Video-Based Face Recognition

9

Anirudh Nanduri, Jingxiao Zheng, and Rama Chellappa

9.1 Introduction

Video-based face recognition is an active research topic because of a wide range of applications including visual surveillance, access control, video content analysis, etc. Compared to still face recognition, video-based face recognition is more challenging due to a much larger amount of data to be processed and significant intra/inter-class variations caused by motion blur, low video quality, occlusion, frequent scene changes, and unconstrained acquisition conditions.

To develop the next generation of unconstrained video-based face recognition systems, two datasets have been recently introduced, IARPA Benchmark B (IJB-B) [54] and IARPA Janus Surveillance Video Benchmark (IJB-S) [23], acquired under more challenging scenarios, compared to the Multiple Biometric Grand Challenge (MBGC) dataset [30] and the Face and Ocular Challenge Series (FOCS) dataset [32] which were collected in relatively controlled conditions. IJB-B and IJB-S datasets were captured in unconstrained settings and contain faces with much more intra/inter-class variations on pose, illumination, occlusion, video quality, scale, etc.

A. Nanduri (✉)
University of Maryland, College Park MD 20742, USA
e-mail: snanduri@umd.edu

J. Zheng
Waymo, Mountain View CA 94043, USA
e-mail: jingxiaozheng@waymo.com

R. Chellappa
Johns Hopkins University, Baltimore MD 21218, USA
e-mail: rchella4@jhu.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

S. Z. Li et al. (eds.), *Handbook of Face Recognition*,
https://doi.org/10.1007/978-3-031-43567-6_9

The IJB-B dataset is a template-based dataset that contains 1845 subjects with 11,754 images, 55,025 frames, and 7,011 videos where a template consists of a varying number of still images and video frames from different sources. These images and videos are totally unconstrained, with large variations in pose, illumination, image quality, etc. Samples from this dataset are shown in Fig. 9.1. In addition, the dataset comes with protocols for 1-to-1 template-based face verification, 1-to-N template-based open-set face identification, and 1-to-N open-set video face identification. For the video face identification protocol, the gallery is a set of still-image templates. The probe is a set of videos (e.g., news videos), each of which contains multiple shots with multiple people and one bounding box annotation to specify the subject of interest. Probes of videos are searched among galleries of still images. Since the videos are composed of multiple shots, it is challenging to detect and associate the faces for the subject of interest across shots due to large appearance changes. In addition, how to efficiently leverage information from multiple frames is another challenge, especially when the frames are noisy.

Similar to the IJB-B dataset, the IJB-S dataset is also an unconstrained video dataset focusing on real-world visual surveillance scenarios. It consists of 202 subjects from 1421 images and 398 surveillance videos, with 15,881,408 bounding box annotations. Samples of frames from IJB-S are shown in Fig. 9.2. Three open-set identification protocols accompany this dataset for surveillance video-based face recognition where each video in these protocols is captured from a static surveillance camera and contains single or multiple subjects: (1) in surveillance-to-single protocol, probes collected from surveillance videos are searched in galleries consisting of one single high-resolution still-image; (2) in surveillance-to-booking protocol, same probes are searched among galleries consisting of seven high-resolution

Fig. 9.1 Example frames of a multiple-shot probe video in the IJB-B dataset. The target annotation is in the red box and face detection results from the face detector are in green boxes

Fig. 9.2 Example frames of two single-shot probe videos in the IJB-S dataset

still face images covering frontal and profile poses. Probe templates in (1) and (2) should be detected and constructed by the recognition system itself; (3) in the most challenging surveillance-to-surveillance protocol, both gallery and probe templates are from videos, which implies that probe templates need to be compared with relatively low-quality gallery templates.

From these datasets, we summarize the four common challenges in video-based face recognition as follows:

1. For video-based face recognition, test data are from videos where each video contains tens of thousands of frames and each frame may have several faces. This makes the scalability of video-based face recognition a challenging problem. In order to make the face recognition system to be operationally effective, each component of the system should be fast, especially face detection, which is often the bottleneck in recognition.
2. Since faces are mostly from unconstrained videos, they have significant variations in pose, expression, illumination, blur, occlusion, and video quality. Thus, any face representations we design must be robust to these variations and to errors in face detection and association steps.

3. Faces with the same identity across different video frames need to be grouped by a reliable face association method. Face recognition performance will degrade if faces with different identities are grouped together. Videos in the IJB-B dataset are acquired from multiple shots involving scene and view changes, while most videos in IJB-S are low-quality remote surveillance videos. These conditions increase the difficulty of face association.
4. Since each video contains a different number of faces for each identity, the next challenge is how to efficiently aggregate a varying-length set of features from the same identity into a fixed-size or unified representation. Exploiting the correlation information in a set of faces generally results in better performance than using only a single face.

In this chapter, we mainly focus on the second and fourth challenges. After face association, video faces from the same identities are associated into sets and the correlation between samples in the same set can be leveraged to improve the face recognition performance. For video-based face recognition, a temporal deep learning model such as Recurrent Neural Network (RNN) can be applied to yield a fixed-size encoded face representation. However, large-scale labeled training data is needed to learn robust representations, which is very expensive to collect in the context of the video-based recognition problem. This is also true for the adaptive pooling method [28, 57] for image set-based face recognition problems. For IJB-B and IJB-S datasets, the lack of large-scale training data makes it challenging to train an RNN-based method. Also, RNN can only work on sequential data, while faces associated from videos are sometimes without a certain order. On the contrary, representative and discriminative models based on manifolds and subspaces have also received attention for image set-based face recognition [50, 52]. These methods model sets of image samples as manifolds or subspaces and use appropriate similarity metrics for set-based identification and verification. One of the main advantages of subspace-based methods is that different from the sample mean, the subspace representation encodes the correlation information between samples. In low-quality videos, faces have significant variations due to blur, extreme poses, and low resolution. Exploiting the correlation between samples by subspaces will help to learn a more robust representation to capture these variations. Also, a fixed-size representation is learned from an arbitrary number of video frames.

