ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0016

Target Detection Algorithm in Complex Scenes

Xing Gu, Tingting Gui, Yanli Shi

Changchun University of Science and Technology National Demonstration Center for Experimental Electrical, Changchun, Jilin, 130022, China.

Abstract

With the development of computer vision, as a basic problem in this field, target detection is to determine the area of interest to us through a certain algorithm. After a long research, the field has developed rapidly and achieved good results in most scenes. However, in some complex scenes, the shape of the target will change to a certain extent due to various reasons such as illumination, occlusion, and angle greatly damage its detection effect. In this paper, the problem of target detection in complex scenes is studied from two directions. In the aspect of feature expression, a local image block descriptor based on CNN features is provided; in the target classification model, a CNN image block and GBRF phase is proposed Combined target classification model. The method proposed in this paper has good results in complex scenes.

Keywords

Target Detection; CNN; GBRF.

1. Introduction

In real scenes, there are often a series of problems such as non-rigid deformation, multiple viewing angles, and partial occlusion of different targets due to overlapping spatial positions [1]. These problems often affect the effect of target detection, and make accurate target detection become a very challenging problem in complex scenes in the field of computer vision. However, since target detection is widely used in various fields [2] such as security monitoring systems, intelligent transportation systems, and scene analysis, it is necessary to solve the existing problems at this stage. The current solution of this situation is mainly embodied in two specific aspects: firstly, the image feature expression method has strong robustness and high discrimination; secondly, using the existing feature expression designs a more robust classification model.

In terms of feature expression, this paper combines the deep image features of the convolutional neural network, and proposes a local image block descriptor based on the convolutional neural network. It not only uses the relatively stable characteristics of the image of the local feature, but also uses the powerful features of the deep layer, which ability to express features. In the target classification model, a target classification model combining CNN and GBRF is proposed.

2. Improved target detection algorithm

2.1 Local image block descriptor based on CNN features

A CNN can be roughly divided into a feature extraction part and a trainable classification part[3]. This article is based on the AlexNet network to study the features of the pooling layer after the fifth convolutional layer of AlexNet, the sixth and seventh fully connected layers are extracted. The research can find that the expression ability of the features calculated by the CNN model mainly comes from the convolutional layer and the pooling layer connected to it, rather than relying on the fully connected layer with huge parameters[4].

Based on the above analysis, this paper proposes a truncated AlexNet to obtain CNN features, in which three improvements are made to the CNN feature extraction network: (1) The last three fully connected layers of AlexNet are removed, and the fifth convolutional layer is also removed. Since the latter pooling layer only down-samples the feature pooling extracted by the fifth convolutional layer, it does not significantly improve the feature extraction effect, so the fifth convolutional layer output feature is finally used as the final CNN output feature. (2) Each time before convolution and pooling are performed with a convolution kernel of size s, the s/2 area around the input feature map of the operation is filled with 0. In this way, the neurons in the feature map can have a direct correspondence with the original region.

After the above improvements, the final output feature of the feature extraction network we obtained is the output of the conv5 layer, which mainly contains 256 feature maps, and each feature map is 1/16 of the original image. That is to say, in the obtained feature map, each pixel corresponds to a maximum of 39*39, 11*11 receptive regions with 4 sliding steps in the original image. Each pixel of the conv5 feature map has a large receptive field in the corresponding original image, and the overlap of adjacent regions is very high. But on the other hand, a pixel in the conv5 feature map is only a local feature descriptor, and different feature channels correspond to the multi-channel high-level features of the original image's receptive area. The position feature value of each corresponding pixel of the output 256 feature maps is serially operated, and the final result is the descriptor of the image corresponding to the receptive area. In short, the collection of local image blocks retains the structural information of the entire target, and each local image block also integrates the multi-channel feature information of the same local range. This method can express both global structure and local information well.

