Learning Discriminative Aggregation Network for Video-based Face Recognition

Yongming Rao[1,2,3], Ji Lin[1,2,3], Jiwen Lu[1,2,3]*, Jie Zhou[1,2,3]
[1]Department of Automation, Tsinghua University, Beijing, China
[2]State Key Lab of Intelligent Technologies and Systems, Beijing, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China
raoyongming95@gmail.com; lin-j14@mails.tsinghua.edu.cn; {lujiwen,jzhou}@tsinghua.edu.cn

Abstract

In this paper, we propose a discriminative aggregation network (DAN) method for video face recognition, which aims to integrate information from video frames effectively and efficiently. Unlike existing aggregation methods, our method aggregates raw video frames directly instead of the features obtained by complex processing. By combining the idea of metric learning and adversarial learning, we learn an aggregation network that produces more discriminative synthesized images compared to raw input frames. Our framework reduces the number of frames to be processed and significantly speed up the recognition procedure. Furthermore, low-quality frames containing misleading information are filtered and denoised during the aggregation process, which makes our system more robust and discriminative. Experimental results show that our method can generate discriminative images from video clips and improve the overall recognition performance in both the speed and accuracy on three widely used datasets.

1. Introduction

Video face recognition has been attracting increasing efforts in the past few years [1, 3, 4, 13, 29, 34, 37, 43, 44, 45], which has many practical applications in surveillance, person identification, video search. Compared to still image recognition, video face recognition is more challenging because there are many noisy frames in videos which contain unfavorable poses and viewing angles. Furthermore, as the video usually consists of more than 100 frames, it brings considerable computational burdens for the state-of-the-art recognition methods such as the deep neural networks. Therefore, it is desirable to propose a framework that can denoise the original video by extracting useful information from noisy data and reduce the overall runtime. In other words, a new framework which can aggregate the information from the video keep the same discriminative ability for efficient face recognition is required.

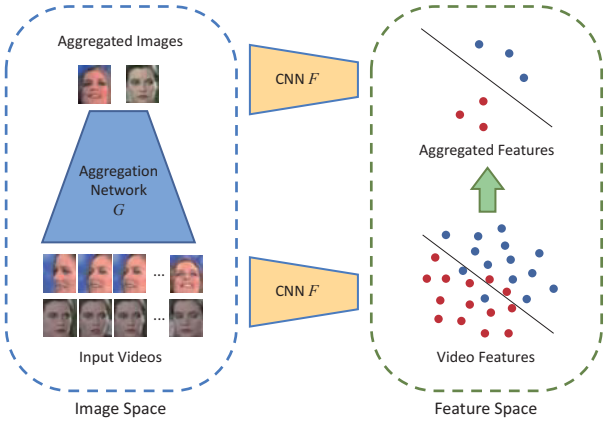

Figure 1. The basic idea of our proposed frames aggregation method. For each video clip, we integrate the information of videos to produce few synthesized images with discriminative aggregation network (DAN). The supervision signal of our proposed framework makes the synthesized images more discriminative than original frames in the feature space. Besides, we only need to pass the few aggregated images into feature extraction network and thus greatly speed up the overall system.

There have been varieties of efforts on integrating information from different frames to represent the whole video [2, 16, 19, 32, 45]. However, most of them focuses on extracting features from raw video frames, which means that feature extraction is performed at first before other operations. This kind of procedure will harm the recognition performance because some frames have low quality will mislead the system into wrong decisions, which cannot be easily distinguished in the feature space because such information is usually lost during the feature extraction stage. Therefore, it is important to conduct aggregation process before feature extraction.

Generative adversarial networks (GAN) have achieved great success in many fields of computer vision [5, 8, 22, 27,

*Corresponding author.

35, 36, 47]. Inspired by the basic idea of adversarial learning, we propose a GAN-like aggregation network which takes the video clip as the input and reconstruct a single image as the output. However, the output image produced by the generative adversarial network is only visually similar to the original data, but does not guarantee any discriminative power. On the other hand, metric learning [9, 13, 37, 43] has been one of the most discriminative techniques in face recognition, which maps samples into a semantic feature space where they can be well distinguished. By combining metric learning with adversarial learning, we are able to train a generative model that can produce photo-realist face images and provides even stronger discriminative ability simultaneously.