To summarize, we describe an automatic system by integrating deep learning components to overcome the challenges in unconstrained video-based face recognition. The proposed system first detects faces and facial landmarks using two state-of-the-art DCNN face detectors, the Single Shot Detector (SSD) for faces [6] and the Deep Pyramid Single Shot Face Detector (DPSSD) [38]. Next, we extract deep features from the detected faces using state-of-the-art DCNNs [38] for face recognition. SORT [4] and TFA [5] are used for face association in single-shot/multiple-shot videos respectively. Finally, in the proposed face recognition system, we learn a subspace representation from each video template and match pairs of templates using principal angles-based subspace-to-subspace similarity metric on the learned subspace representations. An overview of the proposed system is shown in Fig. 9.3.

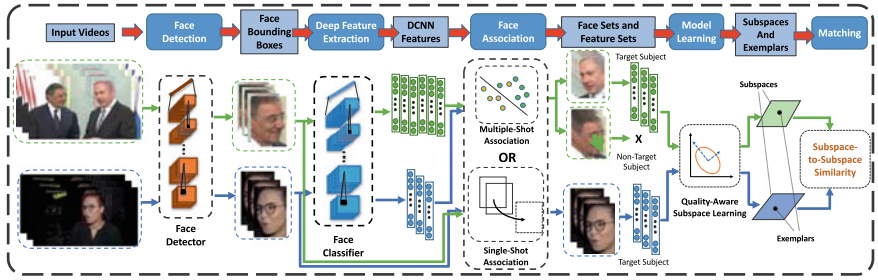

Fig. 9.3 Overview of the proposed system

We present the results of our face recognition system on the challenging IJB-B and IJB-S datasets, as well as MBGC and FOCS datasets. The results demonstrate that the proposed system achieves improved performance over other deep learning-based baselines and state-of-the-art approaches.

9.2 Related Work

9.2.1 Pre Deep Learning Methods

Frame-Based Fusion: An immediate possible utilization of temporal information for video-based face recognition is to fuse the results obtained by a 2D face recognition algorithm on each frame of the sequence. The video sequence can be seen as an unordered set of images to be used for both training and testing phases. During testing one can use the sequence as a set of probes, each of them providing a decision regarding the identity of the person. Appropriate fusion techniques can then be applied to provide the final identity. Perhaps the most frequently used fusion strategy in this case is majority voting [26, 45].

In [35], Park et al. adopt three matchers for frame-level face recognition: FaceVACS, PCA, and correlation. They use the sum rule (with min-max normalization) to fuse results obtained from the three matchers and the maximum rule to fuse results of individual frames. In [25], the concept of identity surface is proposed to represent the hyper-surface formed by projecting face patterns of an individual to the feature vector space parameterized with respect to pose. This surface is learned from gallery videos. In the testing stage, model trajectories are synthesized on the identity surfaces of enrolled subjects after the pose parameters of the probe video have been estimated. Every point on the trajectory corresponds to a frame of the video and trajectory distance is defined as a weighted sum of point-wise distances. The model trajectory that yields the minimum distance to the probe video's trajectory gives the final identification result. Based on the result that images live approximately in a bilinear space of motion and illumination variables, Xu et al. estimate these parameters for each frame of a probe video sequence with a registered 3D generic face model [56]. They then replace the generic model with a person-specific model of each subject in the gallery to synthesize video

sequences with the estimated illumination and motion parameters. Frame-wise comparison is conducted between the synthesized videos and the probe video. A synthesized video is considered as a winner if one of its frames yields the smallest distance across all frames and all the subjects in the gallery.

Ensemble Matching: Without recourse to modeling temporal dynamics, one can consider a video as an ensemble of images. Several methods have focused on utilizing image-ensembles for object and face recognition [1, 14, 16, 41]. For example, it was shown by Jacobs et al. that the illumination cone of a convex Lambertian surface can be approximated by a 9-dimensional linear subspace [3]. Motivated by this, the set of face images of the same person under varying illumination conditions is frequently modeled as a linear subspace of 9-dimensions. In such applications, an object 'category' consists of image sets of several 'instances'. A common approach in such applications is to approximate the image space of a single face/object under these variations as a linear subspace. A simplistic model for object appearance variations is then a mixture of subspaces. Zhou and Chellappa study the problem of measuring similarity between two ensembles by projecting the data into a Reproducing Kernel Hilbert Space (RKHS). The ensemble distance is then characterized as the probabilistic distance (Chernoff distance, Bhattacharyya distance, Kullback–Leibler (KL) divergence, etc.) in RKHS.

