Pedestrian Detection Using Visual Saliency and Deep Learning

Feng Xiao*, Baotong Liu

School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, Shaanxi, China
*Corresponding author(E-Mail: xffriends@163.com)

Abstract
This paper explores the use of Visual Saliency to detect pedestrian for images. Our method computes visual saliency map from the image, then the input image is multiplied by the multiplied visual salient map and product is fed to the Convolutional Neural Network. Using deeply supervised network with short connections (DSS) to compute visual saliency map and pedestrian detection is carried out using Faster R-CNN. DSS Network is trained on MSRA and ECSSD datasets, Faster R-CNN is trained on three challenging databases - Penn-Fudan, INRIA and Daimler datasets. Experimental results demonstrate that the proposed achieves state-of-the-art performance on Penn-Fudan Dataset with 91% detection accuracy and it achieves average miss-rate of 15% on the INRIA Dataset and achieves average miss-rate of 28% on the Daimler Dataset.

Key words: Visual Saliency, Convolutional Neural Network, Pedestrian Detection, Short Connections

1. Introduction

Pedestrian detection is an important research topic in computer vision including intelligent video surveillance, vehicle assisted driving and mobile robots [1], [2], [3]. Pedestrian detection is challenging and needs to overcome complex backgrounds, pedestrian pose, low resolution, light conditions and other factors. Over the last decade, pedestrian detection has made great progress. Many methods describe pedestrians by extracting image features, such as HOG [4], Shapelet [5], and their combination [6]. Some methods use features and classifiers to detect pedestrians, such as Adaboost-HOG [7], SVM-HOG [8], Adaboost [8], Latsvm-v1 [9], MultiFtr+CSS [10], Latsvm-v2 [11]. These methods are based on hand-crafted features.

Deep learning automatically extracts image features by supervised learning or unsupervised learning without the use of hand-crafted features. With the development of artificial intelligence, deep learning has become the focus of current research [12]. Currently, deep learning has achieved rich research results in the field of pedestrian detection [1], [2], [13]. Zhang et al. [14] improved Faster-RCN [15], using a combination of region proposal network (RPN) and Boosted Forest to detect pedestrians and solve the disadvantages of Faster-RCNN for detecting small target pedestrians, this method provides better detection accuracy over multiple datasets. Ouyang et al. [16] proposed a discriminative pedestrian detection model that uses the independence and visibility rules between locations to detect pedestrians. Ouyang et al. [17] proposed a joint deep learning pedestrian detection model, which combines the four components to make the network play its biggest advantage. However, these methods cannot deal with pedestrian detection in strong occlusion and complex scenes.

Pedestrian detection is still an open problem considering complex scenes and strong occlusion. Inspired by the attention mechanism and pedestrian detection using deep learning to achieve competitive results, this paper proposes a pedestrian detection method combining visual saliency and deep learning. The main goal is to improve the detection accuracy in difficult backgrounds and complex scenes. In this paper, the DSS model proposed by Hou et al. [18]is used for saliency prediction. Then the saliency map is multiplied by the original image to obtain a multiplied visual salient image (MVSI). Finally, using MVSI and labels to train Faster R-CNN. The detector using Faster R-CNN [15], and it is pre-trained on ResNet 101 [42].

The major contributions of this paper can be summarized as follows:

- We proposed a new pedestrian detection method to learn features after computing visual saliency, which can accurately locate the position of pedestrians in strong occlusion and complex scenes.
- We achieve state-of-the-art results on three public pedestrian datasets: Penn-Fudan dataset, INRIA dataset and Daimler dataset.

2. Related Work

2.1. Visual Saliency Object Detection

Visual saliency detection is very close to the selection process of the human visual system. It aims to identify the most salient objects or regions in an image. After several decades of development, the visual attention mechanism has proposed and established a variety of visual attention models. Many supervised and
unsupervised visual saliency detection methods are based on these models [19], [20], [21]. The unsupervised method mainly extracts the underlying features of the image for saliency detection.

