ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Fast Deformation Part Model with CNN for Face Detection

Canzhang Guo [a)], Yinwei Zhan

School of Computer, Guangdong University of Technology, Guangzhou 510006, China.

[a)] Corresponding author: gczgdut@foxmail.com

Abstract. This paper proposes a fast deformation part model (DPM) with neural network (FDPDPM) for face detection. It can make fully use of the advantage of high level feature and classifier with spatial location information. In fact, this paper uses a truncated the variant of VVGNET and a pyramid of multi-scale filter method to maintain the scale invariance. As a result, it not only improves the detection efficiency, but also makes a contribution to better detection effect on the face with small size and partial occlusion.

Key words: DPM; filter pyramid; scale invariance; face detection.

INTRODUCTION

The Computer vision (CV) is a study of how to make use of computer simulation of the human visual science, its main task is to analyze or understanding the image (or video), so as to make a judgment or decision. In the past few decades, the CV made great progress and development. Among them, the face detection has always been a hotspot in the field of CV.

Traditional face detection methods generally use facial features with LBP [1], SHTIF [2], HOG [3], etc., these features can represent human face to a certain degree. However, these artificial designing features usually unable to capture high-level semantic information of different tasks, which is restricting the face detection performance to further improve. DPM [4] is a classic model of face detection and based on a spatial relation between different faced parts as well as the overall structure of a face, using low-level features HOG and latent SVM classify. But due to using low-level features HOG, DPM is difficult to capture the significant facial information in different poses and illumination conditions [5].

As a popular machine learning method in recent years, deep learning [6-9] has been widely studied and applied in the fields of CV, speech recognition and natural language processing. At present, the study of face detection, based on deep learning, has become a mainstream research direction. Many universities and research institutions at home and abroad, have conducted extensive and depth studies on face detection based on deep learning [10-12]. Compared with the traditional method of face detection, a method based on deep learning can automatically learn facial representation from mass data and effectively improve the performance of the method. Nevertheless, CNN does not provide a clear relationship between the lower levels of features, such as the features of each part of a face. Therefore, especially when dealing with faces, it may lose potential information on candidate relationship structure, which is the important information to improve accuracy. So, integrating CNN and DPM to enhance their advantage would be a promising approach.

In this paper, we propose a fast deep DPM method to deal for face detection by DeepPyramid DPM [13]. Our main contributions are three points. Firstly, the feature extraction network VGGNET [14] with more layers and smaller size filters is used to replace the Supervision CNN [15] to obtain better features and improve the detection effect of small size faces. Secondly, the pyramid of multi-scale filter is constructed to replace the pyramid of feature, reducing the computing burden and improving the detection efficiency while suppressing the scale change. Finally, we study the impact of the number of parts on the detection effect.

RELATED WORKS

Viola and Jones [16] creatively proposed a face detection method based on the adaboost cascade structure. The cascading detector can detect frontal faces well by using the haar-like [17] feature, but it does not perform well for detection of side faces or partially occluded faces. This is due to the rigidity-based approach that is not flexible enough to handle deformable objects. The face can be seen as a collection of parts (eyes, mouths, etc.), and many face detectors employ DPM framework [4]. The paper [18] uses a hybrid tree structure in the DPM framework to overcome the interference due to different viewpoints. Use HOG, haar-like and other low-level features for classification, which will eliminate many useful undiscovered picture information.

It turns out that in dealing with complex tasks, deep models may be more capable than shallow models [19]. In 2012, Krizhevsky [15] demonstrated the effectiveness of CNN for the first time in image classification tasks. Deep learning flourished in the field of target detection. R-CNN [20] used the first detection of CNN for target detection. With the rapid development based on deep learning, its application in face detection is also increasing. Zhang [21] proposed a multi-task face detection method based on CNN, which improves the performance of face detection by constructing the common learning of two assistant tasks: face pose estimation and face key point detection. Li [11] proposed a face detection method based on cascaded CNN, which quickly filters out non-face image regions on multiple scale images and refines the remaining regions to improve face detection performance.