In this article, the training set is labeled as $\mathrm{D}=\{(x_i,l_i)\}_{i=1}^{N}$, where x_i represents the sample, its size is P*Q, l_i is the category label and $l_i \in \{-1,1\}$, Calculate the feature of each sample through the improved feature extraction network, and express it in the way of image block collection, which can be expressed as:

$$\mathrm{O}(x_i)=\{o_e(x_i): o_e(x_i): \in \mathrm{R}^{256}, \mathrm{e} \in [1, \mathrm{p} * \mathrm{q}]\}$$

Among them, $o_e(x_i)$is the CNN descriptor of a partial image block of the sample; p=P/16 and q=Q/16 respectively represent the length and width of the final output feature layer. After the image is expressed as a collection of partial image blocks based on CNN features, high-level features are beneficial to increase the discrimination ability of partial image blocks, and adjacent overlapping image blocks provide mutually complementary classification information[5]. This method is very helpful to build a more flexible detection model, which can be used to deal with the appearance differences caused by various intra-class changes within the target.

2.2 Build a target detection model based on CNN local image blocks and GBRF

The local classifier formed by combining multiple CNN local image blocks through GBRF constitutes a strong target classification model, and then the common sliding window method is used to detect the target. The above mainly solves the problem of the expression of the target image block, and the following is to find and combine the image blocks with the most distinguishing ability through GBRF, and process them in two aspects.

2.2.1 Tree node splitting function based on local image blocks

In a traditional random forest, tree nodes are defined by selecting a single feature dimension of a data sample, and then comparing the attribute values on this dimension[6,7]. The traditional random forest tree node splitting function can achieve good results when the data of each feature dimension is not highly correlated, but when each feature dimension of the data has a strong correlation, the real data distribution boundary cannot be determined very much. Good performance will eventually lead to low generalization ability. In single-feature division, you can increase the input depth and trees to make the division boundary closer to the real edge, but with it, the parameters of the division become larger, but in the multi-feature division, the smaller forest scale You can get good results.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0016

The specified scenes in this article are complex scenes with high correlation between the local features of the target image, but big differences between the global ones. In other words, no matter whether the division operation is based on a single or based on the overall image features, good results cannot be obtained. In this paper, GBRF-based and CNN-based local feature blocks are used to divide multiple feature comparisons based on local feature blocks. In the tree growth process, local image blocks with high discrimination are selected layer by layer and combined to form a classifier. Based on the above, we define a node splitting function based on local image block comparison. First of all, selecting a local image block with discrimination is beneficial to dividing the sample set; then, selecting the image block set from the sample set and comparing and dividing. Each image block is expressed as a 256-dimensional feature vector, and linear transformation is used to achieve this multi-dimensional feature comparison problem. A linear transformation is used to map the multi-dimensional feature vector of the selected image block to a divisible real-valued space, and the division threshold is found in the mapping space to divide the sample. Finally, the proposed node splitting function is parameterized as O=(h,φ,λ). Among them, h is the index of the selected partial image block in the sample image, φ is the multi-dimensional feature projection vector, λ is the division threshold of the projection space, and the split function is defined as:

$$s(O(x); h, \varphi, \lambda) = \begin{cases} 0 & \text{if } \varphi . o_h\ (x) > \lambda \\ 1 & \text{otherwise} \end{cases}$$

In this paper, by selecting local location blocks at different locations, the maximum sample class purity is found for division, and in each iteration of the optimization process, the comparison and division based on the selected local image block set are completed through a linear transformation, using SVM to learn the required linearity Transformation can make the reception sample have the largest class spacing in the projection space, which is helpful for division. Finally, by adjusting the projection space to realize the division threshold function, the optimal division in an iteration process can be found. After completing the traversal of all the candidate local image blocks, the global optimal division is the division corresponding to the maximum class purity in the entire iteration process.