Appearance Modeling: Most face recognition approaches rely on a model of appearance for each individual subject. The simplest appearance model is a static image of the person. Such appearance models are rather limited in utility in video-based face recognition tasks where subjects may be imaged under varying viewpoints, illuminations, expressions, etc. Thus, instead of using a static image as an appearance model, a sufficiently long video that encompasses several variations in facial appearance can lend itself to building more robust appearance models. Several methods have been proposed for extracting more descriptive appearance models from videos. For example, a facial video is considered as a sequence of images sampled from an "appearance manifold". In principle, the appearance manifold of a subject contains all possible appearances of the subject. In practice, the appearance manifold for each person is estimated from training data of videos. For ease of estimation, the appearance manifold is considered to be a collection of affine subspaces, where each subspace encodes a set of similar appearances of the subject. Temporal variations of appearances in a given video sequence are then modeled as transitions between the appearance subspaces. This method is robust to large appearance changes if sufficient 3D view variations and illumination variations are available in the training set. Further, the tracking problem can be integrated into this framework by searching for a bounding box on the test image that minimizes the distance of the cropped region to the learned appearance manifold.

Basri and Jacobs [3] represent the appearance variations due to shape and illumination on human faces, using the assumption that the 'shape-illumination manifold' of all possible illuminations and head poses is generic for human faces. This means that the shape-illumination manifold can be estimated using a set of subjects exclusive of the test set. They show that the effects of face shape and illumination can be learned using Probabilistic PCA from a small,

unlabeled set of video sequences of faces in randomly varying lighting conditions. Given a novel sequence, the learned model is used to decompose the face appearance manifold into albedo and shape-illumination manifolds, producing the classification decision using robust likelihood estimation.

Wang et al. [52] proposed a Manifold-to-Manifold Distance (MMD) for face recognition based on image sets. In [51], the proposed approach models the image set with its second-order statistic for image set classification.

Chen et al. [9] and [10] proposed a video-based face recognition algorithm using sparse representations and dictionary learning. They used the identity information (face, body, and motion) in multiple frames and the accompanying dynamic signature to recognize people in unconstrained videos. Their approach is based on video-dictionaries for face and body. Video-dictionaries are a generalization of sparse representation and dictionaries for still images. They design the video-dictionaries to implicitly encode temporal, pose, and illumination information. In addition, the video-dictionaries are learned for both face and body, which enables the algorithm to encode both identity cues. To increase the ability to learn nonlinearities, they further apply kernel methods for learning dictionaries. Zheng et al. [60] proposed a hybrid dictionary learning and matching approach for video-based face recognition.

9.2.2 Deep Learning Based Methods

Face Recognition: Taigman et al. [49] learned a DCNN model on the frontalized faces generated from 3D shape models built from the face dataset. Sun et al. [46, 47] achieved results surpassing human performance for face verification on the LFW dataset [21]. Schroff et al. [44] adopted the GoogLeNet trained for object recognition to face recognition and trained on a large-scale unaligned face dataset. Parkhi et al. [36] achieved impressive results using a very deep convolutional network based on VGGNet for face verification. Ding et al. [12] proposed a trunk-branch ensemble CNN model for video-based face recognition. Chen et al. [7] trained a 10-layer CNN on CASIAWebFace dataset [59] followed by the JB metric and achieved state-of-the-art performance on the IJB-A [24] dataset. Chen et al. [8] further extended [7] and designed an end-to-end system for unconstrained face recognition and reported a very good performance on IJB-A, JANUS CS2, LFW, and YouTubeFaces [55] datasets. In order to tackle the training bottleneck for the face recognition network, Ranjan et al. [37] proposed the crystal loss to train the network on very large-scale training data. Zheng et al. [61] achieved good performance on video face datasets including IJB-B [54] and IJB-S [23]. Deng et al. [11] introduced sub-center Additive Angular Margin Loss (ArcFace) loss which significantly increases the discriminative power of the model and also makes it less susceptible to label noise by encouraging one dominant sub-class that contains the majority of clean faces and non-dominant sub-classes that include hard/noisy faces.

Video Face Recognition: Most deep-learning-based video face recognition methods extract the features from each frame and take a weighted average of them. [14, 16, 29, 58] use attention weights or quality scores to aggregate the features. Some methods like [27, 31, 42] model the spatio-temporal information with an attention mechanism to find the focus of video frames. [34, 41] propose synthesizing representative face images from a video sequence.

9.3 Method

For each video, we first detect faces from video frames and align them using the detected fiducial points. Deep features are then extracted for each detected face using our DCNN models for face recognition. Based on different scenarios, we use face association or face tracking to construct face templates with unique identities. For videos with multiple shots, we use the face association technique TFA [5] to collect faces from the same identities across shots. For single-shot videos, we use the face tracking algorithm SORT introduced in [4] to generate tracklets of faces. After templates are constructed, in order to aggregate face representations in videos, subspaces are learned using quality-aware principal component analysis. Subspaces along with quality-aware exemplars of templates are used to produce the similarity scores between video pairs by a quality-aware principal angle-based subspace-to-subspace similarity metric. In the following sections, we discuss the proposed video-based face recognition system in detail.

9.3.1 Face/Fiducial Detection

The first step in our face recognition pipeline is to detect faces in images (usually for galleries) and videos. We use two DCNN-based detectors in our pipeline based on different distributions of input.

For regular images and video frames, faces are relatively bigger and with higher resolution. We use SSD trained with the WIDER face dataset as our face detector [6]. For small and remote faces in surveillance videos, we use DPSSD [38] for face detection. DPSSD is fast and capable of detecting tiny faces, which is very suitable for face detection in videos.