Only a few works achieved the complementarity between DPM and CNN. Ouyang Wanli [19] constructs CNN with a combined input of HOG and YUV images. This CNN structure also has a deformation layer to handle occlusion. However, the structure of this deformed layer is constructed on the basis of pedestrian detection and is not suitable for face detection and does not take into account the effects of scale changes. Girshick [13] theoretically stated that DPM can be represented using a CNN network. However, this is for general object detection and needs to be modified before it can be used in face detection. Both Ranjan [5] and Dinh-Luan [22] applied DeepPyramid DPM [13] to face detection. Ranjan added a layer of normalization to improve the scale invariance of the algorithm. Dinh-Luan has constructed a new comprehensive face characterization model. At the same time, Intuitive non-maximum suppression has been introduced to improve detection accuracy, but the computational complexity is higher. So, we construct a fast deep DPM for face detection.

THE STRUCTURE OF FDPDPM

In this section, we mainly introduce the structure of FDPDPM, which mainly includes feature extraction network, MULTI-DPM-CNN, and non-maximum suppression, as shown in Fig.1

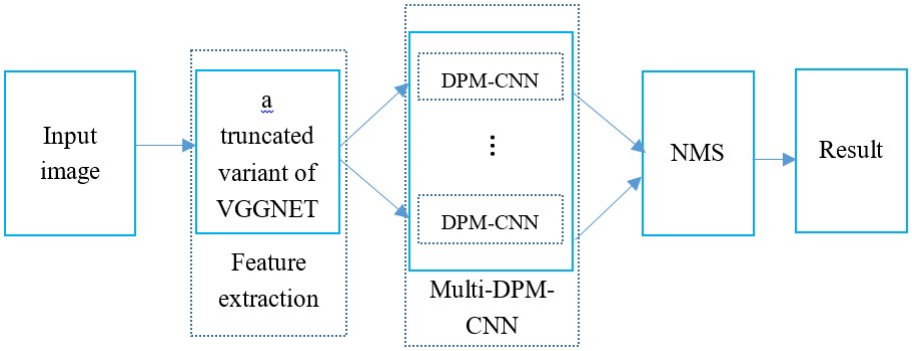

FIGURE 1. The structure of FDPDPM.

Feature Extraction Network

Since an object can appear in the image in many scales, it is necessary to increase the robustness of the algorithm to multi-scale. One way to solve the multi-scale problem criterion is to generate a feature pyramid from the image pyramid so that the detector can detect targets of different scales. This method is used by many algorithms in deep learning, including OverFeat [23], DeepPyramid DPM [13]. As shown in Fig. 2, there is a conv5 map generated at the input of different scales for a truncated variant of Supervision CNN [15] (with five convolutional layers and four

pools) in DeepPyramid DPM. This shows that this network is sensitive to scale changes and needs to consider multi-scale changes. In addition, the SuperVision CNN has a stride of 4×2×2=16, which means that one cell at the position (x,y) of the conv5 layer map corresponds to the pixel (16x, 16y) on the image. This extraction of features is not sufficient because it ignores any bounding box less than 16×16, which is detrimental to the detection of small faces.

