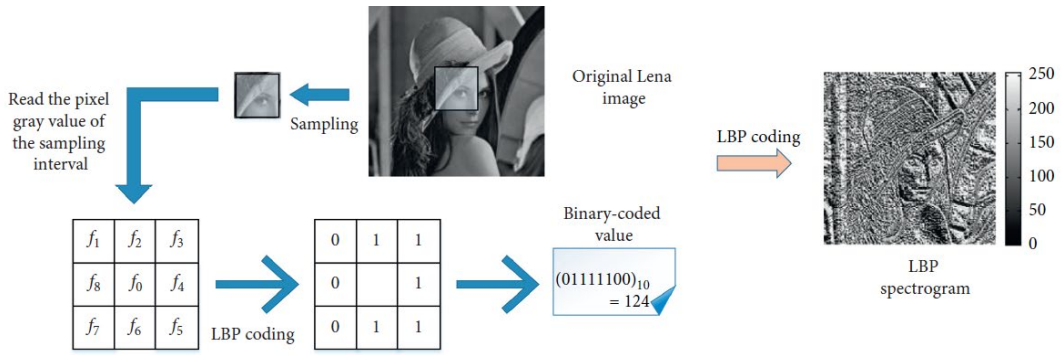

Figure 3. Schematic diagram of CS-LBP operator

CS-LBP can reduce the dimension of the graph, but the anti-jamming performance against noise is not enough. In order to improve the robustness of the algorithm and make it have stronger anti-jamming ability to noise, this paper proposes a CS-LBP with a noise threshold T. The texture features are shown in formula (5).

$$CS-LBP_{P,R}^{T}=\sum_{i=0}^{(P/2)-1}s(g_i-g_{i+(P/2)},T)2^i$$
$$s(x,T)=\begin{cases}1 & x\geq T\\ 0 & otherwise\end{cases} \tag{5}$$

The texture features of CS-LBP with noise threshold have grayscale invariance and rotation invariance of traditional LBP, and at the same time, it is symmetrical and has better anti-noise performance. However, this feature does not reflect the saliency feature of the graph well [6]. In this paper, a saliency operator LST with a threshold is proposed to have better adaptability to the pixel gray value, and the operator is expressed as formula (6).

$$LST_{P,R}^{t}=\sum_{i=0}^{P-1}f(g_i-g_c,t)$$
$$f(x,t)=\begin{cases}1 & |x|\geq t\\ 0 & otherwise\end{cases} \tag{6}$$

Taking the 3*3 template as an example, the P value is 8, the R value is 1, gc is the gray value of the center pixel, and gi is the gray value of the 8 pixels in the field. Binarize the template, compare gc with gi, when gc is greater than gi greater than a certain threshold t, its value is 1, otherwise its value is 0. Combining the CS-LBP feature with noise threshold and the thresholded saliency operator LST, the two-dimensional saliency texture operator TS-LBP in this paper is constructed as shown in Eq. (7). Using this operator to represent the texture feature of the image can not only reflect the texture structure of the image, but also display its local saliency features, and at the same time, it has better anti-noise performance.

$$TS-LBP=(CS-LBP_{P,R}^{T},LSN_{P,R}^{t}) \tag{7}$$

2.4 Feature fusion

First, extract the color histogram of the target area of interest according to formula (2), then extract the texture histogram of the target area of interest according to formula (7), and fuse the two features together to form a fusion feature histogram as shown in Figure 4. (Image referenced in Race classification from face images using local descriptors). The fused features can reflect the texture features, saliency features and color features of the target area, and have good anti-noise performance.

Compared with a single-color feature or texture feature, this feature has stronger anti-interference and higher robustness.

