FCFD: TEACH THE MACHINE TO ACCOMPLISH FACE DETECTION STEP BY STEP

Shilun Lin[1], *Fei Su*[1,2]

[1]School of Information and Communication Engineering
[2]Beijing Key Laboratory of Network System and Network Culture
Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

In this paper, a novel progressive strategy is proposed to teach the machine to accomplish face detection in the wild. Firstly, deep model named Fully-connected Face Classifier (FCFC) is built up. With the targeted training data, FCFC learns the knowledge corresponding to distinguish face in various pose, facial expression, occlusion proportion , and blur degree from background gradually. Secondly, FCFC is converted to Fully Convolutional Face Detector (FCFD) which is able to handle a face image in arbitrary size. Finally, FCFDs with different receptive fields and sliding strides are combined to detect faces in various sizes for a given image scale. Experiments on FDDB face detection benchmark dataset and social network selfies show that the proposed FCFD achieves promising results, especially in the detection of blurred or occluded face. Generalization of our FCFD to social network selfies is instructive to analyze and retrieve those photos, since more discriminative information of one person (*e.g.*, gender, identity) can be extracted once his or her face is located.

Index Terms— Teaching step by step, Deep neuron networks, Multi-model fusion, Blurred or occluded face detection

1. INTRODUCTION

Face detection is one of the fundamental technologies to all facial analysis algorithms, including face alignment [1], face recognition [2] and face attribute prediction [3]. The goal of face detection is to get the location and extent of each face in an arbitrary image [4]. Although effortless for human, this task is very difficult for machines. There are many challenges in face detection like variations in pose, scale, expression, occlusion proportion and blur degree. Seminal work by Viola and Jones [5] has made significant progress in face detection. Their detector could detect near frontal faces rapidly and are widely used in portable electronic devices, such as digital camera and mobile phone. Nevertheless it often misses faces with rotation and occlusion which could degrade its performance in practical applications obviously. Hence, in the past decade, extensive research efforts have been devoted to address this problem. Wu *et al.* [6] extended Viola and Jones cascade-based method through training a detector cascade for each view of the face and combining their results in test. Mathias *et al.* [7] achieved the state-of-the-art performance with just a vanilla deformable part models (DPM) [8] based detector which was robust to partial occlusion. For the purpose of further improving the performance of face detection in the wild, where a number of intricate factors such as extreme pose, exaggerated expressions, varying degrees of blur and large portion of occlusion need to be taken into account, recent research focuses on the unconstrained face detection with deep neural network [9, 10].

To address above challenges, we propose a knowledge based Fully Convolutional Face Detector (FCFD), which can be taught step by step through merging knowledge that is used for detecting face in different conditions (pose, expressions, blur and occlusion) into this network structure. FCFD models with different receptive fields and sliding strides are fused to detect faces in various sizes precisely. Our proposed FCFD is verified on face detection benchmark dataset FDDB and collective selfies captured from social networks randomly. The experimental results demonstrate that FCFD could accomplish face detection task excellently, even faces are extremely blurred or occluded. Relevant information of one person (*e.g.*, gender, age and identity) can be extracted once his or her face is located in the image, which is implemental to develop practical selfie management system.

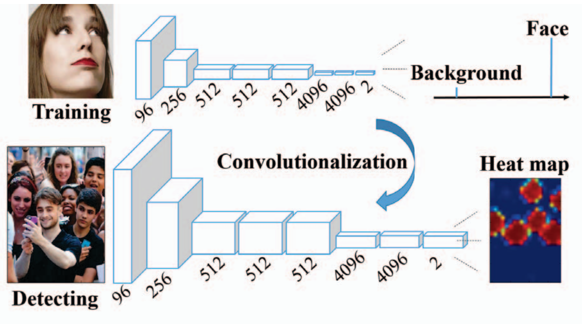

Fig. 1. The pipeline of Fully Convolutional Face Detector.

This work is supported by Chinese National Natural Science Foundation (61372169, 61532018, 61471049) and Special Funds of Beijing Municipal Co-construction Project.