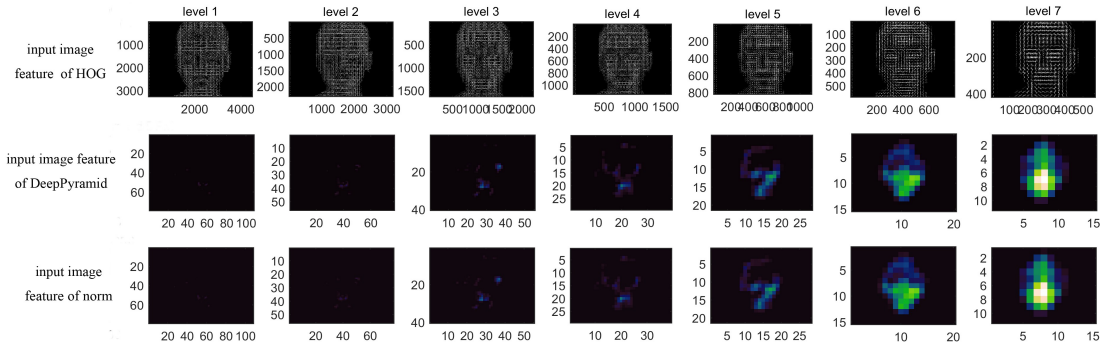

FIGURE 2. Comparison between HOG (in raw 1), conv5 (in raw 2) and norm (in raw 3) feature pyramids. In contrast to conv5 features which are scale sensitive, norm features have almost uniform activation intensities in some levels.

Based on the above analysis, we inherited and modified the top 7 layers of a vtruncated variant of VGGNET [4] to construct a feature extraction network and output a 28×28×256-dimensional feature map. Mainly based on two considerations, (1) Deeper networks and smaller filter sizes have the effect of implicit rules, which can make the discriminant function more accurate [14] (2) The stride of the center of the receptive field is 2×2×2=8, and the smaller target can be focused on. Here, we do not use the image pyramid as the input of the feature extraction network and reduce the computational complexity by adding a normalized layer and multiple DPM-CNNs to reduce the sensitivity of the algorithm to scale changes.

We inherit the method of DP2MFD [5] to add a normalized layer behind the conv7 layer. For a feature point f(x,y) at channel i and position (x,y), the normalized value f ' (i,x,y) is

$$f'(i,x,y) = \frac{f(i,x,y) - \mu_i}{\sigma_i} \tag{1}$$

where μi, σi are the mean and standard deviation of the i-channel feature map, and the norm feature map is used as the input of DPM-CNN for testing and training.

Multi-DPM-CNN

We constructed our three face models based on DPM [4]. The first model has four parts, including 2 eyes, 1 nose, and 1 mouth. The second model has 5 parts, including 1 forehead, 2 eyes, 1 nose, and 1 mouth. The third model has 7 parts, including 1 forehead, 2 eyes, 1 nose, 2 cheeks, and 1 mouth. As shown in Fig. 3.

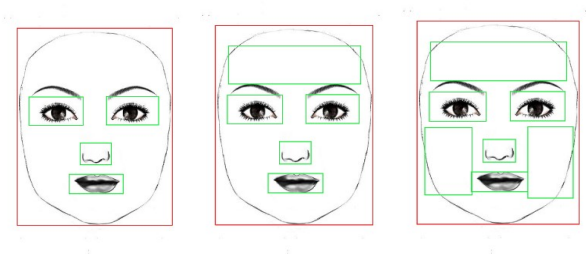

FIGURE 3. Comparison between 4, 5, and 7 parts models (from left to right)

In the DPM model [4], an object model containing n (in this paper, n=7) parts can be defined as n+2 elements (F_0, P_1…, P_n, b). F_0 is the root filter and P_i is i-th part filter. In the model, b is a real value representing the deviation, so at the position (x0, y0), the score (x0, y0) is

$$\text{score}(x_0, y_0) = O_0(x_0, y_0) + \sum_{i=1}^{n} D_i((x_0, y_0) + \upsilon_i) + b \tag{2}$$

where O_0(x0, y0) represents the root filter score, and O(.) is obtained by convolving the feature with the filter. $D_i((x_0, y_0) + \upsilon_i)$ represents the score of all the part filter, and vi indicates the offset of t Pi relative to the upper left corner of F_0. So, the score $D_i(x, y)$ of i-th part filter Pi is

$$D_i(x, y) = \max_{dx, dy}(O_i(x + dx, y + dy) - d_i(dx, dy)) \tag{3}$$

where d_i (x, y) is a convex quadratic function that represents the cost of deformation. The d_i (x, y) can expressed as

$$d_i(x, y) = a_i(x^2 + y^2) + b_i(x^2 + y^2)^{0.5} \tag{4}$$

where a_i>0, bi are all parameters that can be learned.