In this paper, we propose a discriminative aggregation network (DAN) method for video-based face recognition, where the overall framework is shown in Figure 1. By combining metric learning and adversarial learning, DAN can aggregate the useful information of an input video into one or few more discriminative images in the feature space for face recognition. Since the number of images to be processed is greatly reduced, our framework significantly speeds up video-based face recognition. Unlike existing methods which extract features from raw video frames before other fusion operations, our framework directly fuses the information from one raw video into several images, and can thus distinguish low quality frames and denoise the input video simultaneously. Experimental results on the YouTube Face dataset [44], Point-and-Shoot Challenge [1] and the YouTube Celebrities dataset [24] show that DAN can accelerate the recognition speed and improve the recognition performance simultaneously.

2. Related Work

Video Face Recognition: Existing video face recognition methods [15, 17, 18, 30, 31, 32, 34, 37, 39, 40, 42, 43, 45] can be mainly categorized into two classes: image-based and video-based. For the first category, each video is considered as a set of images and the relationship between frames is exploited for recognition. These methods are designed to solve the general face recognition problem, where they are easily applied into video face recognition [6, 34, 37, 39, 40, 43]. We consider this type of methods as the basis of video face recognition, and our model is built upon these methods. For the second category, each video is usually modeled as an image set, and the distance or similarity between videos is computed by the properties of image sets. In previous works, image-set-based models have a variety of forms. Cevikalp *et al*. modeled image sets as affine hulls [2]. Huang *et al*. calculated the distances between image sets using the distances between SPD manifolds [16, 18]. Lu *et al*. represented image sets as a set of n-order statistics [31, 42]. Yang *et al*. proposed an attention-based model to aggregated features of image sets [45]. For methods from both classes, the key challenge is how to represent a video as a single feature. In their works, they first represent frames in videos using handcraft by feature vectors or deep neural networks, and then aggregate these features. In this work, we represent videos in a different way, where we aggregate frames at the beginning and speed up the recognition process.

Deep Metric Learning: Many metric learning algorithms have been proposed in recent years, and many of them have been successfully applied to improve the recognition performance [9, 13, 37, 43]. Deep metric learning [13] aims to produces discriminative features through the combination of deep learning and metric learning. For example, Hu *et al*. employed a fully connected network to achieve parametric metric learning. Schroff *et al*. presented a triplet loss function for feature embedding. Wen *et al*. proposed the center loss function to improve the faces distribution in feature space. Different from these feature embedding methods, we propose a new *image embedding* method to guide the aggregation network to synthesize discriminative images.

Adversarial Learning: Goodfellow *et al*. proposed the idea of generative adversarial networks (GAN) [8]. After that, adversarial learning has attracted great attention in recent years [5, 22, 27, 35, 36, 47]. Compared to conventional generative models, GAN has shown promising performance for generating more sharper images, and the ability of photo-realistic image synthesis has been applied in many areas. Larsen *et al*. combined a variational autoencoder (VAE) [26] with a GAN to take the advantages from both models and learned a high-level abstract visual features embedding [27]. Zhang *et al*. developed the idea of deep convolutional GAN [35] and text-to-image synthesis [36] and achieved impressive results on image synthesis [47], Isola *et al*. studied on a variety of image-to-image translation applications in computer vision by combining the traditional n-norm distance loss and adversarial loss [22]. Ledig *et al*. employed a GAN-like network together with a loss function defined by high-level features to improve the perceptual quality of image super-resolution. However, little progress has been made in adversarial learning for recognition task. In our work, we combine the idea of adversarial learning with metric learning to aggregate photo-realistic images for boosting face recognition performance.

3. Approach

3.1. Problem Definition

Video-based face recognition aims to recognize whether a face video belongs to a certain subject. Such videos usually contain more than 100 frames (like videos in the YouTube Face dataset) and brings considerable computation-