For the node sample set$D = \{(x_i, l_i)\}_{i=1}^{N}$, firstly use the CNN network to calculate the conv5 layer output features of each sample, and define each sample in each D as a CNN local image block feature set; secondly, select an image block index, and then remove it according to the index The feature and label of the image block corresponding to each sample in the sample set, and the linear SVM is used to learn the projection vector in the collected image block data set[8]; again, the projection vector is used to map the image block to the projection space, and different division thresholds are selected. The samples are divided, and the division with the highest-class purity is recorded; finally, all possible selected partial image blocks are traversed, and the above operations are performed. In the whole process, the image block index, projection vector, and division threshold corresponding to the maximum class purity division are the optimal division parameters for the node. In particular, this chapter uses the LIBSVM toolbox to learn linear transformations, and the penalty coefficient is set to 0.5 during the learning process. In this method, the CNN local image descriptor provides a stable expression of the local area of the image, especially in the local area deformation problem. The leaf node division function alternately selects the most distinguishable image block and the Learn the linear projection with the optimal division plane. The combination of the two allows the tree to achieve the optimal division of samples when the depth is not large.

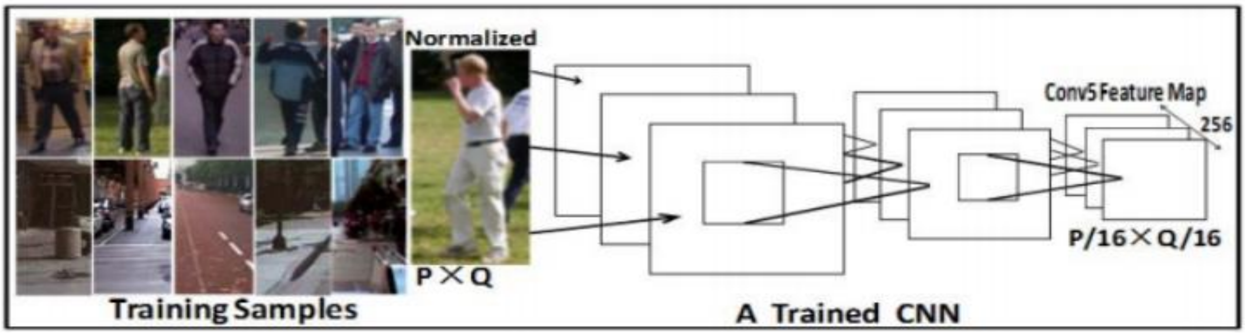
After expressing each sample image as a set of local data blocks and defining the node splitting function based on the local image blocks, the required detectors are generated using the above-mentioned GBRF-based target detection system construction method. In the execution stage, first slide the detection window of a predetermined size in the test image, use the trained GBRF classification model to classify and judge each candidate window, and determine the position of the detection frame judged as the target. The whole process is divided into the following steps:

(1) In the data preparation stage, first calculate the conv5 layer feature map of each sample in the training set through the network; then use the formula to express each sample image as a CNN image