Itti et al. [20] proposed visual attention based on saliency which extracts features such as color and brightness on multiple scales, and establishes a visual model through fusion strategy. Harel et al. [22] proposed Graph-Based Visual Saliency (GBVS), which uses Markov random field to calculate Markov chain to enhance the saliency of images. Zhu et al. [19] calculated the saliency map using the background method with image region spatial distribution. Yan et al. [23] used a multilayer structure for salient detection.

Inspired by human visual attention, in recent years, methods that combine attention mechanism and deep learning are increasingly applied in image recognition [24], fine-grained classification [25], machine understanding [26]. Mavani et al. [24] recognize facial expressions using visual saliency and deep learning, which uses a deep multi-layer network [41] to compute facial saliency maps and feed them into a face recognition network to further observe the differences between different facial expressions. Gurnani et al. [27] classify age, gender, and facial expressions using a convolutional neural network for visual saliency, which uses a deep multi-layer network to compute a saliency map of the input image and feed it into a classification network. This approach simplifies visual saliency and classification studies. Gajjar et al. [28] detected humans using a visual attention mechanism under a convolutional neural network, which also uses a deep multi-layer network to calculate the significance of the image, and then uses the detection network to detect the human body. Zhang et al. [29] detecting occluded pedestrians through guided attention in CNNs, the method also proposed three attention mechanisms: self-attention network, visible frame attention network and part attention network.

The self-learning ability of the convolutional neural network to feature makes the advantage of the visual attention mechanism in extracting the object region of interest more prominent. Therefore, in order to automatically predict visual saliency maps and learning features, we use the depth-supervised network proposed by Hou et al. which uses a series of short connections to connect the deep-side output to the shallow-side output. It not only highlights significant targets but also pinpoints target boundaries.

2.2. Pedestrian Detection

After more than ten years of development, pedestrian detection has accumulated considerable research results. CNN has achieved great success in object detection [30][31]. The early application of CNN to pedestrian detection is based on R-CNN [32] structure [33], [34]. These methods rely on high-quality region proposal to obtain good detection results. The influence of Faster R-CNN [15] in the field of object detection cannot be ignored.

Zhang et al. [14] combined the RPN with Boosted Forest to solve the problem that Faster R-CNN can't handle small target pedestrians well. This method provides competitive accuracy and speed on multiple datasets. Zhang et al. [29] added three attention mechanisms to Faster R-CNN [15] to explore the relationship between feature channels and pedestrian body parts, and to solve the pedestrian detection problem under occlusion. This method has a good performance on multiple complex datasets. Zhang et al. [3] proposed the CityPersons pedestrian dataset, which is more diverse. By changing the RPN scale and optimizer of Faster R-CNN, the accuracy of Faster R-CNN in pedestrian detection is improved.

Ren et al. [15] used RPN to share convolutions with the detection network and generate high quality regional proposals with little time consuming. Therefore, we used Faster R-CNN as a follow-up detection network for visual salient.

3. The Proposed Method

The detailed representation of our method is shown in Figure 1. The process uses a DSS network [18] to compute a saliency map of the three dataset images, and the CNN-based network is used to extract the necessary features of the image. After computing the saliency maps, we multiply it to the input image to generate a multiplied visual salient image (MVSI). Train the Faster R-CNN [15] by using labels and MVSI.
3.1. The Saliency Model

In recent years, significant progress has been made in visual attention, but there are still problems in processing scale space. It is found that the Holistically-Nested Edge Detector (HED) [35] can solve the scale space problem better. It mainly solves the edge detection problem, and its application to the visual attention mechanism still needs to be improved. After a further study, it is found that the short connection can solve the visual saliency problem of the HED model. Therefore, this paper establishes a top-down visual attention model through short connections to solve the visual saliency problem of the pedestrian scale space.

HED Model [35] detects edges by learning rich multi-scale features. The main idea of the HED model is to add a side output layer after the convolutional layer of the VGG Network [36]. Each side output layer outputs an edge map, and finally merges the edge maps of all side outputs to obtain edge detection results. The HED model is shown in Figure 2.