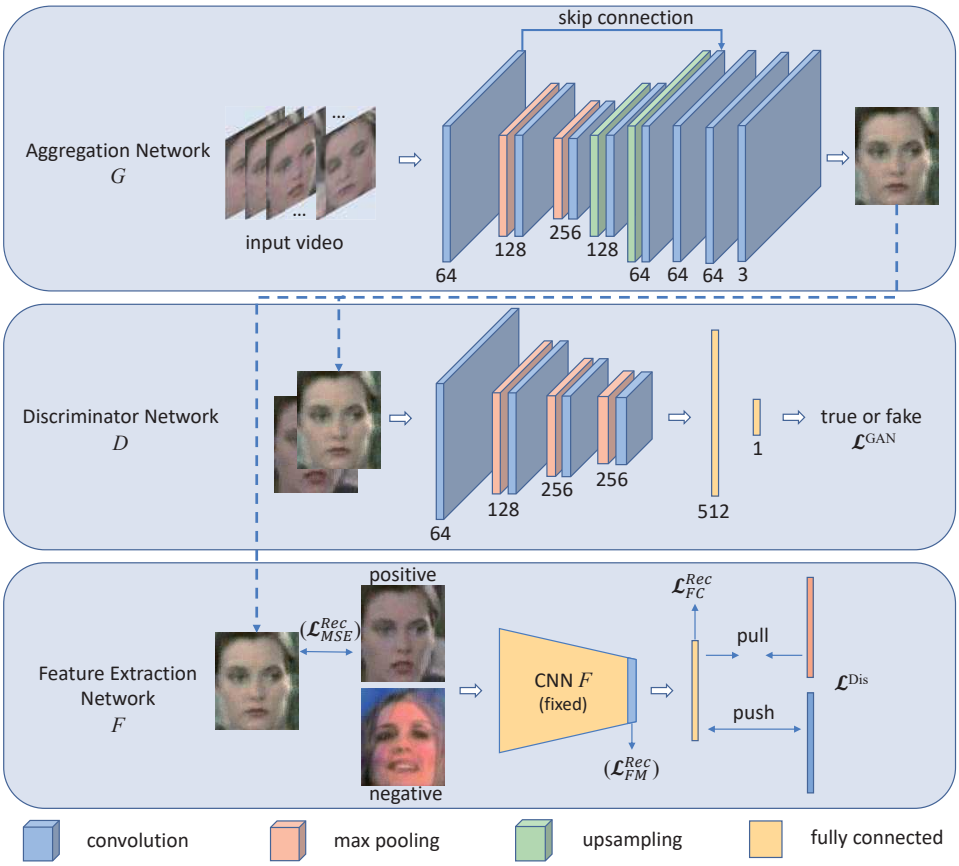

Figure 2. Detailed architecture of our proposed framework. The numbers are either the feature map channel for convolutional blocks or feature dimension for fully connected layers. The output of aggregation network is then fed into discriminative network for adversarial learning, and the feature extraction network to increase discrimination. Different losses are applied at different places as illustrated in the figure.

al burdens for existing methods. The goal of our framework discriminative aggregation network (DAN) is to aggregate a long video into one or few frames while still remains or increases the discriminative power, which can be used for efficient face recognition.

We denote our goal as the following objectives:

$$\begin{aligned} & V^m \rightarrow X^n \\ \text{subject to.} \quad & m > n, \\ & Dis(F(X_p), F(X_n)) > Dis(F(V_p), F(V_n)) \end{aligned} \tag{1}$$

where V^m is the input video with m frames and X^n is the aggregated n images, with m much greater than n. The subscripts p and n refer to positive and negative samples and F is the feature extraction network. We used a function Dis to evaluate the discriminative ability between positive and negative samples. This means with DAN we can greatly reduced the number of images to be processed, while the aggregated images still have more discriminative ability in the feature space of certain CNN F.

3.2. Overall Framework

The overall framework of our proposed discriminative aggregation network (DAN) is presented in Figure 2. DAN consists of 3 sub-networks. We define them as aggregation (generator) network G, discriminator network D and feature extraction network F. We denote the whole video as V. For the ease of implementation, at each time we aggregate a subset S of V into a single image, so the input of G is a subset S and the output is a single discriminative image X. The discriminator D tries to judge whether the image is generated by G or selected from the original video, forming adversarial learning with G. The feature generator network F extracts features from the aggregated images, and tries to make the feature discriminative in the feature space.