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

2. PROPOSED METHOD

An illustration of our FCFD architecture is shown in Fig.1. In the training phase, a Fully-connected Face Classifier (FCFC) similar to VGG-M [11] is trained to discern face in different conditions. Then the fully-connected layers of the trained FCFC is converted into convolutional layers for the purpose of building FCFD. During the detecting process, we use the heat map produced by FCFD to locate faces in an image of any size.

2.1. Teaching step by step

As mentioned above, unconstrained face detection is a challenging task due to dramatic appearance changes under various pose, scale, expression, occlusion and blur. Thus we give full consideration to those complicated factors in the process of training, and teach the machine to distinguish human faces in different conditions step by step. In this section, details in the training process of FCFC, which is the foundation of FCFD, is provided. The architecture, which bears a resemblance to VGG-M, is selected to construct FCFC. VGG-M architecture is semblable to the one Zeiler and Fergus used [12]. It is characterised by the decreased stride and smaller receptive field in Conv1 layer, which is beneficial to extracting facial features. Using larger stride (2 instead of 1) in Conv2 layer is helpful to maintain the computation time reasonable.

Donahue *et al.* [13] empirically validated that a generic visual feature based on a convolutional network trained on ImageNet outperformed a host of conventional representations on standard object recognition tasks. Yosinski *et al.* [14] documented that even transferring features from distant tasks could be better than using random features. For our face detection task, model pre-trained by classifying massive general object categories has good generalization capability on handling complex background clutters. Hence, the VGG-M model available in the Caffe [15] is applied as the original FCFC which is pre-trained with 1,000 general object categories from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [16], containing 1.2 million training images and 50 thousands validation images.

Pre-training by classifying massive general object brings some prior knowledge of different backgrounds to FCFC. In the next stage, FCFC is taught to distinguish multi-pose face with various expression from background through fine-tuning. Positive training examples extracted from AFLW dataset [17], which contains multi-angle faces with different facial expression, are resized to 224×224. The base learning rate is set to 0.001, and it is increased by 10 times for a brand new layer. The number of stochastic gradient descent (SGD) iterations is chosen as 60K. In each iteration, as mentioned in [18], we uniformly sample 32 positive examples and 96 background examples to construct a mini-batch of size 128.

After fine-tuning, FCFC has the ability to recognize multi pose face with various facial expression. In practice, however, occluded or blurred faces are exceedingly common. So, in the final stage, our FCFC is taught to accomplish this task by further training the network with more challenging training data. We integrate training examples extracted from WIDER FACE dataset [19] which contains faces with different degrees of occlusion or blur. Further teaching with those targeted data could significantly improve the ability of FCFC in occluded or blurred face detection.

2.2. Convolutionalization

During detecting, the face detector should be able to handle input of arbitrary size. However, FCFC obtained in the last step requires a fixed input size (e.g., 224×224), which limits both the scale of the input image and the detection range. A CNN mainly consists of two parts: convolutional layers, and fully-connected layers that follow. As described in [20], the convolutional layers do not require a fixed input size and could generate feature maps of any size, because the operation is in a sliding-window manner. On the other hand, the fully-connected layers need to have fixed length input according to their definition. Hence, the fixed size constraint only comes from the fully-connected layers, which exist at a deeper stage of the network. To address this issue, the fully-connected layers are converted into convolutional layers by reshaping layer parameters [21], which converts FCFC into FCFD.

Detailed pipeline of convolutionalization is illustrated in Fig.2. Consider FCFC architecture (Fig.2, left, gives an example) that takes a 224×224×3 input image, and then utilizes a series of convolutional layers and pooling layers (zero padding is used in some layers) to reduce the image to an activations volume of size 6×6×512 (after Pool5 layer). From there, FCFC uses two fully-connected layers in size of 4096 and the last fully-connected layers with 2 neurons that compute the class scores (probability of face or background). We convert each of these three fully-connected layers into convolutional layer by following measures: At the beginning, replacing the first fully-connected layer which looks at 6×6×512 volume with a convolutional layer that uses filter size 6×6, giving output volume 1×1×4096; Next, replacing the second fully-connected layer with a convolutional layer that uses filter size 1×1, giving output volume 1×1×4096; Finally, replacing the last fully-connected layer similarly, giving final output volume 1×1×2.

