Learning Discriminative Aggregation Network for Video Face Recognition Supplementary Material

Anonymous ICCV submission

Paper ID 1600

Table 1. The 5 landmarks coordinates of face template.

	x	y
left eye	30.3	51.5
right eye	65.5	51.5
nose	48.0	71.7
left mouth corner	33.5	92.2
right mouth corner	62.7	92.2

1. Face Alignment

We used the MTCNN [6] to detect 5 points landmarks (two eyes, nose and mouth corners) and aligned faces by similarity transformation from detected landmarks to face template. The original images and aligned faces are shown in Figure 1. All faces were cropped and resized to 96×112. The 5 landmarks coordinates of face template are presented in Table 1, which is decided according to the mean of selected front faces.

2. Feature Extraction Network

We used the feature extraction network [1] provided by authors of [5]. Architecture of the network is illustrated in Figure 2, where the kernel size and stride of convolutional layers are 3×3 and 1 respectively, and the kernel size and stride of max pooling layers are 2×2 and 2 respectively. The network use PReLU [2] as activations. We present the detailed information of the network in Table 2. For face verification results, we followed the standard protocol of the LFW [3] and the YTF [3] dataset, which is the same as mentioned in experiment section. Note that we computed the cosine similarity by using feature vectors of frames or images directly, and we did not use the horizontal flip, cropping or PCA tricks for all experiments in experiment section and supplementary material.

[1] https://github.com/ydwen/caffe-face

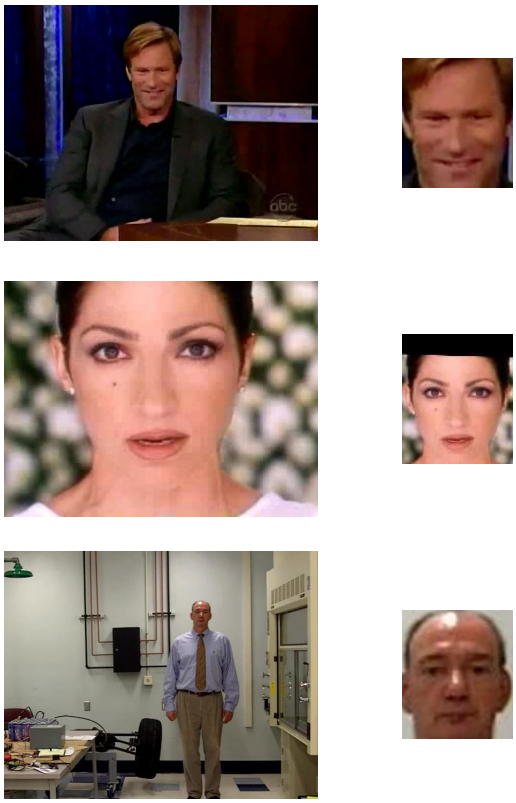

Figure 1. Face alignment examples from the Youtube face dataset (YTF) [3], the Youtube Celebrities dataset (YTC) [4] and Point-and- Shoot Challenge (PaSC) [1].

Table 2. Detailed information of feature extraction network. We present the number of parameters and face verification accuracy (%) on widely used LFW and YTF dataset.

# Parameters	2.75×10^7
LFW accuracy	97.96
YTF accuracy	93.16

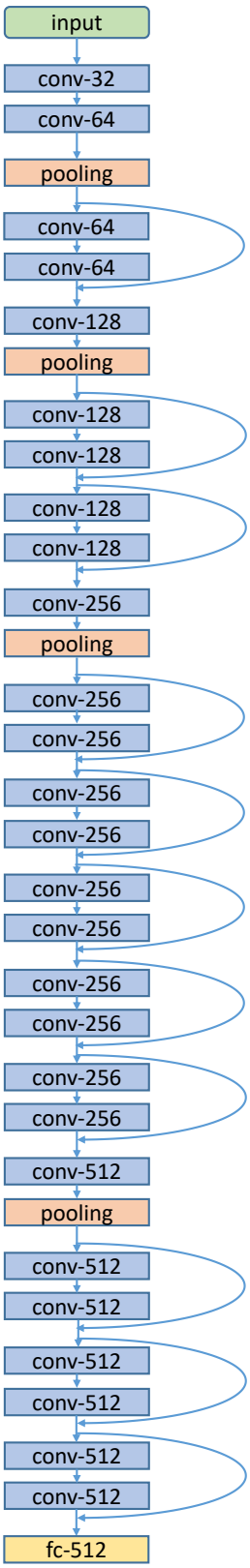

Figure 2. Network architecture. The numbers are either the feature map channel for convolutional layers or feature dimension for fully connected layers.