The aggregation network G starts with several convolution blocks into smaller feature maps, and then reconstructs the aggregated output image with several deconvolution

blocks. We also add a skip connection between the first and high-level feature map following another GAN based framework [28]. The discriminator network D consists of several convolution blocks and finally produces 1 output denoting whether the image is generated or selected from the original video. For aggregation network, each convolutional block consists of a standard convolution layer and a batch normalization layer [20]. For discriminator network, each convolutional block is a standard convolution layer. The kernel size and stride of convolutional layers is 3×3 and 1×1 respectively. All max pooling layers have kernel size 2×2 and stride 2. For upsampling layer, we use bilinear filter with stride 2 to upscale feature maps. All hidden layers in both G and D use PReLU [11] as activations, and the output layer of G uses `tanh` nonlinearity to produce normalized pixel values. The output layer of D use `sigmoid` nonlinearity to produce the possibility of whether the input image is true or synthesized. G and D are trained iteratively such that they can provide loss signal to each other to reach an optimal balance, where the generated output of G cannot be distinguished from ground-truth images. For the feature extraction network F, we used the network provided by the author of [43], which is a residual convolutional network [12]. The detailed structure of F is provided in the supplementary. We keep F unchanged during the training process.

3.3. Loss Functions

We hope our framework DAN can aggregate video clip into single image while at the same time gain more discriminative power. Special designed loss functions are needed to achieve such goal. We use the following loss function:

$$\mathcal{L} = \lambda \mathcal{L}^{Dis} + \eta \mathcal{L}^{Rec} + 0.01 \mathcal{L}^{GAN} \tag{2}$$

where $\mathcal{L}^{Dis}$ is the discriminative loss, $\mathcal{L}^{Rec}$ is the reconstruction loss and $\mathcal{L}^{GAN}$ is the adversarial loss. The parameter 0.01 is referred from [22].

3.3.1 Discriminative Loss

Samples from face recognition datasets consists of positive video pairs and negative video pairs. We use the term (X, P) for positive pairs and (X, N) for negative pairs, where X is the aggregated image, P and N are positive and negative samples randomly chosen from the other video clips respectively.

To make the generated image discriminative, we propose the discriminative loss as:

$$\mathcal{L}^{Dis} = \begin{cases} (||F(X) - F(P)||^2 - \alpha)_+ & y = 1 \\ (\beta - ||F(X) - F(N)||^2)_+ & y = 0 \end{cases} \tag{3}$$

and

$$\alpha = \min_{A \in S} ||F(A) - F(P)||^2 \tag{4}$$

where y is the label either 1 or 0 denoting positive or negative pairs, F is the feature extraction network. S is the subset clip to be aggregated and A is one of a frame in it. We use Euclidean distance to measure the distance between two feature representations. α is the smallest distances between all frames in S and P. β is a manually set constant margin. The subscript $+$ means $\max(0, \cdot)$.

The basic idea is that if we sample a positive video pair from training data, and take a subset S of one video for aggregation and randomly sample a frame P from the other video, we hope the aggregated image X is closer to P than any other frame from the original video subset S in the feature space of F. Contrarily, if negative sample is considered, we hope the distance between generated X and N is greater than a certain margin. With such loss function, we can guarantee the feature of aggregated image extracted by F is more discriminative than original frames.

3.3.2 Reconstruction Loss

Since we reconstruct a face image from a compressed representation, we need to exert reconstruction loss on the output image. Here we compared 3 forms of reconstruction loss as shown in Figure 2.

Pixel-wise MSE loss is the most widely used objective function for existing frameworks like [7, 38], which is calculated as:

$$\mathcal{L}^{Rec}_{MSE} = \frac{1}{N_I} ||I - X||^2_{\mathcal{F}} \tag{5}$$

where I is the original image and X is the reconstructed one. N_I is the number of total pixels in an image.

Another reconstruction loss proposed in [28] focuses on the feature map difference between reconstructed or original image, as shown in the bottom part in Figure 2. The loss function is listed below:

$$\mathcal{L}^{Rec}_{FM} = \frac{1}{N} \sum_{i=1}^{n} ||\phi_i(I) - \phi_i(X)||^2_{\mathcal{F}} \tag{6}$$

where ϕ maps image to its high-level feature maps, and in our case, we use the convolutional part of feature extraction network F as ϕ. The subscript i denotes the index of channel, with totally n feature maps. N_{FM} is the number of total entries of feature maps.