HED [35] model is suitable for edge detection of images and still needs improvement in visual saliency detection. We analyzed the DSS model and found that the deep side output can better locate the significant regions, but the detail information is lost; the shallow side output can extract the underlying features of the
image better, but the global information of the image is lost. The deep side output and the shallow side output are properly combined to construct a top-down visual saliency model to refine the salient objects. In the short connection structure, the deep side output is up-sampling using bilinear interpolation to enable fusion of feature maps of different scales. Figure 4 shows the short connection structure.

Based on the above analysis, we redefine the side output activation function, the side output loss function, the fusion loss function, and the final loss function. Equation 1 is the side output activation function. Equation 1 is the side output activation function.

\[ R_{\text{side}}^{(m)} = \begin{cases} \sum_{i \leq m} r_{i}^{m} R_{\text{side}}^{(i)} + A_{\text{side}}^{(m)}, & m = 1, \ldots, 5 \\ A_{\text{side}}^{(m)} , & m = 6 \end{cases} \]  \hspace{1cm} (1)

Where \( r_{i}^{m} \) represents the short connection weight from side output \( i \) to side output \( m \) \( (r > m) \), and \( A_{\text{side}}^{(m)} \) are activation of the \( m \) th output. The short connection weight is zero if the current layer does not have a short connection. The side output loss function and the fusion loss function are defined by Equation 2 and Equation 3, respectively.

\[ L_{\text{side}} (W, w, r) = \sum_{m=1}^{M} \alpha_{m} f_{\text{side}}^{(m)}(W, w^{(m)}, r) \]  \hspace{1cm} (2)

\[ L_{\text{fuse}} (W, w, f, r) = \sigma(Z, \sum_{m=1}^{M} f_{m} R_{\text{side}}^{(m)}) \]  \hspace{1cm} (3)

Where \( \alpha_{m} \) is the weight loss of the \( m \) th side output, \( r = \{ r_{i}^{m} \}, i > m \), \( f_{\text{side}}^{(m)} \) is cross-entropy loss function for the \( m \) th side output.

\( W \) is a collection of all network layer parameters. \( Z_{\text{m}} \) is a Ground Truth (GT) associated with the input image \( X_{\text{m}} \). \( \sigma(\cdot, \cdot) \) represents the distance between the GT and the side output prediction, which is set to the image-level class-balanced cross-entropy loss. Based on the definition of the loss function of Equation 2 and Equation 3, the total loss function of top-down visual attention is redefined by Equation 4.

\[ L_{\text{final}} (W, w, f) = L_{\text{side}} (W, w, r) + L_{\text{fuse}} (W, w, f, r) \]  \hspace{1cm} (4)

The short connection directly affects model performance. How to connect the deep-side output to the shallow-side output to increase model learning performance is difficult. So we explore two different short connections, one that connects the deep side output to the upper side output and the other that connects the deep side output to the multilevel side output. Different short connections are shown in Figure 4. Figure (a) is a simple connection, Figure (b) is a multi-layer connection. For different short connection methods, this paper made a comparative test, as shown in Figure 5. Experiments have found that the visual saliency effect of multilayer connections is better than the visual saliency of simple connections.

Figure 3. Short connection structure

\[ 16 \times 16 \]  \hspace{1cm} \( \text{conv} \)  \hspace{1cm} 16 \times 16 \hspace{1cm} \text{loss} \\
\[ 32 \times 32 \]  \hspace{1cm} \( \text{conv} \)  \hspace{1cm} 32 \times 32 \hspace{1cm} \text{loss} \\
\]
Pedestrian Detection Using Visual Saliency and Deep Learning

Figure 4. Short connection structure

Figure 5. The result of the different short connection

Based on the above study of the top-down visual attention mechanism model, Figure 6 is a visual attention model structure with short connections.

Figure 6. Visual attention model flow chart

The short connection test results are shown in Figure 7. Experimental data was obtained from the Penn-Fudan pedestrian database [37]. We can see from the experimental results that the short connection visual attention model can extract significant regions well even in complex scenes and occlusions.
We multiply the significant image by the original image to obtain a significant pedestrian region, so that subsequent detection models can detect pedestrians in the image. In this paper, the results of the salient and original images are recorded as multiplied visual saliency images, and the pedestrian multiplied visual saliency map is shown in Figure 8.