In DeepPyramid DPM [13], the cost of deformation of each part is obtained by introducing Distance transform pooling layer, which is essentially a max pooling. The biggest difference from max pooling is that Pi transforms the distance di (x, y) is performed on the entire convolution map, so no fixed pool window is specified in advance and the pool area shape can be learned from the data. In the construction of DPM-CNN, the feature map is convolved with the root filter and the part filter respectively, and after obtaining the respective, the convolution map of the partial filter passes through the DT-pooling layer, and then the root convolution map and all DT map are stacked to obtain 8 channels of stacked maps. Finally, the target geometry filter is used to obtain the score map. The role of the geometric filter is to define the offset vi of the part filter relative to the top-left corner of the root filter.

In this way, a DPM-CNN is formed. Unlike the Deep Pyramid DPM, we construct 7 (level=1, 2..., 7) DPM-CNN parallel processing feature maps. The difference between these DPM-CNNs is the filter size. The difference is that this is to reduce the sensitivity of the system to scale changes and is another method to solve multi-scale problems—fixed pictures correspond to different sizes of filters. The advantage of constructing a feature pyramid is that it reduces the computation. We can use the number of convolutions to make a rough estimate. The depth of the convolutional layer passed by DeepPyramid DPM is 7×5+7×(8+1+7) =147. The cost of our method is 7+1+7×(8+1+7) =120.

Non-Maximal Suppression

If the score(j,x,y) at score map level=j and position(x,y) is greater than a certain threshold, the point is mapped back to the region of the picture, the upper left corner of the region is (8x, 8y), The coordinates in the lower right corner are (8x×filter size(j), 8y×filter size(j)). In the end, the non-maximum suppression is performed. When the Intersection-Over-Union (IOU) overlap between the areas is greater than 0.5, the lower scored areas in the overlapping areas are removed. The specific formula of IOU overlaps between region R1 and R2 is as follows:

$$\frac{S(R_1 \cap R_2)}{S(R_1 \cup R_2)} \geq 0.5 \tag{5}$$

Where S(.) indicates area of a region.

EXPERIMENTAL RESULTS

We use the Pascal VOC dataset for training and testing. The PASCAL database has a large number of faces with complex sense and our collected faces as positive samples, and non-face data as negative samples to train face detection.

We studied the influence of the number of different part filters on the face detection effect, and applied a root filter and 4, 5, and 7-part filters to detect the target. Table 1 shows the detection accuracy of our system under different number of part filters.

TABLE 1. Comparison between different numbers of partial filters

The Number of Partial Filters	True Positive Rate at 1000 False Positive Images
4	79.23%
5	80.34%
7	83.11%

After experimental comparison and analysis, it can be found that when using a root filter and 7-part filters for face detection, more effective detection can be achieved, and the robustness, high detection rate. This is because the multi-part filter model can capture more information than less model.

TABLE 2. Comparison between different DPM algorithms

Different DPM Algorithms	True Positive Rate at 1000 False Positive Images
Hog+DPM	66.10%
SuperVision CNN+DeepPyramid DPM	81.28%
VGGNET+DeepPyramid DPM	82.55%
Ours method (7 parts)	83.11%

In addition, we compared the detection effects of different DPM algorithms. It can be seen from the comparison between raw 1 and 2 that the use of high-level depth feature pyramids instead of low-level traditional feature hogs can improve the accuracy of DPM algorithm for face detection. From 66.10% to 81.29%. (VGGNET) DeepPyramid DPM increased by 1.2% compared with (SuperVision CNN) DeepPyramid DPM, verifying that VGGNET is more suitable for extracting features than SuperVision CNN, and that deeper networks and smaller filter sizes have implicit rules. Although our method is only slightly improved compared to DeepPyramid DPM, there is a good improvement in speed. We tested our algorithm using Intel Core I7 CPU, 8G memory computer, operating system Ubuntu16, software caffe+matlab2014a. Our model averagely detects a picture as 220.3s, which has an increase in 33.7 compared to the 254s consumed by the DeepPyramid DPM.