We cannot naively define the above two forms of reconstruction loss, as there are multiple images in the input S. For implementation, we choose I according to the following rule:

$$I = \begin{cases} \text{argmin}_{A \in S} ||F(A) - F(P)||^2 & y = 1 \\ \text{argmax}_{A \in S} ||F(A) - F(N)||^2 & y = 0 \end{cases} \tag{7}$$

However, the two forms of reconstruction loss both focus on visually similarity, from shallow to upper level. They

can guarantee visual characteristics but not semantic information or discriminative power. DAN focuses on the feature representation extracted from the aggregated image, so it is naturally to apply reconstruction loss to the feature embedding. We propose our new reconstruction loss as below:

$$\mathcal{L}_{FC}^{Rec} = ||F(X) - \text{mean}(F(V^m))||^2 \quad (8)$$

where F is the feature extraction network as mentioned above, and V^m is the original video consisting of m frames. We hope the feature of reconstructed image is close to the mean of features extracted from V per frame to reduce the intra-class distances.

We give detailed analysis on the three forms of reconstruction loss in Section 4.3.

3.3.3 Adversarial Loss

In addition to the reconstruction loss, we also add the adversarial loss to our framework as widely adopted in GAN-based frameworks [5, 22, 27, 35, 36, 47]. This encourages G to generate aggregated outputs that are close to the natural distribution, by forming adversarial learning with G. The loss is defined based on the possibility whether an image comes from the original video, denoted as:

$$\begin{aligned} \mathcal{L}^{GAN} &= \mathbb{E}_{A\sim p_{\text{train}}(A)}[\log D(A)] \\ &+ \mathbb{E}_{V^m\sim p_{\text{train}}(V^m)}[\log(1 - D(G(V^m))] \end{aligned} \quad (9)$$

Here $D(G(V^m))$ is the probability that the aggregated image $G(V^m)$ is a natural image taken from the original video V^m. The goal of D is to maximize L^{GAN} while G tends to minimize it. D and G play the minimax game until reaching a balanced state. D and G are trained iteratively following commonly used settings.

The overall training procedure of our method is summarized in **Algorithm 1**.

Algorithm 1 Minibatch stochastic gradient descent training of our DAN.

Input: Training video pairs, learning rate lr, iterative number I_t, and parameter λ, η.
Output: Aggregation network G

1: Initialize G with MSE pretrained model.
2: Initialize D with pretrained model.
3: Load model of F.
4: **for** $iter < I_t$ **do**
5: **for** k steps **do**

- Sample a video V from the training set, and aggregate a subset S into image $X = G(S)$.
- Sample a frame A from the subset S
- Update the discriminator by ascending its stochastic gradient:

$$\nabla\mathcal{L}^{GAN}$$

6: **end for**

- Sample a video sampling V from the training set, and aggregate a subset S into image $X = G(S)$.
- Calculate the reconstruction target of $\mathcal{L}^{Rec}$ from selected V
- Update the aggregation by descending its stochastic gradient:

$$\nabla\mathcal{L} = \nabla(\mathcal{L}^{Dis} + \mathcal{L}^{Rec} + \mathbb{E}_{\text{batch}}[\log(1 - D(G(S))])$$

7: **end for**
8: **return** Neural network G

4. Experiments

4.1. Datasets and Protocols

We evaluated the proposed discriminative aggregation network (DAN) on three widely used datasets including the YouTube Face dataset (YTF) [44], the Point-and-Shoot Challenge (PaSC) [1] and the YouTube celebrities (YTC) [24]. Here we give a brief description of these three datasets.

YTF: The YouTube Face (YTF) dataset is a widely used video face dataset, which contains 3,425 videos of 1,595 different subjects. In this dataset, there are many challenging videos, including amateur photography, occlusions, problematic lighting, pose and motion blur. The length of face videos in this dataset vary from 48 to 6,070 frames, and the average length of videos is 181.3 frames. In experiments, we follow the standard verification protocol and test our method for unconstrained face $1:1$ verification with the given 5,000 video pairs. These pairs are equally divided into 10 splits, and each split has around 250 intra-personal pairs and around 250 inter-personal pairs.

PaSC: The Point-and-Shoot Challenge (PaSC) dataset contains 2,802 videos of 265 subjects. In this dataset, videos have different distances to the camera, viewpoints, the sensor types and *etc*. The dataset is composed of two parts, in which videos are taken by control and handheld cameras respectively. Compared to the YTF dataset, PaSC is more challenging because faces in this dataset have full pose variations. We followed the standard $1:N$ verification protocol and tested our method on both control and handheld parts of the dataset.