After the conversion, FCFD which can "slide" across many spatial positions in a larger image with a single forward pass is obtained. For instance (Fig.2, right), forwarding an image of size 500×500 through FCFD would give the volume in the size of 15×15×512 after Pool5 layer. Forwarding through the next 3 convolutional layers that were just converted from fully-connected layers would give the final volume of size 10×10×2, since (15 - 6)/1 + 1 = 10. Note that instead of a single vector of class scores of size 1×1×2, two 10×10 array of class scores across the 500×500 image is obtained. The array of face is employed as the heat map.

Each point in the heat map contains the probability of having a face which could be used to locate face in its corresponding 224×224 region in the input image. Then, those detected regions are processed by non-maximal suppression (NMS) to get the accurate face locations. Although evaluating FCFC independently across 224×224 crops of the 500×500 image in strides of 32 pixels gives an identical result to forwarding the converted FCFD one time. While the latter is much more efficient, since the 100 evaluations of the former share computation.

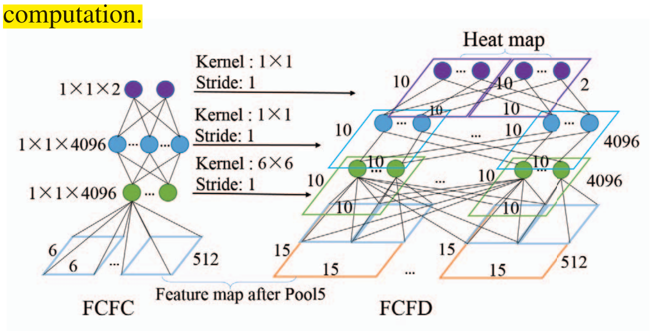

Fig. 2. Detailed pipeline of convolutionalization.

2.3. Multi-model Fusion

FCFD can detect face in an arbitrary image according to the heat map which is obtained in a equivalent sliding-window manner. Faces in a given image often have a variety of sizes, and the distances between those faces are also different. Sliding-window size of 224×224 and stride of 32 pixels, as discussed above, is too large for faces in small sizes and distances. Fusion FCFD is proposed to address this problem, which combines model with different receptive fields and sliding strides. In addition to the 224×224/32s FCFD, we train another FCFC based on the pre-trained VGG-M model with 100×100 training examples. And Pool5 layer is discarded for the purpose of reducing the stride of final FCFD. Because of the change in input size and down-sampling factor, we throw all fully-connected layers of the pre-trained VGG-M model and retrain three fully-connected layers with 100×100 training data. Eventually, the 100×100/16s FCFD converted from that FCFC works together with the 224×224/32s FCFD to detect faces in different sizes and distances.

3. EXPERIMENTS

The proposed FCFD is evaluated on the face detection benchmark dataset FDDB and collective selfies captured from social networks to verify its performance. We demonstrate that teaching step by step and multi-model fusion play an important role in improving the performance of face detection, especially under the wild conditions. Our experiments are based on Caffe [15], a popular deep learning framework.

3.1. Dataset

Training examples extracted from AFLW dataset, which consists of 21,000 images with 24,000 face annotations, is employed in the first stage of training. As mentioned in [9], to increase the number of positive examples, we utilize the randomly sampled patches with ⩾ 0.5 IoU overlap with a ground-truth box and their mirror version as positive examples. As shown in Fig.3 (top), positive examples extracted from AFLW dataset contain a variety of face poses and facial expression, while most of the faces in that dataset are not occluded or blurred. So it is tough for FCFC to identify occluded or blurred faces, which will reduce the performance of the final FCFD. In response to this issue, in the second stage, we introduce examples extracted from WIDER FACE dataset which consists of 393,703 labeled face bounding boxes in 32,203 images and has a high degree of variability in scale, pose, occlusion and blur. Fig.3 (bottom) gives some instances. Obviously positive examples sampled from WIDER FACE contain a variety of occluded and blurred faces. Integrating such training data and further training the network could improve the ability of FCFC to identify various types of human faces, and thus adapt our proposed FCFD to face detection in the wild.