CONCLUSION

In this paper, face detection is implemented based on the deep deformable part model, and the influence of the number of part models and occlusion on the face detection effect is studied. By adding the normalized layer and constructing the filter pyramid to replace the feature pyramid, the invariance of the scale is achieved. Without reducing the detection accuracy, the computational complexity is reduced, and the detection efficiency is improved. At the same time, by using VGGNET, the stride is reduced, and the detection effect on the face of a small size is improved.

Since we did not adopt an end-to-end training approach, the feature extraction network and DPM-CNN were separated and did not make good use of the back-propagation mechanism. Therefore, our future work will focus on how to end-to-end training network.

ACKNOWLEDGMENTS

This work was supported by Project of Science and Technology Program of Guangdong (grant no. 2017B010110015) and Project of Natural Science Foundation of China (grant no. 61502159).

REFERENCES

1. T. Ahonen, A. Hadid, M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," IEEE Transactions on Pattern Analysis & Machine Intelligence, (2006), pp. 2037–2041.
2. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, (2004), pp. 91–110.
3. N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection," Computer Vision and Pattern Recognition–2005, IEEE Computer Society Conference, (2005), pp. 886–893.
4. P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, et al. "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis & Machine Intelligence, (2010), pp. 1627–1645.
5. R. Ranjan, V. M. Patel, R. Chellappa, "A deep pyramid Deformable Part Model for face detection," (2015), pp.1–8.
6. Y. Lecun, Y. Bengio, G. Hinton, "Deep learning," Nature, (2015), pp. 436.
7. Y. Bengio, A. Courville, P.Vincent, "Representation Learning: A Review and New Perspectives," IEEE Transactions on Pattern Analysis & Machine Intelligence, (2013), pp. 1798–828.
8. K. He, G. Gkioxari, P. Dollár, et al. "Mask R–CNN," Eprint Arxiv, (2017).
9. G. Ian, B. Yoshua, C. Aron, "Deeplearning," in Mas- sachusetts (MIT Press, 2016).
10. K. Zhang, Z. Zhang, H. Wang, et al, "Detecting Faces Using Inside Cascaded Contextual CNN," IEEE International Conference on Computer Vision, (2017), pp. 190–3198.
11. H. Li, Z. Lin, X. Shen, et al, "A convolutional neural network cascade for face detection," IEEE Computer Vision and Pattern Recognition, (2015), pp. 5325–5334.
12. B. Yang, J. Yan, Z. Lei, et al, "Convolutional Channel Features," IEEE International Conference on Computer Vision, (2015), pp. 82–90.
13. R. Girshick, F. Iandola, T. Darrell, et al, "Deformable part models are convolutional neural networks," IEEE Computer Vision and Pattern Recognition, (2015) pp. 437–446.
14. K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, (2014).
15. A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," International Conference on Neural Information Processing Systems, (2012), pp. 1097–1105.
16. P. Viola, M. Jones, "Robust Real-time Face Detection," International Journal of Computer Vision, (2004), pp. 137–154.
17. R. Lienhart, "An extended set of Haar–like feature for rapid object detection," Proc of Icip, (2002), pp. 900–903.
18. D. Ramanan, H. Pirsiavash, "Steerable part models," IEEE Conference on Computer Vision and Pattern Recognition, (2012), pp. 3226–3233.
19. W. Ouyang, H. Zhou, H. Li, et al, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," IEEE Transactions on Pattern Analysis & Machine Intelligence, (2017), pp. 99.
20. R. Girshick, J. Donahue, T. Darrell, et al, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, (2014), pp 580–587.
21. C. Zhang, Z. Zhang, "Improving multi–view face detection with multi-task deep convolutional neural networks," Applications of Computer Vision, (2014), pp. 1036–1041.
22. D. L. Nguyen, V. T. Nguyen, M. T. Tran, et al, "Deep Convolutional Neural Network in Deformable Part Models for Face Detection," Pacific-Rim Symposium on Image and Video Technology. (Springer International Publishing, 2015), pp. 669–681.
23. P. Sermanet, D. Eigen, X. Zhang, et al, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," Eprint Arxiv, (2013).