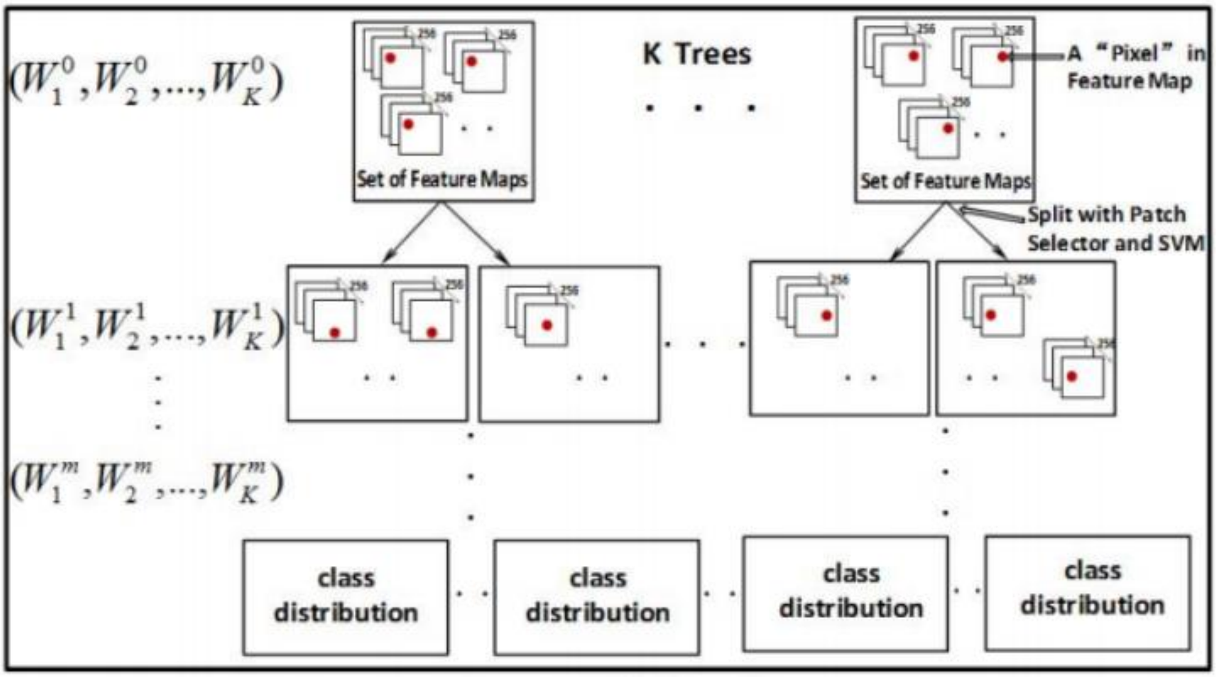
ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0016

collection, and mark the processed sample as $\overline{D}=\{(O(x_i),l_i)\}_{i=1}^{N}$; finally, use the Bootstrap method to Draw a sample set $\overline{D_K}$ for each tree T_k separately.

(2) Construct the GBRF model. First initialize the sample weight W_K^0 of the data set $\overline{D_K}$ of each tree (representing the weight vector of the middle sample in the root node layer): Then, starting from the root node layer of the forest, use the node split method proposed in this article to split each node of this layer. The latter forest with a depth of 1 constitutes a weak classifier, and its classification structure is defined by the decision result of each tree with a depth of 1 in the forest; then, the weak classifier is used to predict the class probability of each sample, which is defined based on the Tangent loss function The prediction error of the sample is updated by the Gradient Boosting algorithm in the classification weight of the node in the first layer; finally, the node that splits the current layer is continuously updated alternately to update the classification weight of the underlying layer until the leaf node meets the termination condition, or The maximum depth of the tree is reached. The entire training process is shown in Fig.1.



(a) Calculate CNN features



(b) Training GBRF based on CNN image block features

Fig. 1 Construction process of GBRF target classification model based on CNN partial image blocks

(3) After training the GBRF classification model, the detection method adopts the standard sliding window method. It is specifically embodied in input the test image into the convolutional neural

network to obtain a 256-channel conv5 feature map; then slide the detection frame pixel by pixel in the feature map, and the detection frame is in the form of a collection of CNN image blocks; finally, the detection frame The CNN image block collection method is placed in each tree of GBRF. Starting from the root node, the left and right sub-nodes are continuously selected to traverse the tree according to the division parameters stored in the tree nodes such as the index number of the image block, linear projection, and division threshold, etc. A certain leaf node is finally predicted based on the class distribution probability of reaching the leaf node. The prediction result of GBRF is the average of all tree prediction structures.

$$T_k(\tau) = \mathrm{p}(\mathrm{y} = 1|\mathrm{Leaf}_k(\tau))$$

$$\mathrm{Score}(\tau) = \frac{1}{\mathrm{K}}\sum_{k=1}^{K} T_k(\tau)$$

Among them, the frame under test τ in the conv5 feature map traverses the tree T_k and reaches the leaf node, $\mathrm{p}(\mathrm{y} = 1|\mathrm{Leaf}_k(\tau))$ represents the probability of a positive sample in the leaf node. According to the GBRF prediction result Score (τ) and the classification threshold, it is judged whether the frame to be tested is the target.