After raw face detection bounding boxes are generated using either SSD or DPSSD detectors, we use All-in-One Face [40] for fiducial localization. It is followed by a seven-point face alignment step based on the similarity transform on all the detected faces.

9.3.2 Deep Feature Representation

After faces are detected and aligned, we use the DCNN models to represent each detected face. The models are state-of-the-art networks with different architectures for face recognition. Different architectures provide different error patterns during testing. After fusing the

results from different models, we achieve performance better than a single model. Design details of these networks along with their training details are described in Sect. 9.4.2.

9.3.3 Face Association

In previous steps, we obtain raw face detection bounding boxes using our detectors. Features for the detected bounding boxes are extracted using face recognition networks. The next important step in our face recognition pipeline is to combine the detected bounding boxes from the same identity to construct templates for good face recognition results.

For single-shot videos, which means the bounding boxes of a certain identity will probably be contiguous, we rely on SORT [4] to build the tracklets for each identity. For multi-shot videos, it is challenging to continue tracking across different scenes. In the proposed system, we use [5] to adaptively update the face associations through one-shot SVMs.

9.3.4 Model Learning: Deep Subspace Representation

After deep features are extracted for each face template, since each template contains a varying number of faces, these features are further encoded into a fixed-size and unified representation for efficient face recognition.

The simplest representation of a set of samples is the sample mean. However, video templates contain faces with different quality and large variations in illumination, blur, and pose. Since average pooling treats all the samples equally, the outliers may deteriorate the discriminative power of the representation. Different from other feature aggregation approaches that require a large amount of extra training data which are not available for datasets like IJB-B and IJB-S, we propose a subspace representation for video face templates.

9.3.4.1 Subspace Learning from Deep Representations

A d-dimensional subspace S can be uniquely defined by an orthonormal basis $\mathbf{P} \in \mathbb{R}^{D\times d}$, where D is the dimension of features. Given face features from a video sequence $\mathbf{Y} \in \mathbb{R}^{D\times N}$, where N is the sequence length, $\mathbf{P}$ can be found by optimizing:

$$\underset{\mathbf{P},\mathbf{X}}{\text{minimize}} \ \|\mathbf{Y} - \mathbf{P}\mathbf{X}\|_F^2 \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{9.1}$$

which is the reconstruction error of features $\mathbf{Y}$ in the subspace S. It is exactly the principal component analysis (PCA) problem and can be easily solved by eigenvalue decomposition. Let $\mathbf{Y}\mathbf{Y}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be the eigenvalue decomposition, where $\mathbf{U} = \left[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_D\right]$ are eigenvectors and $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_D\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ are the corresponding eigenvalues, we have $\mathbf{P} = \left[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d\right]$ consisting of the first d basis in $\mathbf{U}$. We use **Sub** to denote this basic subspace learning algorithm (9.1).

9.3.4.2 Quality-Aware Subspace Learning from Deep Representations

In a face template from videos, faces contain large variations in pose, illumination, occlusion, etc. Even in a tracklet, faces have different poses because of head movement, or being occluded in some frames because of the interaction with the environment. When learning the subspace, treating the frames equally is not an optimal solution. In our system, the detection score for each face bounding box provided by the face detector can be used as a good indicator of the face quality, as shown in [37]. Hence, following the quality pooling proposed in [37], we propose quality-aware subspace learning based on detection scores. The learning problem is modified (9.1) as

$$\underset{\mathbf{P},\mathbf{X}}{\text{minimize}} \sum_{i=1}^{N} \tilde{d}_i \|\mathbf{y}_i - \mathbf{P}\mathbf{x}_i\|_2^2 \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{9.2}$$

where $\tilde{d}_i = softmax(ql_i)$ is the normalized detection score of face i, q is the temperature parameter and

$$l_i = \min(\frac{1}{2}\log\frac{d_i}{1-d_i}, t) \tag{9.3}$$

which is upper bounded by threshold t to avoid extreme values when the detection score is close to 1.

Let $\tilde{\mathbf{Y}} = \left[\sqrt{d_1}\mathbf{y}_1, \cdots, \sqrt{d_N}\mathbf{y}_N\right]$ be the normalized feature set, and the corresponding eigenvalue decomposition be $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}^T$. We have

$$\mathbf{P}_D = \left[\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \cdots, \tilde{\mathbf{u}}_d\right] \tag{9.4}$$

which consists of the first d bases in $\tilde{\mathbf{U}}$. The new subspace is therefore learned by treating samples differently according to their quality. This quality-aware learning algorithm is denoted as **QSub**.

9.3.5 Matching: Subspace-to-Subspace Similarity for Videos

After subspace representations are learned for video templates, inspired by a manifold-to-manifold distance [52], we measure the similarity between two video templates of faces using a subspace-to-subspace similarity metric. In this part, we first introduce the widely used metric based on principal angles. Then we propose several weighted subspace-to-subspace metrics which take the importance of basis directions into consideration.

9.3.5.1 Principal Angles and Projection Metric

One of the most used subspace-to-subspace similarities is based on principal angles. The principal angles $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_r \leq \frac{\pi}{2}$ between two linear subspaces S_1 and S_2 can be computed by Singular Value Decomposition (SVD).

