Multi-Prototype Networks for Unconstrained Set-based Face Recognition Supplementary Materials

Anonymous IJCAI'19 Submission
Paper ID 110

Abstract

In this supplementary material, we present fully detailed information on 1) multi-scale pyramid construction and random cropping strategy for MPNet data layer; 2) learning algorithm of the MPNet; 3) implementation details; 4) details on IJB-A [Klare *et al.*, 2015] and evaluation metrics; 5) details on YTF [Wolf *et al.*, 2011] and evaluation metrics; 6) details on IJB-C [Maze *et al.*, 2018] and evaluation metrics; 7) high-resolution visualized verification results and discussions for IJB-A split1.

Appendix A

This appendix provides the additional information on the multi-scale pyramid construction and random cropping strategy for MPNet data layer (Sec. 3.1).

MPNet learns face representations at multi-scale for gaining strengthened robustness to scale variance in real-world faces. Specifically, for each medium within a face media set, a multi-scale pyramid is constructed by resizing the image or video frame to four different scales. To handle the error of face detection, MPNet performs random cropping to collect local and global patches from each scale of the multi-scale pyramid with a fixed size, as illustrated in Fig. 1. Some implementation details are as follows. For each medium with the provided face bounding box, we first crop the facial RoI accordingly and then resize it to multiple $r \times r \times 3$ sizes to build the multi-scale pyramids, where r=$224, 256, 384$ and 512. The size of inputs to MPNet is fixed as $224 \times 224 \times 3$ by randomly cropping local and global patches of compatible size from images/video frames. No 2D or 3D face alignment is used.

Appendix B

This appendix provides the additional information on the learning algorithm of the MPNet (Sec. 3.3).

We summarize the details of multi-prototype learning of the MPNet with joint supervision of the ranking loss and the auxiliary DSG loss in Algorithm 1. Clearly, the MPNet is end-to-end trainable and can be optimized with **B**ack**P**ropogation (BP) and **S**tochastic **G**radient **D**escent (SGD) algorithm.

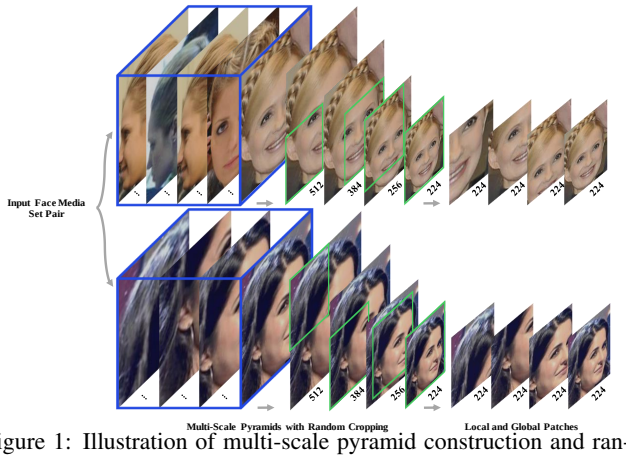

Figure 1: Illustration of multi-scale pyramid construction and random cropping strategy for the MPNet. For each medium within a face media set, a multi-scale pyramid is constructed by resizing the medium to four different scales. Local and global patches are then randomly cropped from each multi-scale pyramid with a fixed size. Best viewed in color.

Appendix C

This appendix provides the additional information on the implementation details (Sec. 4).

We initialize the CNN module of MPNet for deep set-based facial representation learning with the VGGface model [Parkhi *et al.*, 2015], and fine-tune it on the target dataset. For each medium with the provided face bounding box, we first crop the facial RoI accordingly and then resize it to multiple $r \times r \times 3$ sizes to build the multi-scale pyramids, where r=$224, 256, 384$ and 512. The size of inputs to MPNet is fixed as $224 \times 224 \times 3$ by randomly cropping local and global patches of compatible size from images/video frames. No 2D or 3D face alignment is used. The threshold R for balancing input data distribution is set as 128 for trading-off recognition accuracy and computation cost. The weights of the 1^{st} layer (implemented with a 1D convolution layer with sigmoid activation function) of the DSG sub-net are initialized by normal distribution with an std 0.001. The number of total prototypes K is set as 500. We also conduct experiments to illustrate how the K influences the overall performance in Sec. 4.2. The bandwidth parameter β in Eq. (3) is set to 10, the margin τ of the ranking loss is fixed as 0.8, and the