YTC: The YouTube Celebrities (YTC) dataset contains 1,910 videos of 47 subjects and the number of frames varies from 8 to 400. We followed the protocol of standard ten-fold cross validation and selected 3 videos for training and 6 videos for testing randomly for each subject in each fold. We used the dataset to evaluate the performance of our method on the video face classification task.

4.2. Implementation Details

Pre-processing: Following [43], we employed a recent algorithm MTCNN [48] to detect 5 points landmarks for faces in videos. If detection fails, we used the landmarks provided by datasets. We used similarity transformation to align faces according to the landmarks, and then cropped and resized faces to 112×96 to remove the background information. In order to reduce the influence of video lengths, each video was resampled to 200 frame as pre-processing. In practice, our aggregation network can be applied to images of arbitrary size since it is fully convolutional.

Training: We set the input of aggregation network as 20 video frames in implementation, which is a trade-off of efficiency and complexity. When training the adversarial networks, we followed the standard approach [8] and set k as 1. We alternately updated one step for discriminator network and one step for aggregation network. To optimize proposed networks, we employed the mini-batch stochastic gradient descent (SGD) with batch size 16 and applied the Adam [25] optimizer. We set learning rate, β_1 and β_2 as 0.0001, 0.9 and 0.999 respectively. We used the aggregation and discriminator networks pretrained by the supervision signal of MSE loss as initialization before using other reconstruction losses and discriminative loss to avoid local optima. Each network was trained for $10,000$ iterations. We turned the update of batch-normalization parameters off during test time to ensure that the output depends only on the input [20]. The parameter λ and η in Eq. [37] is set as 1.0, and the weight of GAN loss is set as 0.01, as [22].

We trained our DAN on the training set of the YTF dataset. For PaSC and YTC, which have relatively small training set, we fine-tuned the YTF trained model to report the result.

Testing: For all of the three datasets, we firstly used our proposed method to aggregate the whole video into 10 face images. Then we use the face recognition network and mean-pooling to represent each video as a single feature vector. For both $1:1$ and $1:N$ verification task, we used the cosine similarity and threshold comparison, where thresholds were computed from the training set. For classification task, we computed the cosine similarity between examples in training set and testing set and decided the categories according to the nearest neighbor.

Experiment environment: Our implementation is based on Caffe [23] with python wrapper. Our hardware configuration comprises of a 2.40 GHz CPU and a 32.0 GB RAM. We use a Tesla K80 GPU for neural network acceleration.

4.3. Results and Analysis

Results: The results of YTF, PaSC and YTC datasets are given in Tables 1-3 respectively. For all these three datasets, we report the average accuracy obtained by our framework, denoted as **DAN** in the three tables. We also report the result by directly passing all the video frames through feature extraction CNN F with mean pooling for comparison, denoted as **CNN**. To investigate into the influence of frame numbers, we formed two subsets of original video frames by: (1) randomly sampling the same number of frames as generated by DAN from the original video, which is 10 in

Table 1. Comparisons of the average verification accuracy (%) with the state-of-the-art results on the YTF dataset.

Method	Accuracy
LM3L [14]	81.3 ± 1.2
DDML [13]	82.3 ± 1.2
EigenPEP [29]	84.8 ± 1.4
DeepFace-single [40]	91.4 ± 1.1
DeepID2+ [39]	93.2 ± 0.2
FaceNet [37]	95.12 ± 0.39
Deep FR [34]	97.3
NAN [45]	95.72 ± 0.64
Wen *et al*. [43]	94.9
CNN	93.16 ± 0.97
Random + CNN	92.80 ± 1.17
Hierarchical Pooling	93.15 ± 1.12
DAN	94.28 ± 0.69

Table 2. Comparisons of the verification rate (%) with the other state-of-the-art results on the PaSC at a false accept rate(FAR) of 0.01.

Method	Control	Handheld
PittPatt	48.00	38.00
DeepO2P [21]	68.76	60.14
VGGFace	78.82	68.24
SPDNet [16]	80.12	72.83
GrNet [19]	80.52	72.76
CNN	90.78	78.67
Random+CNN	89.12	78.03
Hierarchical Pooling	89.83	78.23
DAN	92.06	80.33

Table 3. Comparisons of the classification accuracy (%) with the other state-of-the-art results on the YTC.