Let $\mathbf{P}_1 \in \mathbb{R}^{D\times d_1}$, $\mathbf{P}_2 \in \mathbb{R}^{D\times d_2}$, denoting the orthonormal basis of S_1 and S_2, respectively. The SVD is $\mathbf{P}_1^T\mathbf{P}_2 = \mathbf{Q}_{12}\mathbf{\Lambda}\mathbf{Q}_{21}^T$, where $\mathbf{\Lambda} = diag\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$. $\mathbf{Q}_{12}$ and $\mathbf{Q}_{21}$ are orthonormal matrices. The singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ are exactly the cosine of the principal angles as $\cos\theta_k = \sigma_k,\ k = 1, 2, \ldots, r$.

Projection metric [13] is a popular similarity metric based on principal angles:

$$s_{PM}(S_1, S_2) = \sqrt{\frac{1}{r}\sum_{k=1}^{r}\cos^2\theta_k} \tag{9.5}$$

Since $\|\mathbf{P}_1^T\mathbf{P}_2\|_F^2 = \|\mathbf{Q}_{12}\mathbf{\Lambda}\mathbf{Q}_{21}^T\|_F^2 = \|\mathbf{\Lambda}\|_F^2 = \sum_{k=1}^{r}\sigma_k^2 = \sum_{k=1}^{r}\cos^2\theta_k$, we have

$$s_{PM}(S_1, S_2) = s_{PM}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\frac{1}{r}\|\mathbf{P}_1^T\mathbf{P}_2\|_F^2} \tag{9.6}$$

and there is no need to explicitly compute the SVD. We use **PM** to denote this similarity metric (9.6).

9.3.5.2 Exemplars and Basic Subspace-to-Subspace Similarity

Existing face recognition systems usually use cosine similarity between exemplars to measure the similarity between templates. The exemplar of a template is defined as its sample mean, as $\mathbf{e} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{y}_i$, where $\mathbf{y}_i$ are samples in the template. Exemplars mainly capture the average and global representation of the template. On the other hand, the projection metric we introduced above measures the similarity between two subspaces, which models the correlation between samples. Hence, in the proposed system, we make use of both of them by fusing their similarity scores as the subspace-to-subspace similarity between two video sequences.

Suppose subspaces $\mathbf{P}_1 \in \mathbb{R}^{D\times d_1}$ and $\mathbf{P}_2 \in \mathbb{R}^{D\times d_2}$ are learned from a pair of video templates $\mathbf{Y}_1 \in \mathbb{R}^{D\times L_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{D\times L_2}$ in deep features respectively, by either **Sub** or **QSub** methods introduced in Sect. 9.3.4. Their exemplars are $\mathbf{e}_1 = \frac{1}{L_1}\sum_{i=1}^{L_1}\mathbf{y}_{1i}$ and $\mathbf{e}_2 = \frac{1}{L_2}\sum_{i=1}^{L_2}\mathbf{y}_{2i}$ respectively. Combining the orthonormal bases and exemplars, the subspace-to-subspace similarity can be computed as

$$\begin{aligned} s(\mathbf{Y}_1, \mathbf{Y}_2) &= s_{Cos}(\mathbf{Y}_1, \mathbf{Y}_2) + \lambda s_{PM}(\mathbf{P}_1, \mathbf{P}_2) \\ &= \frac{\mathbf{e}_1^T\mathbf{e}_2}{\|\mathbf{e}_1\|_2\|\mathbf{e}_2\|_2} + \lambda\sqrt{\frac{1}{r}\|\mathbf{P}_1^T\mathbf{P}_2\|_F^2} \end{aligned} \tag{9.7}$$

where $s_{Cos}(\mathbf{Y}_1, \mathbf{Y}_2)$ is the cosine similarity between exemplars, denoted as **Cos**, and $s_{PM}(\mathbf{P}_1, \mathbf{P}_2)$ is computed by (9.6). Since the DCNN features are more robust if we keep their signs, instead of using $s_{Cos}^2(\mathbf{Y}_1, \mathbf{Y}_2)$ as in [52] where the sign information is lost, we use $s_{Cos}(\mathbf{Y}_1, \mathbf{Y}_2)$ in our formulation. Accordingly, we also take the square root of the principal angle term to keep the scale consistent. λ here is a hyperparameter that balances the cosine

similarity and principal angle similarity. If $\mathbf{P}_i$'s are learned by **Sub**, we denote the whole similarity metric (including exemplars computing and subspace learning) as **Cos+Sub-PM**. If $\mathbf{P}_i$'s are learned by the proposed **QSub**, we denote the similarity as **Cos+QSub-PM**.

9.3.5.3 Quality-Aware Exemplars

In either **Cos+Sub-PM** or **Cos+QSub-PM** we are still using simple average pooling to compute the exemplars. But as discussed in Sect. 9.3.4, templates consist of faces of different quality. Treating them equally in pooling will let low-quality faces deteriorate the global representation of the template. Therefore, we propose to use the same normalized detection score as in Sect. 9.3.4 to compute the quality-aware exemplars by $\mathbf{e}_D = \frac{1}{L}\sum_{i=1}^{L} \tilde{d}_i \mathbf{y}_i$, where $\tilde{d}_i = softmax(ql_i)$ and l_i are computed by (9.3). Then, the cosine similarity between the quality-aware exemplars is

$$s_{QCos}(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\mathbf{e}_{D1}^T \mathbf{e}_{D2}}{\|\mathbf{e}_{D1}\|_2 \|\mathbf{e}_{D2}\|_2} \tag{9.8}$$

and we denote it as **QCos**. Using the new cosine similarity, the similarity becomes

$$s(\mathbf{Y}_1, \mathbf{Y}_2) = s_{QCos}(\mathbf{Y}_1, \mathbf{Y}_2) + \lambda s_{PM}(\mathbf{P}_1, \mathbf{P}_2) \tag{9.9}$$

If P_i's are learned by **QSub**, the similarity is further denoted by **QCos+QSub-PM**.