**Figure 7. Saliency map of visual attention model**

**Figure 8. Pedestrian multiplied visual saliency map**

### 3.2. Pedestrian Detection Model

The top-down visual attention model with short connection can better extract the pedestrian saliency region in the image. For the obtained visual saliency map, the saliency region is detected by the subsequent object detection network to give the pedestrian position and confidence. In recent years, deep learning has developed very rapidly in the field of object detection, and has achieved remarkable results.

Faster R-CNN [15] improves the detection accuracy by calculating the candidate frame of the feature map through the Region Proposal Network (RPN). It solves the bottleneck of the regional proposal and implements end-to-end network training. YOLO [31] meshes the input image and predicts the object position and confidence within the grid at one time, solving the real-time problem of object detection. R-FCN [30] addresses image location sensitivity issues using a shareable location-sensitive score graph. SSD [38] combines multi-feature mapping with different resolutions to process objects of different sizes, and its time and detection accuracy are greatly improved. Faster R-CNN [15] and YOLO [30] directly connect to the fully connected layer after the convolutional layer, using only the highest layer feature map, while the SSD [38] network uses the feature pyramid structure for object detection, and performs Softmax classification and position regression simultaneously on multiple feature layers.
The advantage of Faster R-CNN [15] is that the Region Proposal Network (RPN) shares the convolution with the detection network and generates high-quality regional recommendations with little time consuming. Therefore, we use Faster R-CNN [15] as a visually significant follow-up detection network. The loss of RPN is composed of classification loss and regression loss. The classification loss is calculated by binary cross entropy loss, the regression loss is calculated by using Smooth L1 loss, and the RPN loss function is defined as formula (5).

\[
L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg}(t_i, t_i^*)
\]

(5)

Where \(i\) is the index of the anchors, \(p_i\) is the probability value of the anchor. The label \(p_i^*\) of ground-truth is 1 if the anchor is a positive sample, 0 otherwise. \(t_i^*\) is the parameterized coordinate of the predicted border, and \(t_i\) is the coordinate of the ground-truth. \(L_{cls}\) is the loss of the target/non-target class, and \(L_{reg}\) is the regression loss.

4. Experiments and Results

4.1. Datasets

We evaluate our approach in three public pedestrian datasets: Penn-Fudan dataset [37], Daimler dataset [39], and INRIA dataset [4]. There are 170 samples in the Penn-Fudan dataset, 15560 train samples and 4800 test samples in the Daimler dataset, and 1832 train samples and 741 test samples in the INRIA dataset. The images in the Penn-Fudan dataset mainly come from campus scenes and urban street scenes. The images in the Daimler dataset mainly come from road scenes. The images in the INRIA dataset mainly scenes include cities, mountains, airports and beaches. Our experimental data has 23,103 images, of which 17,522 are used for model training (130 from Penn-Fudan, 15560 from Daimler, 1832 from INRIA), and 5,598 from model tests (40 from Penn-Fudan, 4800 from Daimler and 741 from INRIA).

4.2. Evaluation Metrics

We used two experimental evaluation metrics for experimental analysis: detection accuracy and Miss Rate. We mainly detect pedestrians in static images without using video information. Penn-Fudan dataset uses detection accuracy as an evaluation metric, and the Daimler dataset and the INRIA dataset use the missing rate as an evaluation metrics. We compared the other 12 methods to verify our method. Among them, there are 7 comparison methods in the Penn-Fudan dataset, 7 comparison methods in the Daimler dataset, and 8 comparison methods in the INRIA dataset.

4.3. Ablation Experiments

We used three datasets to train the model in order to improve detection performance and then tested on three data sets. Because different dataset scenarios and perspectives are different, the network can learn more pedestrian characteristics.

We explored the performance of the model using the detection accuracy for Penn-Fudan Dataset and compared with the other six methods: HOG [4], Adaboost-HOG [7], SVM-HOG [7], Adaboost [8], R-FCN [30], YOLO [31]. The experimental results are shown in Table 1. Our method improved accuracy by over 3% compared to the YOLO, and even more for HOG.