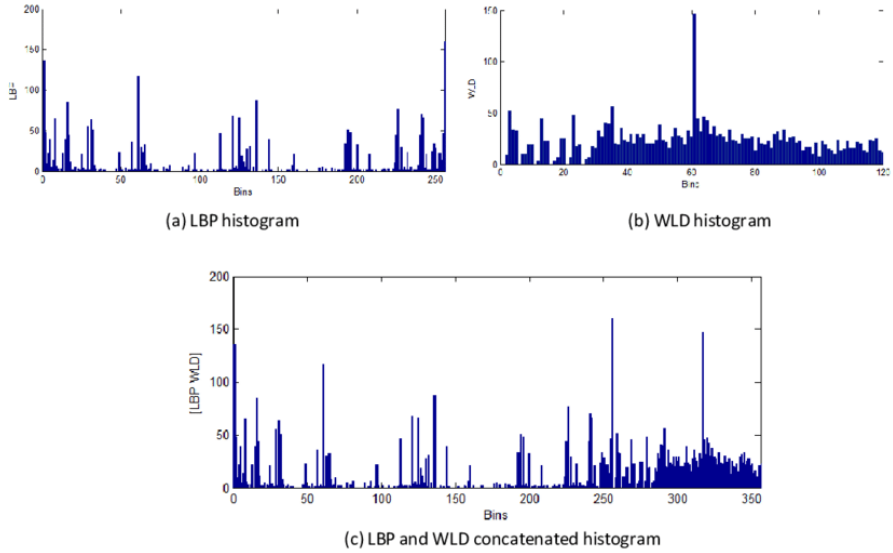

Figure 4. Feature fusion histogram

2.5 Introduction to Adaboost Algorithm

Adaboost algorithm is an iterative algorithm. The core idea is to use the same training set to obtain weak classifiers with general classification performance through training, and cascade these weak classifiers to form a strong classifier with better classification performance [7]. Pedestrians are classified by this strong classifier. The flow of the algorithm is as follows:

A1): Given a training set: $F=\{(x_1,y_1),...,(x_N,y_N)\}$, where $y_i\in\{0,1\}$.

2): Initial distribution of training samples: $D_1(i)=\frac{1}{N}$

3): For $t=1,...,T$, calculate the error of weak classifier $h_t:X\to\{0,1\}$ on distribution D_t: $\varepsilon_t=P_{D_t}\left(h_t(x_i)\neq y_i\right)$ Calculate the weight of the weak classifier: $\alpha_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

4): Update the distribution of training samples: $D_{t+1}(i)=\frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$, where Z_t is a normalization constant.

5): The final strong classifier is: $H(x)=sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$

3. Algorithm detection

The dataset used in the experiment is the PASCALVOC dataset, which includes 20 categories, namely aero, bike, bird, boat, bottle, bus, car, cat, chair, cow, table, dog, horse, bike, person, plant, sheep, sofa, train, tv, a total of 27088 pictures [8]. This paper uses the PASCALVOC2019 and PASCALVOC2012 datasets for training, and the PASCALVOC2019 dataset for testing. It is based on the TensorFlow framework and uses a graphics processor (GPU) for accelerated operations. The experimental environment is Intel□CoreTMi7-8750HCPU@2.20GHz processor, 8GB memory, Nvidia GeForce GTX1050Ti graphics card. In object detection, the map metric is usually used to evaluate the accuracy. Based on the original SSD model, the resolution of the input image is set to 300×300, the Batch size is 8, and the initial learning rate is 0.0001. First, iterate 80,000 times and then reduce the learning rate to 0.00001. Iterate 100,000 times to obtain the final network model. The

experimental results comparing the original SSD algorithm (the input image resolution is 300×300), the DSSD algorithm and the algorithm in this paper are shown in Table 1.

Table 1. Comparison of detection accuracy of three target detection algorithms under 20 categories

Algorithm		SSD	DSSD	This article
The internet		VGG-16	ResNet	VGG-16
mAP/%		77. 5	78. 6	79. 7
Detection accuracy/%	aero	79. 5	81. 9	83. 5
	bike	83. 9	84. 9	86. 0
	bird	76. 0	80. 5	78. 1
	boat	69. 6	68. 4	74. 8
	bottle	50. 5	53. 9	53. 4
	bus	87. 0	85. 6	87. 9
	car	85. 7	86. 2	87. 3
	cat	82.4	88. 9	88. 6

It can be seen from the detection results that the TS-LBP detection algorithm of the occlusion-aware pedestrian 2D saliency texture operator combined with the dual attention mechanism proposed in this paper improves mPA by 2.2% compared with the original SSD algorithm [9]. The comparison results are shown in Figure 5.