9.3.5.4 Variance-Aware Projection Metric

As previously discussed, the projection metric $S_{PM}(S_1, S_2)$ is the square root of the mean square of principle angles between two subspaces and it treats each basis direction in each subspace equally. But these basis vectors are actually eigenvectors of an eigenvalue decomposition problem. Different basis vectors correspond to different eigenvalues, which represent the variance of data in the corresponding direction. Obviously, those basis directions with larger variances contain more information than those with smaller variances. Therefore, based on the variance of each basis direction, we propose a variance-aware projection metric:

$$s_{VPM}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\frac{1}{r}\|\tilde{\mathbf{P}}_1^T \tilde{\mathbf{P}}_2\|_F^2} \tag{9.10}$$

where

$$\tilde{\mathbf{P}}_i = \frac{1}{tr(\log(\mathbf{\Lambda}_i))} \mathbf{P}_i \log(\mathbf{\Lambda}_i) \tag{9.11}$$

$\mathbf{\Lambda}_i$ is a diagonal matrix whose diagonals are eigenvalues corresponding to eigenvectors in $\mathbf{P}_i$. $\frac{1}{tr(\log(\mathbf{\Lambda}_i))}$ is the normalization factor. We use the logarithm of variance to weigh different basis directions in a subspace. This similarity metric is inspired by the Log-Euclidean distance used for image set classification in [51]. Empirically, we use $\max(0, \log(\mathbf{\Lambda}_i))$ instead of $\log(\mathbf{\Lambda}_i)$ to avoid negative weights. We use **VPM** to denote this similarity metric (9.10).

9.3.5.5 Quality-Aware Subspace-to-Subspace Similarity

By combining the quality-aware subspace learning, quality-aware exemplars and variance-aware projection metric, we propose the quality-aware subspace-to-subspace similarity between two video templates as

$$s(\mathbf{Y}_1, \mathbf{Y}_2) = s_{QCos}(\mathbf{Y}_1, \mathbf{Y}_2) + \lambda s_{VPM}(\mathbf{P}_{D1}, \mathbf{P}_{D2}) \tag{9.12}$$

where s_{QCos} is defined in (9.8), $\mathbf{P}_{Di}$'s are learned by (9.4) and s_{VPM} is defined in (9.10). This similarity metric is denoted as **QCos+QSub-VPM**. Comparisons of the proposed similarity metrics and other baselines on several challenging datasets are discussed in Sect. 9.4.

9.4 Experiments

In this section, we report video-based face recognition results for the proposed system on two challenging video face datasets, IARPA Janus Benchmark B (IJB-B) and IARPA Janus Surveillance Video Benchmark (IJB-S), and compare them with other baseline methods. We also provide results on Multiple Biometric Grand Challenge (MBGC), and Face and Ocular Challenge Series (FOCS) datasets, to demonstrate the effectiveness of the proposed system. We introduce the details of datasets, protocols, and our training and testing procedures in the following sections.

9.4.1 Datasets

IARPA Janus Benchmark B (IJB-B): IJB-B dataset is an unconstrained face recognition dataset. It contains 1845 subjects with 11,754 images, 55,025 frames, and 7,011 multiple-shot videos. IJB-B is a template-based dataset where a template consists of a varying number of still images or video frames from different sources. A template can be either an image-only, video-frame-only, or mixed-media template. Sample frames from this dataset are shown in Fig. 9.1.

In this work, we only focus on the 1:N video protocol of IJB-B. It is an open-set 1:N identification protocol where each given probe is collected from a video and is searched among all gallery faces. Gallery candidates are ranked according to their similarity scores to the probes. Top-K rank accuracy and True Positive Identification Rate (TPIR) over False Positive Identification Rate(FPIR) are used to evaluate the performance. The gallery templates are separated into two splits, G_1 and G_2, all consisting of still images. For each video, we are given the frame index with a face bounding box of the first occurrence of the target subject, as shown in Fig. 9.1. Based on this anchor, all the faces in that video with the same identity should be collected to construct the probes. The identity of the first occurrence bounding box will be considered as the template identity for evaluation.

IARPA Janus Surveillance Video Benchmark (IJB-S): Similar to IJB-B, the IJB-S dataset is also a template-based, unconstrained video face recognition dataset. It contains faces in two separate domains: high-resolution still images for galleries and low-quality, remotely captured surveillance videos for probes. It consists of 202 subjects from 1421 images and 398 single-shot surveillance videos. The number of subjects is small compared to IJB-B, but it is even more challenging due to the low-quality nature of surveillance videos.

Based on the choices of galleries and probes, we are interested in three different surveillance video-based face recognition protocols: surveillance-to-single protocol, surveillance-to-booking protocol, and surveillance-to-surveillance protocol. These are all open-set 1:N protocols where each probe is searched among the given galleries. Like IJB-B, the probe templates are collected from videos, but no annotations are provided. Thus raw face detections are grouped to construct templates with the same identities.