Method	Accuracy
MDA [41]	67.2 ± 4.0
LMKML [32]	70.31 ± 2.52
MMDML [30]	78.5 ± 2.8
GJRNP [46]	81.3 ± 2.0
DRM-WV [10]	88.32 ± 2.14
CNN	96.79 ± 1.27
Random + CNN	96.63 ± 1.31
Hierarchical Pooling	96.78 ± 1.25
DAN	97.32 ± 0.71

Table 5. Investigation of different loss functions and the corresponding accuracy.

Adversarial loss $\mathcal{L}^{GAN}$	Discriminative loss $\mathcal{L}^{Dis}$	Reconstruction loss $\mathcal{L}^{Rec}_{MSE}$	Reconstruction loss $\mathcal{L}^{Rec}_{FM}$	Reconstruction loss $\mathcal{L}^{Rec}_{FC}$	Accuracy (%)
		✓			91.38 ± 0.74
✓					92.50 ± 0.96
✓		✓			92.36 ± 0.90
✓			✓		92.46 ± 0.97
✓				✓	92.92 ± 0.81
✓	✓	✓			93.02 ± 0.88
✓	✓		✓		93.16 ± 0.93
✓	✓			✓	$\mathbf{94.28 \pm 0.69}$

Table 4. Runtime analysis.

Method	Runtime(ms)	Processed frames
CNN	819.7	181
Random + CNN	42.0	10
DAN	126.1	200

our case; (2) mean pooling on similar faces and summarizing the whole video as 10 frames. We measured the performance by mean pooling on corresponding CNN features and denoted results as **Random + CNN** and **Hierarchical Pooling** respectively.

The results show that on all three datasets, DAN outperforms the original CNN for dense feature extraction. This is a strong baseline with high computation complexity, showing that the aggregated images produced by DAN are more discriminative than original video frames. Randomly sampling 10 frames from the original video will lead to significantly performance drop. This, on the contrary, shows the robustness and denoising ability of proposed DAN results.

Compared to previous state-of-the-art methods, DAN outperforms all of them on PaSC and YTC. On YTF, DAN achieves competitive but not the best result. This is largely due to the baseline CNN which is comparatively weaker than those of [37] and [45]. But the gained improvement over baseline CNN result has already proven the effectiveness.

Runtime Analysis: Efficiency is one of the advantages of our framework, and here we give a short analysis of runtime. For dense feature extraction baseline, we calculated that the average frame number of the YTF dataset is 181.3, which is used to measure the runtime. For random CNN protocol, we randomly selected 10 frames from the video and measure the forward time. For our DAN, we measured the overall time including aggregating 200 frames into 10 images with DAN and the forward time of CNN. The results are listed in Table 4.

From the table we can see that our DAN is much faster than the baseline method with dense feature extraction, and has only little overhead compared to the random sampled baseline, while achieves better results than both of them. This shows the effectiveness of proposed DAN framework.

Investigation of Loss Functions: We proposed three loss functions above: discriminative loss, reconstruction loss and adversarial loss. Here we analyze the effects of each loss functions with detailed experiments on YTF. The results are given in Table 5.

As shown in the table, training with only MSE reconstruction loss provides the basic baseline, where the discriminative ability is harmed, leading to significant performance drop. Introducing adversarial loss will contribute to more realistic and therefore more discriminative images than the MSE loss, but are below dense CNN feature extraction baseline. Combining adversarial loss and reconstruction loss will improve the performance slightly.

As for the reconstruction loss, we provide comparison between three forms of $\mathcal{L}^{Rec}$: MSE loss, feature map loss and feature embedding loss. MSE loss focuses on low level visual characteristics, and thus can make little contribution to the discriminative power of the extracted feature. Feature map loss exert supervision on high level activation map and is closer to perceptual similarity. Such characteristics can help to distinguish person in some degree, but still cannot guarantee the distribution in the final feature embedding. On the contrary, our proposed $\mathcal{L}^{Rec}_{FC}$ directly supervises the feature embedding itself, and introduces metric learning into the training, thus can make the aggregated images even more dividable in the feature space. The results prove our theory.