Since the detection is performed in the conv5 layer feature map, it is assumed that the detection frame with the center pixel position (u, v) is judged as a positive example. According to the correspondence between the conv5 feature map and the original image coordinates, the center of the target area in the original detection image It is (16u, 16v), the size is 16p×16q, and the target detection is completed.

3. Experiments and Results Analysis

3.1 Dataset

(1) TUD pedestrian data set[9]: This database is a commonly used database for pedestrian detection, including 400 pedestrian training samples and 250 test images. The number of positive samples in the database is too small, and the background image is relatively single, so when training the pedestrian detection model, the training samples of the TUD pedestrian data set and the training samples of the INRIA data set are merged to finally train the common pedestrian detection model on the two databases.

(2) INRIA pedestrian data set[10]: The pedestrians in this data set are affected by factors such as angle, clothing, lighting, occlusion, and their appearance characteristics vary greatly. At the same time, the background image contains various scenes in daily life, such as streets, shopping malls, and airports. Wait. In this training set, there are 614 images containing pedestrians, among which 1208 pedestrians are calibrated, and there are 1218 background images without pedestrians. In order to enhance the training data, we flipped the 1208 calibrated pedestrians horizontally, and finally obtained 2416 positive samples, and negative samples were randomly selected from 1218 background images. The INRIA pedestrian test set contains 288 images, in which the number of identifiable pedestrians marked is 589.

(3) UIUC vehicle data set: This data set contains a training set and two test data (single-scale test set and multi-scale test set respectively). There are 550 vehicle samples and 500 negative samples in the training set. The single-scale test set has 170 test images and the number of vehicles to be detected is 200. The multi-scale test set has 108 test images and the number of vehicles to be detected is 139.

3.2 Experiments and Results

When pedestrians are detected, in order to ensure that each pixel in the conv5 layer feature map calculated by CNN has a different perception area in the original image, we standardize the size of the image sample to 192*192, and the corresponding output feature map is 12*24. Similar When training the vehicle detector, the sample graph is regularized to 432*176, and the output feature map is 27*11. Set the number of trees in GBPF to 100, the maximum depth of the forest to 5, and the tree node split termination condition, split the minimum number of samples in the node is 10, and the maximum class purity is defined as 99%. According to the characteristics of the data set, the test

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0016

image scale scaling factor in TUD pedestrians is set to 24 different scales between 1 and 7.5, and the test image scale scaling factor in INRIA pedestrians is set to 30 different scales ranging from 0.8 to 8.5. On the scale. Corresponding to the UIUC vehicle multi-scale test set, the image scale scaling factor is set at 20 different scales between 0.9 and 2.5. The experimental platform of this chapter: Inter Core (TM) i5 3.2GHz CPU, 16G RAM, Matlab R2012b, Window 64-bit OS.

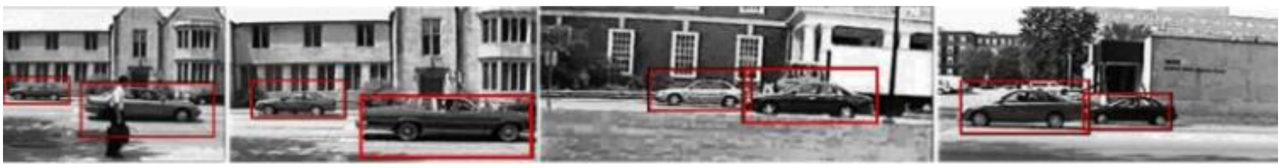
In the detection process, we follow the PASCAL protocol for the suspected detection frame that exceeds the detection threshold, that is, if the detection frame and the target real area (given by the label information of the test set) exceed 50%, it is considered a success Detection. In order to avoid multiple detection frames for the same target, this chapter uses the non-maximum suppression (NMS) method to discard the detection frames that have low scores and appear inside the detection frames with high scores.