Galleries consist of only single frontal high-resolution images for surveillance-to-single protocol. Galleries are constructed by both frontal and multiple-pose high-resolution images for surveillance-to-booking protocol. For the most challenging surveillance-to-surveillance protocol, galleries are collected from surveillance videos as well, with given bounding boxes. In all three protocols, gallery templates are split into two splits, G_1 and G_2. During evaluation, the detected faces in videos are first matched to the ground truth bounding boxes to find their corresponding identity information. The majority of identities that appear in each template will be considered as the identity of the template and will be used for further identification evaluation. Example frames are shown in Fig. 9.2. Notice the remote faces are of very low quality.

Multiple Biometric Grand Challenge (MBGC): The MBGC Version 1 dataset contains 399 walking (frontal face) and 371 activity (profile face) video sequences from 146 people. Figure 9.4 shows some sample frames from different walking and activity videos. In the testing protocol, verification is specified by two sets: target and query. The protocol requires the algorithm to match each target sequence with all query sequences. Three verification experiments are defined: walking-vs-walking (WW), activity-vs-activity (AA), and activity-vs-walking (AW).

Face and Ocular Challenge Series (FOCS): The video challenge of FOCS is designed for frontal and non-frontal video sequence matching. The FOCS UT Dallas dataset contains 510 walking (frontal face) and 506 activity (non-frontal face) video sequences of 295 subjects with a frame size of 720×480 pixels. Like MBGC, FOCS specifies three verification protocols: walking-vs-walking, activity-vs-walking, and activity-vs-activity. In these experiments, 481 walking videos and 477 activity videos are chosen as query videos. The size of target sets ranges from 109 to 135 video sequences. Sample video frames from this dataset are shown in Fig. 9.4.

IJB-MDF: The IARPA JANUS Benchmark Multi-domain Face (IJB-MDF) dataset consists of images and videos of 251 subjects captured using a variety of cameras corresponding to visible, short-, mid-, and long-wave infrared and long-range surveillance domains. There

(a) MBGC Walking (b) MBGC Activity

(c) FOCS Walking (d) FOCS Activity

Fig. 9.4 Examples of MBGC and FOCS datasets

are 1,757 visible enrollment images, 40,597 short-wave infrared (SWIR) enrollment images, and over 800 videos spanning 161 hours.

9.4.2 Implementation Details

In this section, we discuss the implementation details for each dataset respectively.

9.4.2.1 IJB-B

For the IJB-B dataset, we employ the SSD face detector [6] to extract the face bounding boxes in all images and video frames. We employ the facial landmark branch of All-in-One Face [40] for fiducial detection on every detected bounding box and apply facial alignment based on these fiducials using the seven-point similarity transform.

The aligned faces are further represented using three networks proposed in [39]. We denote them as Network A, Network B, and Network C. Network A modifies the ResNet-101 [20] architecture. It has an input size of dimensions 224×224 and adds an extra fully connected layer after the last convolutional layer to reduce the feature dimensionality to 512. Also, it replaces the original softmax loss with the crystal loss [37] for more stable training. Network B uses the Inception-ResNet-v2 [48] model as the base network. Similar to Network A, an additional fully-connected layer is added for dimensionality reduction.

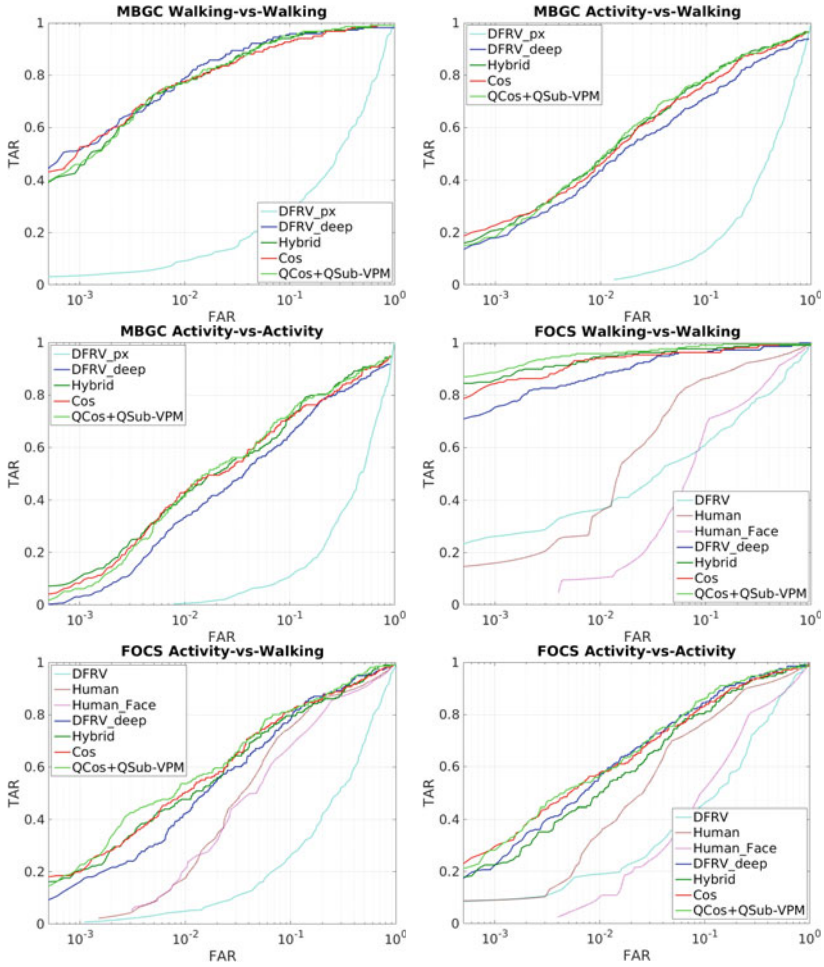

Fig. 9.5 Verification results on MBGC and FOCS datasets

Naive softmax followed by cross-entropy loss is used for this network. Network C is based on the face recognition branch in the All-in-One Face architecture [40]. The branch consists of seven convolutional layers followed by three fully-connected layers.