Table 1. Comparison with the different approach performance on Penn-Fudan Dataset

<table>
<thead>
<tr>
<th>Approach</th>
<th>Detection accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG[4]</td>
<td>67%</td>
</tr>
<tr>
<td>Adaboost-HOG[7]</td>
<td>85%</td>
</tr>
<tr>
<td>SVM-HOG[7]</td>
<td>78%</td>
</tr>
<tr>
<td>Adaboost[8]</td>
<td>72%</td>
</tr>
<tr>
<td>R-FCN[30]</td>
<td>86%</td>
</tr>
<tr>
<td>YOLO[31]</td>
<td>88%</td>
</tr>
<tr>
<td>Ours</td>
<td>91%</td>
</tr>
</tbody>
</table>

We explored the performance of the model using the miss rate for INRIA Dataset and compared with the other six methods: VJ [40], HOG [4], Latsvm-v1 [9], HOG+LBP [6], MultiFtr+CSS [10], Latsvm-v2 [11]. The experimental results are shown in Table 2. From the experimental results, the Miss rate of our method is 15%, which is 57% lower than the VJ method, 24% lower than the HOG+LBP method, and 5% lower than the Latsvm-v2.
Table 2. Comparison with the different approach performance on INRIA Dataset

<table>
<thead>
<tr>
<th>Approach</th>
<th>Miss Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>VJ[40]</td>
<td>72%</td>
</tr>
<tr>
<td>HOG[4]</td>
<td>46%</td>
</tr>
<tr>
<td>Latsvm-v1[9]</td>
<td>44%</td>
</tr>
<tr>
<td>HOG+LBP[6]</td>
<td>39%</td>
</tr>
<tr>
<td>MultiFtr+CSS[10]</td>
<td>25%</td>
</tr>
<tr>
<td>Latsvm-v2[11]</td>
<td>20%</td>
</tr>
<tr>
<td>Ours</td>
<td>15%</td>
</tr>
</tbody>
</table>

We explored the performance of the model using the miss rate for Daimler Dataset and compared with the other seven methods: VJ[40], Shapelet[5], HOG[4], Latsvm-v1[9], HOG+LBP[6], MultiFtr+CSS[10], Latsvm-v2[11]. The experimental results are shown in Table 3. The miss rate of our method is 28%, 30% lower than the Latsvm-v1 method, 11% lower than the MultiFtr+CSS method, and 10% lower than the Latsvm-v2.

Table 3. Comparison with the different approach performance on Daimler Dataset

<table>
<thead>
<tr>
<th>Approach</th>
<th>Miss Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>VJ[40]</td>
<td>95%</td>
</tr>
<tr>
<td>Shapelet[5]</td>
<td>94%</td>
</tr>
<tr>
<td>HOG[4]</td>
<td>60%</td>
</tr>
<tr>
<td>Latsvm-v1[19]</td>
<td>58%</td>
</tr>
<tr>
<td>HOG+LBP[6]</td>
<td>49%</td>
</tr>
<tr>
<td>MultiFtr+CSS[10]</td>
<td>39%</td>
</tr>
<tr>
<td>Latsvm-v2[11]</td>
<td>38%</td>
</tr>
<tr>
<td>Ours</td>
<td>28%</td>
</tr>
</tbody>
</table>

The proposed method achieves optimality on all three pedestrian datasets, and our model adapts to strong occlusion and complex scenarios. Figure 9 shows the results of the model in the Penn-Fudan database and the INRIA database. These test samples have strong occlusion and complex scene characteristics.

Figure 9. Qualitative results on some of the test images
5. Conclusions

Our method uses visual saliency to train a detection network, which can improve the performance of the model by feeding the multiplied visual saliency map directly into the detection network. Multiplying the saliency map eliminates background information while enhancing the pedestrian area. Multiply the saliency map into the network, and detect the network you can easily extract pedestrian features. The experimental results show that the performance of pedestrian detection is improved by visual saliency. The detection accuracy on the Penn-Fudan pedestrian data set is 91%, the Daimler pedestrian dataset deletion rate is 28%, and the INRIA pedestrian dataset deletion rate is 15%.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 61572392) and Shaanxi Provincial Natural Science Foundation (Grant No. 2017JC2-08).

References