In order to verify the feasibility and effectiveness of the scheme, we conducted related experiments on three common target detection data sets of TUD pedestrians, INRIA pedestrians and UIUC vehicles. The results are shown in Fig 2.

(a) Test results on the TUD pedestrian test set

(b) The structure on the INRIA pedestrian test set

(c) The detection results on the UIUC vehicle single-scale test set



(d) The detection results on the UIUC vehicle multi-scale test set

Fig. 2 The detection results of the algorithm in this paper on each data set

Fig. 2(a) and Fig.2(b) show the detection examples of the proposed algorithm on two pedestrian datasets, TUD and INRIA. It can be seen that although the test images contain pedestrians with different postures, different viewing angles, partial occlusions and different lighting conditions, the proposed detection algorithm shows strong adaptability.

 Volume 7 Issue 9, 2021
DOI: 10.6919/ICJE.202109_7(9).0016

In order to verify its effectiveness, we use different feature expression and node division strategies to design the following four detection models, and compare the detection performance on the TUD pedestrian data set.

(1) The method proposed in this paper is established based on the node splitting function of CNN feature and image block comparison. Here, the method is marked as GBRF+P-CNN.

(2) When training the GBRF model, use a similar CNN network to calculate the sample features, the difference is that the conv5 pooling layer pool5 and the 6th fully connected layer fc6 will be retained, and the 4096-dimensional CNN feature vector output by f c6 will be used to express the sample image . Based on such image expression, the node splitting function uses a single feature dimension value to compare and define. Finally, the GBRF classification model is trained using the above image expression and node splitting method, which is labeled GBRF+S-CNN.

(3) The training samples are regularized to 48×96, and the HOG feature is an image descriptor based on local blocks. Each sample is expressed as a 36-dimensional HOG local image block set. Each local block is composed of 2×2 adjacent 8×8 pixel cells, and each pixel in the cell is digitized to 9 main directions according to its gradient direction and intensity, that is, each local block can be expressed as 36 dimensions HOG feature vector, one cell overlaps between adjacent blocks, and finally each sample is expressed as a set of 5×11 36-dimensional HOG local image blocks. With such image expression, the node splitting function is defined based on the 36-dimensional HOG partial image block comparison, which is the same as the node splitting function definition method in method (1), except that 256 is the CNN partial image block used in (1). Use the above image expression and node splitting method to train the GBRF classification model, which is labeled GBRF+P-HOG.

(4) The training samples are regularized to 48×96, and the global HOG feature of each sample is calculated, that is, each sample is expressed as a 1980-dimensional feature vector. Based on such image expression, the node splitting function uses a single feature dimension value to compare and define. Finally, the GBRF classification model is trained using the above image expression and node splitting method, which is labeled GBRF+S-HOG.

Fig.3 shows the detection performance of the GBRF detection model on the TUD pedestrian data set under different feature expression and node division strategies (PR curve description).