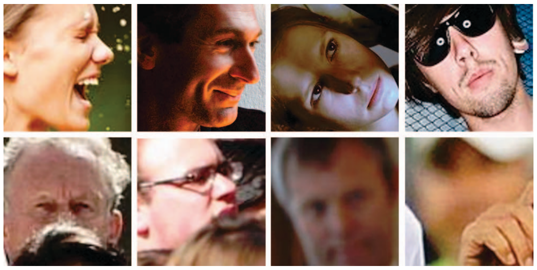

Fig. 3. Positive training examples in AFLW dataset (top) and WIDER FACE dataset (bottom).

3.2. Multi-scale detection

Due to the combination of 100×100/16s and 224×224/32s models, the fusion FCFD is sensitive to faces in size of 100×100 and 224×224. As shown in Fig.4 (a) and (b), utilizing different sizes of sliding windows corresponding to different strides ensures that face in suitable size have a strong response in heat map and the response of relatively close faces could be well separated. However, faces in nonsensitive sizes may be missed (such as faces at the bottom). To detect faces of various sizes, we apply the multi-scale detection which scaled the test image up and down to resize the faces into sensitive sizes approximately. We can see in Fig.4 (c) and (d), along with the change in test image size (down-sampling in this example), response of faces in sensitive sizes (at the bottom) is enhanced while the others almost disappeared.

3.3. Performance of FCFD

We evaluate FCFD on face detection benchmark dataset FDDB and get an average precision of 85.1% (qualitative results are depicted in Fig.5). DDFD [9] achieves a relatively close result of 84.0%, however, the network does not have explicit

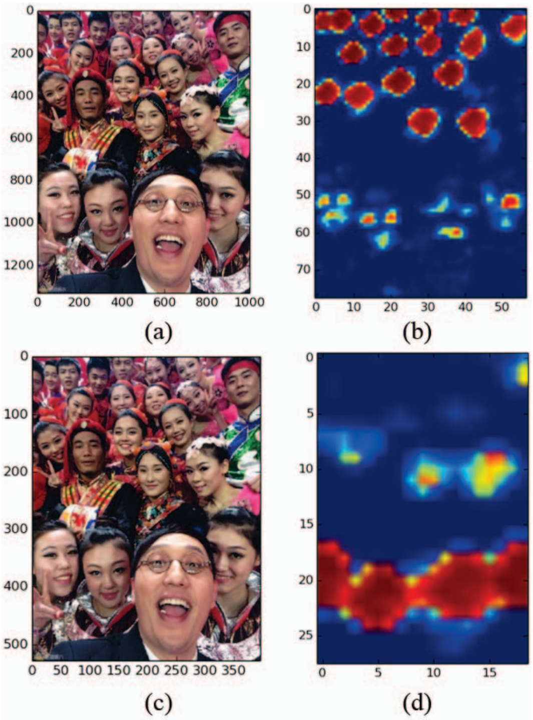

Fig. 4. Performance of FCFD with multi-model fusion and multi-scale detection. Best viewed in color.

mechanism to handle occlusion or blur. DPM-based method [22] achieves 86.4% average precision which uses extra information of face pose annotation during the training and does not pay much attention to serious occluded or blurred faces. Faceness [10] has its first stage designated to handle partial occlusions and achieves a high average precision of 90.99% with a complex pipeline which consists of 5 attribute-aware CNNs. And labeling training data with specific part-level binary attributes (*e.g.*, each training image for hair CNN with hair attributes including black hair, blond hair, brown hair, gray hair, bald, wavy hair, straight hair, receding hairline, bangs) is a time consuming and labor intensive process. In order to verify the generalization of FCFD, we test it on the collective selfies captured from social networks randomly as well (Fig.6 gives some qualitative results). From the detection results we found that our method can detect faces in the case of even more than half of the face region occluded. What's more, faces in different degree of blur could be detected as well. Experimental results demonstrate that teaching gradually helps FCFC to learn specific knowledge for distinguishing faces in varied conditions. So FCFD is able to detect faces under pose variations, occlusions and blurs.