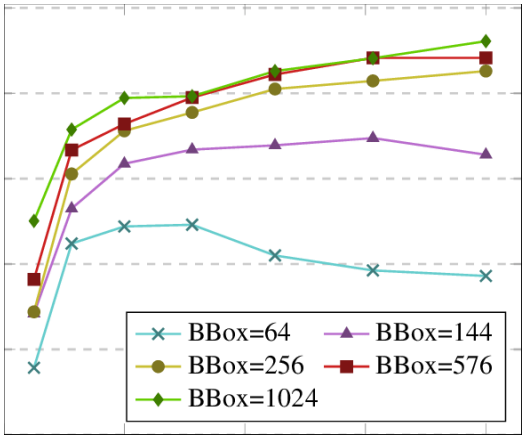

Figure 5. MAP comparison of three target detection algorithms on the PASCALVOC2019 test set

It can be seen from Figure 6 that the detection algorithm in this paper has improved the accuracy of the target frame location for pedestrian detection compared with the original SSD algorithm. At the same time, for relatively distant targets, the detection algorithm in this paper can still identify them, but the original SSD algorithm cannot detect them. Therefore, compared with the original SSD algorithm, the detection effect of the detection algorithm in this paper is better.

(a) The detection system under the SSD algorithm

(b) The target frame localization method for pedestrian detection proposed in this paper

Figure 6. Comparison of test results

4. Conclusion

This paper proposes a pedestrian detection algorithm based on the fusion of color features and salient texture features in the region of interest, and conducts pedestrian detection experiments through Adaboost classifier. This intelligent pedestrian detection method in traffic scenes can effectively improve the recognition of occlusion-aware pedestrian systems through the 2-dimensional saliency texture operator TS-LBP algorithm for occlusion-aware pedestrians with dual attention mechanism, avoiding the need for the entire image to be detected. The image is scanned globally, which is fast and robust. Improved target frame localization accuracy for pedestrian detection. This method greatly reduces the computational complexity of the model and improves the detection speed.

References

[1] Feng Yuping, Guan Yuyu, Liu Ning, et al. Real-time pedestrian detection method fused with attention mechanism. Electronic Measurement Technology, vol. 44, pp. 87-92, Seventeen 2021.

[2] Zhao Bin, Wang Chunping, Fu Qiang, et al. Multi-scale infrared pedestrian detection based on deep attention mechanism. Acta Optics, vol. 40, pp. 12-17, May 2020.

[3] Xue Yongjie, Ju Zhiyong. Indoor scene recognition method based on attention mechanism and deep neural network. Small Microcomputer System, vol. 42, pp. 77-81, May 2021.

[4] Liu Ziyan, Wan Peipei. A feature extraction method for pedestrian re-identification based on attention mechanism. Computer Applications, vol. 40, pp. 672-676, March 2020.

[5] Zheng Xin, Lin Lan, Ye Mao, et al. Pedestrian re-identification combining attention mechanism and multi-attribute classification. Chinese Journal of Image Graphics, vol. 25, pp. 10-15, May 2020.

[6] Wang Lihui, Yang Xianzhao, Liu Huikang, et al. Pedestrian detection and tracking algorithm based on GhostNet and attention mechanism. Data Collection and Processing, vol. 37, pp. 14-22, January 2022.

[7] Li Jingyu, Yang Jing, Kong Bin, et al. Multi-scale vehicle pedestrian detection algorithm based on attention mechanism. Optical Precision Engineering, vol. 29, pp. 11-14, June 2021.

[8] Zhou Dake, Song Rong, Yang Xin. Occlusion-aware pedestrian detection combined with dual attention mechanism. Journal of Harbin Institute of Technology, vol. 53, pp. 81-88, September 2021.

[9] He Xumei, Shu Xiaohua, Gu Zhiru, et al. Pedestrian detection method in traffic scene based on deep learning. Electronic Products World, vol. 28, pp. 44-48, March 2021.