The most important observation is that bringing discriminative loss $\mathcal{L}^{Dis}$ to the system can greatly boost the recognition performance, which is also the main contribution of

Figure 3. The examples of original video frames and the aggregated images (on the left), and the distribution of their features after t-sne [33] (on the right). The crossings refer to original video frames and the dots refer to synthesized images. From the distribution we can see that DAN can decrease the intra-class distance while increase the inter-class distance.

our article. By combining discriminative loss $\mathcal{L}^{Dis}$ and feature embedding based reconstruction loss $\mathcal{L}_{FC}^{Rec}$ we can obtain the best result beyond CNN baseline.

Visualization: To investigate into the effectiveness of the proposed DAN, we visualize some of the results in Figure 3. The visualization consists of 2 parts: raw video frames and aggregated images, and their distribution in the feature space after reducing dimension to two with t-sne [33] operations. We can see from the figure that the aggregated images are visually similar to the original data and of very good quality, including good positions, viewing angles, illuminations, *etc*., which are very important for recognition. Bad quality frames with blurring or unfavorable viewing angles are denoised during the process.

As for the feature space representation, we plot the distribution of original video with 200 frames and the aggregated 10 images. We can see from the figure that DAN helps to enlarge the margin between negative video pairs, especially the first example, and reduce the intra-class distance. This well demonstrates that DAN aggregated images can have better discriminative power and robustness than the original video.

5. Conclusion

In this paper, we propose a discriminative aggregation network (DAN) method for effective and efficient video-based face recognition. By combining metric learning and adversarial learning, our DAN can aggregate the useful information of an input video into one or few more discriminative images in the feature space, which can be used for face recognition. DAN is one of the first aggregation frameworks that takes raw video frames as input instead of feature embeddings, further utilizing the raw information. After the aggregation, the generated images have smaller intra-class distances and greater inter-class distances in the feature space, contributing to the discriminative power and robustness of the system. Furthermore, runtime is greatly reduced as we only need to pass few output images through feature extraction network for recognition. Experimental results on three widely used datasets demonstrate the effectiveness of our framework.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments and helpful advice. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001004, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

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, 5

[2] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010. 1, 2

[3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *WACV*, pages 1–9, 2016. 1

[4] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *ICCVW*, pages 118–126, 2015. 1

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180, 2016. 1, 2, 5

[6] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *PAMI*, 2017. 2

[7] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *PAMI*, 38(2):295–307, 2016. 4

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 2, 6

[9] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009. 2

[10] M. Hayat, M. Bennamoun, and S. An. Deep reconstruction models for image set classification. *PAMI*, 37(4):713–727, 2015. 6

[11] 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. 4

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[13] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014. 1, 2, 6

[14] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *ACCV*, pages 252–267, 2014. 6

[15] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011. 2

[16] Z. Huang and L. Van Gool. A riemannian network for spd matrix learning. *arXiv preprint arXiv:1608.04233*, 2016. 1, 2, 6

[17] Z. Huang, R. Wang, S. Shan, and X. Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *CVPR*, pages 1677–1684, 2014. 2

[18] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, pages 720–729, 2015. 2

[19] Z. Huang, J. Wu, and L. Van Gool. Building deep networks on grassmann manifolds. *arXiv preprint arXiv:1611.05742*, 2016. 1, 6

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4, 6

[21] C. Ionescu, O. Vantzos, and C. Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *ICCV*, pages 2965–2973, 2015. 6

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 1, 2, 4, 5, 6

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM-MM*, pages 675–678, 2014. 6

[24] 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. 2, 5

[25] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[27] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 1, 2, 5

[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 4

[29] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, pages 17–33, 2014. 1, 6

[30] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015. 2, 6

[31] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013. 2

[32] J. Lu, G. Wang, and P. Moulin. Localized multifeature metric learning for image-set-based face recognition. *TCSVT*, 26(3):529–540, 2016. 1, 2, 6

[33] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 8

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015. 1, 2, 6

[35] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2, 5

[36] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, volume 3, 2016. 1, 2, 5

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 1, 2, 6, 7

[38] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016. 4

[39] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015. 2, 6

[40] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 2, 6

[41] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009. 6

[42] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012. 2

[43] 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, 2, 4, 6

[44] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011. 1, 2, 5

[45] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016. 1, 2, 6, 7

[46] M. Yang, X. Wang, W. Liu, and L. Shen. Joint regularized nearest points for image set based face recognition. *IVC*, 2016. 6

[47] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 1, 2, 5

[48] 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. 6