4. CONCLUSION

In this paper, a framework based on deep fully convolutional network, namely FCFD, is proposed to accomplish face detection in complicated conditions. Challenges associated with face detection, such as variations in pose, facial expression, occlusion proportion and blur degree, are carefully considered. And then, the model learns the knowledge corresponding to these challenges from the thoughtfully prepared training data step by step. This novel teaching strategy ensures that our FCFD has the ability to complete the detection of specific types of faces. Moreover, combining multi FCFD is proved to be conducive to detect face of different sizes meticulously in a given scale. Experiments on FDDB dataset and social network selfies show that FCFD achieves promising results, especially in the detection of blurred or occluded faces.

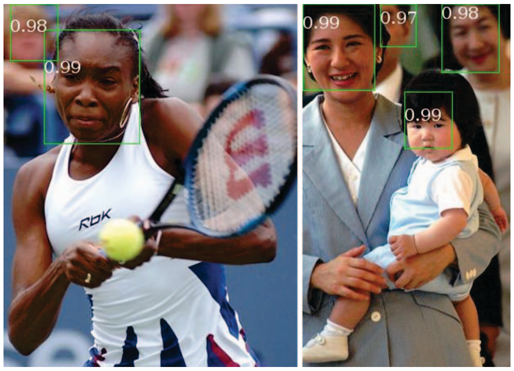

Fig. 5. Qualitative results of FCFD on FDDB.

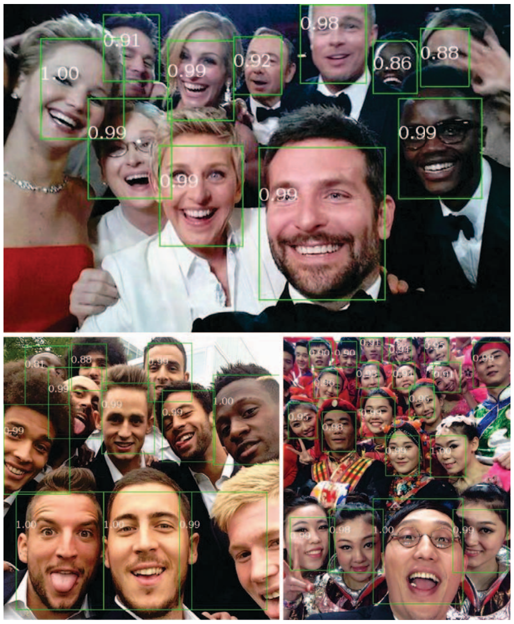

Fig. 6. Qualitative results of FCFD on collective selfies captured from social networks randomly.

5. REFERENCES

[1] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 918–930, 2016.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3730–3738.

[4] Ming-Hsuan Yang, David J Kriegman, and Narendra Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[5] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] Bo Wu, Haizhou Ai, Chang Huang, and Shihong Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 79–84.

[7] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *Computer Vision–ECCV 2014*, pp. 720–735. Springer, 2014.

[8] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*. IEEE, 2008, pp. 1–8.

[9] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 643–650.

[10] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3676–3684.

[11] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Computer Science*, 2014.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[13] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *University of California Berkeley Brigham Young University*, pp. 647–655, 2013.

[14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 248–255.

[17] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Region-based convolutional networks for accurate object detection and segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 1, pp. 142–158, 2016.

[19] Shuo Yang, Ping Luo, Change Loy Chen, and Xiaoou Tang, "Wider face: A face detection benchmark," *Computer Science*, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision–ECCV 2014*, pp. 346–361. Springer, 2014.

[21] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *Eprint Arxiv*, 2014.

[22] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.