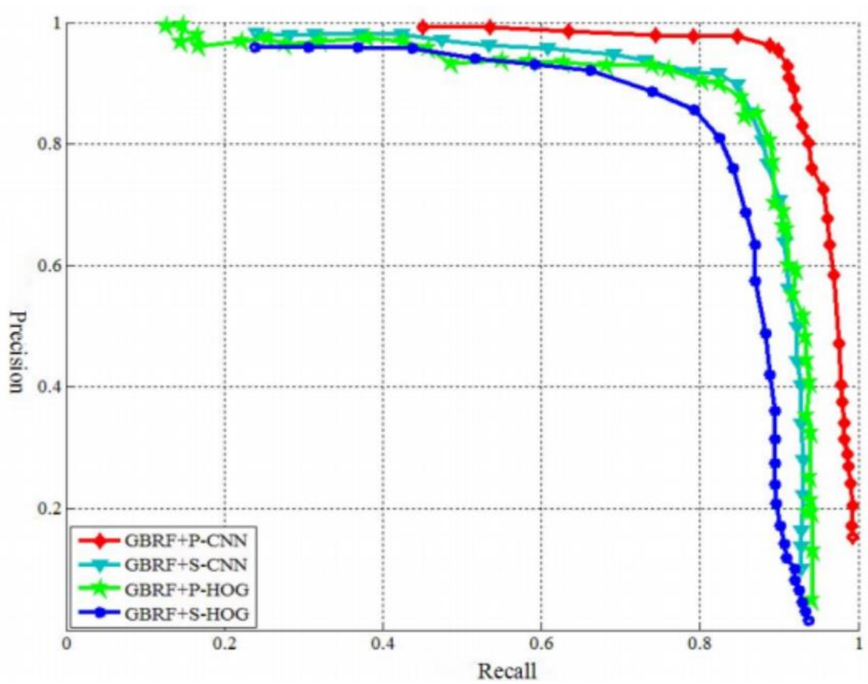


Fig. 3 Comparison of algorithm detection performance under different features and partition strategies

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0016

It can be seen that GBRF+P-CNN (that is, the solution proposed in this article) has achieved the best detection results. GBRF+S-HOG has the worst detection results. By comparing the detection results between GBRF+P-CNN and GBRF+P-HOG, GBRF+S-CNN and GBRF+S-HOG, it effectively proves that CNN features have stronger expressive power than traditional artificial design features. By comparing the detection results between GBRF+P-CNN and GBRF+S-CNN, GBRF+P-HOG and GBRF+S-HOG, it shows that the image block can effectively retain the local structure information of the image. The node division strategy based on the image block is as follows: Stronger classification capabilities.

4. Summary

In this paper, under the framework of "GBRF+sliding window" target detection, the target detection method is studied from the perspective of improving image expression ability. First, we will study in depth how to use CNN to extract high-level image expressions, and use it to replace traditional artificially designed image features; then, in order to cope with the appearance changes caused by changes in the target class, we propose to model the target by discovering and combining local image blocks. To this end, we define a local image block descriptor based on CNN features; finally, use GBRF to discover and combine the local image blocks defined by CNN layer by layer. In order to optimize the results of each node division, this chapter proposes a linear SVM-based image block multi-dimensional feature mapping method, and defines the node division method in the projection space. The experimental results on three common target detection data sets of TUD pedestrians, INRIA pedestrians and UIUC vehicles verify the effectiveness of the proposed method.

References

[1] R.B.Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Columbus: IEEE Conference on Computer Vision and Pattern Recognition, 2014:580–587.

[2] Y.Zhang, K.Sohn, R.Villegas,et al. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction [C]// Boston: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[3] C.Szegedy, A.Toshev, D.Erhan. Deep neural networks for object detection [C]// Nevada: Advances In Neural Information Processing Systems, 2013:676–685.

[4] R.B.Girshick, F.Iandola, T.Darrell, et al. Deformable part models are convolutional neural networks [C]// Boston: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[5] W. Nam, P.Dollar, J.H. Han. Local Decorrelation For Improved Pedestrian Detection [C]// Quebec: Advances In Neural Information Processing Systems, 2014:1127–1136.

[6] Y. Freund. Experiments with a new boosting algorithm [C]// Morgan Kaufmann: International Conference on Machine Learning, 1996:148–156.

[7] C.Y.Dong, F. K.Yan, L.W.Cong, et al. Using LogitBoost classififier to predict protein structural classes[J]. Journal of Theoretical Biology, 2006, 238(1):172-176.

[8] H. Masnadi-Shirazi,V. Mahadevan,N. Vasconcelos. On the design of robust classifiers for computer vision [C]//San Francisco: IEEE Conference on Computer Vision and Pattern Recognition,2010:779–786.

[9] J.Li, T.Wang, Y.Zhang. Face detection using SURF cascade [C]// Barcelona: IEEE International Conference on Computer Vision Workshops, 2011:2183–2190.

[10] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. Improving neural networks by preventing co-adaptation of feature detectors. arXiv, 2012.