3. Visual Results

More examples of DAN are presented in Figure 3.

References

[1] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, pages 1–8, 2013. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 1

[3] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1

[4] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008. 1

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 1

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016. 1

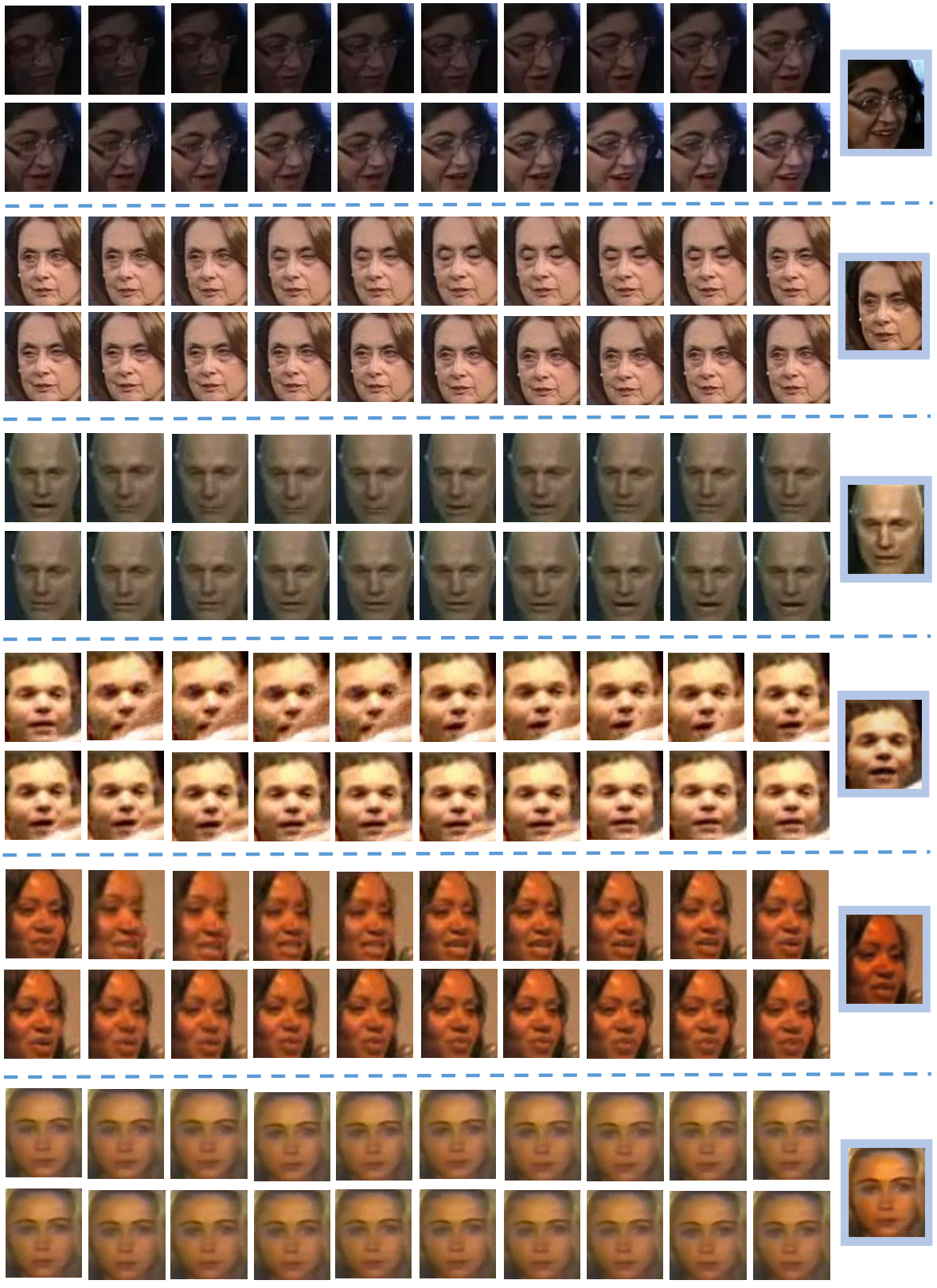

Figure 3. The examples of original video frames (on the left) and the aggregated images (on the right). It can be observed that the synthesized images are visually better than input frames and our proposed DAN can denoise the low-quality frames.