Algorithm 1 Multi-prototype learning algorithm

Input: Training data $X_p = \{(X^{p1}, X^{p2}, y^p)\}$. Initialized parameters θ, W in the CNN module and DSG sub-net, respectively. Hyperparameters $R, K, \beta, \tau, \lambda$ and learning rate μ^t. The number of iteration $t \leftarrow 0$;
Output: The parameters θ and W.
while not converge **do**
$t \leftarrow t + 1$;
Compute the joint loss by $\mathcal{L}^t = \mathcal{L}^t_{Ranking} + \lambda \mathcal{L}^t_{DSG}$;
Compute the backpropagation error for each p by $\frac{\partial \mathcal{L}^t}{\partial X^t_p} = \frac{\partial \mathcal{L}^t_{Ranking}}{\partial X^t_p} + \lambda \cdot \frac{\partial \mathcal{L}^t_{DSG}}{\partial X^t_p}$;
Update the parameters θ by $\theta^{t+1} = \theta^t - \mu^t \Sigma^m_p \frac{\partial \mathcal{L}^t}{\partial X^t_p} \cdot \frac{\partial X^t_p}{\partial \theta^t}$;
Update the parameters W by $W^{t+1} = W^t - \mu^t \Sigma^m_p \frac{\partial \mathcal{L}^t}{\partial X^t_p} \cdot \frac{\partial X^t_p}{\partial W^t}$;
end while

trade-off parameter λ is set as 0.01 by 5-fold cross-validation. Different values of λ lead to different deep feature distributions. With proper λ, the discriminative power of deep features can be significantly enhanced. $\lambda = 0.01$ is large enough for balancing the scales of two loss terms as the sub-graph loss calculates summations over more pairs. The proposed network is implemented based on the publicly available Caffe platform [Jia *et al.*, 2014], which is trained on three NVIDIA GeForce GTX TITAN X GPUs with 12G memory. During training, the learning rate is initialized to 0.01, and during fine-tuning, the learning rate is initialized to 0.001. We train our model using SGD with a batch size of 1 face media set pair, momentum of 0.9, and weight decay of 0.0005.

Appendix D

This appendix provides the additional information on the details on IJB-A and evaluation metrics (Sec. 4).

IJB-A contains 5,397 images and 2,042 videos from 500 subjects, captured from in-the-wild environment to avoid near frontal bias. For training and testing, 10 random splits are provided by each protocol, respectively. It contains two tasks, face verification and identification. The performance is evaluated by TAR@FAR, FNIR@FPIR and Rank metrics, respectively.

Appendix E

This appendix provides the additional information on the details on YTF and evaluation metrics (Sec. 4).

YTF contains 3,425 videos of 1,595 different subjects. The average length of a video clip is 181.3 frames. All the video sequences were downloaded from YouTube. We follow the unrestricted with labeled outside data protocol and report the result on 5,000 video pairs.

Appendix F

This appendix provides the additional information on the details on IJB-C and evaluation metrics (Sec. 4).

IJB-C contains 31,334 images and 11,779 videos from 3,531 subjects, which are split into 117,542 frames, 8.87 images and 3.34 videos per subject, captured from in-the-wild environments to avoid the near frontal bias. For fair comparison, we follow the template-based setting and evaluate models on the standard 1:1 verification protocol in terms of TAR@FAR.

Appendix G

This appendix provides the additional information on the high-resolution visualized verification results and discussions for IJB-A split1 (Sec. 4.1).

Finally, we visualize the verification results in Fig. 2 for IJB-A split1 to gain insight into unconstrained set-based face recognition. After computing the similarities for all pairs of probe and reference sets, we sort the resulting list. Each row represents a probe and reference set pair. The original sets within IJB-A contain from one to dozens of media. Up to 8 individual media are shown with the last space showing a mosaic of the remaining media in the set. Between the sets are the set IDs for probe and reference as well as the best matched and best non-matched similarities. Fig. 2 (blue, left) shows the best matched cases. In the top-30 scoring correct matches, we immediately note that every reference set contains dozens of media. The probe sets either contain dozens of media or one medium that matches well. Fig. 2 (blue, right) shows the worst matched cases, representing failed matching. The thirty lowest matched results from single-medium probe sets are all under extremely challenging unconstrained conditions. These extremely difficult cases cannot be solved even using the specific operations designed in MPNet. Fig. 2 (green, left) showing the worst non-matched cases highlights the understandable errors involving single-medium probe sets representing impostors in challenging orientations. Fig. 2 (green, right) showing the best non-matched cases shows the most certain non-mates, again often involving large sets with enough guidance from the relevant information of the same subject.

References

[Jia *et al.*, 2014] 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 *ACM MM*, pages 675–678, 2014.

[Klare *et al.*, 2015] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain. Pushing the frontiers

Figure 2: Verification results analysis for IJB-A split1. (blue, left) The best matched cases, (blue, right) The worst matched cases, (green, left) The worst non-matched cases (green, right) The best non-matched cases. For better viewing of this figure, please see the original zoomed-in color pdf file.

of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *CVPR*, pages 1931–1939, 2015.

[Maze *et al.*, 2018] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. IARPA Janus Benchmark-C: face dataset and protocol. In *ICB*, 2018.

[Parkhi *et al.*, 2015] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[Wolf *et al.*, 2011] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.