Network A and Network C are trained on the MSCeleb-1M dataset [19] which contains 3.7 million images from 57,440 subjects. Network B is trained on the union of three datasets called the Universe dataset: 3.7 million still images from the MSCeleb-1M dataset, 300,000 still images from the UMDFaces dataset [2], and about 1.8 million video frames from the UMDFaces Video dataset. For each network, we further reduce its dimensionality to 128 by triplet probabilistic embedding (TPE) [43] trained on the UMDFaces dataset.

For face association, we follow the details in [5]. Then, features from associated bounding boxes are used to construct the probe templates. We use quality-aware pooling for both gallery and probe templates to calculate their exemplars (**QCos**) where $t = 7$ and $q = 0.3$ are used for detection score normalization. Subspaces are built by applying the quality-aware subspace learning method (**QSub**) on each template and taking the top three eigenvectors with the largest corresponding eigenvalues. When fusing the cosine similarity and variance-aware projection similarity metric (**VPM**), we use $\lambda = 1$ so two similarity scores are fused equally. We compute the subspace-to-subspace similarity score for each network independently and combine the similarity scores from three networks by score-level fusion. We also implement baseline methods using combinations of exemplars from vanilla average pooling (**Cos**), subspaces learned by regular PCA (**Sub**), and projection similarity metric (**PM**).

9.4.2.2 IJB-S

For the IJB-S dataset, we employ the multi-scale face detector DPSSD to detect faces in surveillance videos. We only keep face bounding boxes with detection scores greater than 0.4771, to reduce the number of false detections. We use the facial landmark branch of All-in-One Face [40] as the fiducial detector. Face alignment is performed using the seven-point similarity transform.

Different from IJB-B, since IJB-S does not specify the subject of interest, we are required to localize and associate all the faces for different subjects to yield the probe sets. Since IJB-S videos are single-shot, we use SORT [4] to track every face appearing in the videos. Faces in the same tracklet are grouped to create a probe template. Since some faces in surveillance videos are of extreme pose, blur, and low resolution, to improve precision, tracklets consisting of such faces should be rejected during the recognition stage. By observation, we find that most of the short tracklets are of low quality and not reliable. The average of the detection score provided by DPSSD is also used as an indicator of the quality of the tracklet. On the other hand, we also want to take the performance of face detection into consideration to strike a balance between recall and precision. Thus in our experiments, we use two configurations for tracklets filtering: (1) We keep those tracklets with lengths greater than or equal to 25 and an average detection score greater than or equal to 0.9 to reject low-quality tracklets and focus on precision. It is referred to as **with Filtering**. (2) Following the settings in [23], we produce results without any tracklets filtering and focusing on both precision and recall. It is referred to as **without Filtering**.

Because of the remote acquisition scenario and the presence of blurred probes in the IJB-S dataset, we retrain Network A with the same crystal loss but on the Universe dataset used by Network B. We denote it as Network D. We also retrain Network B with the crystal loss [37] on the same training data. We denote it as Network E. As a combination of high-capacity network and large-scale training data, Networks D and E are more powerful than Networks A, B, and C. As before, we reduce feature dimensionality to 128 using the TPE trained on the UMDFaces dataset.

In IJB-S, subspace learning and matching parts are the same as IJB-B except that we combine the similarity score by score-level fusion from Network D and E. Notice that for the surveillance-to-surveillance protocol, we only use single Network D for representation as Network E is ineffective for low-quality gallery faces in this protocol.

9.4.2.3 MBGC and FOCS

For MBGC and FOCS datasets, we use All-in-One Face for both face detection and facial landmark localization. The MBGC and FOCS datasets contain only one person in a video in general. Hence, for each frame, we directly use the face bounding box with the highest detection score as the target face. Similar to IJB-S, bounding boxes are filtered based on detection scores. From the detected faces, deep features are extracted using Network D. Since MBGC and FOCS datasets do not provide training data, we also use the TPE trained on the UMDFaces dataset to reduce feature dimensionality to 128. For MBGC and FOCS, subspace learning and matching parts are the same as IJB-B and IJB-S.

9.4.3 Evaluation Results

In the following section, we first show some face association results on IJB-B and IJB-S datasets. Then we compare the performance of the proposed face recognition system with several baseline methods. For each dataset, all the baseline methods listed below use deep features extracted from the same network and with the same face detector.

- **Cos:** We compute the cosine similarity scores directly from the exemplars with average pooling.
- **QCos:** We compute the cosine similarity scores from the exemplars with quality-aware average pooling.
- **Cos+Sub-PM:** Subspace-to-subspace similarity is computed by fusing the plain cosine similarity and plain projection metric, and subspaces are learned by plain PCA.
- **QCos+Sub-PM:** Subspace-to-subspace similarity is computed by fusing the quality-aware cosine similarity and plain projection metric, and subspaces are learned by plain PCA.
- **QCos+QSub-PM:** Subspace-to-subspace similarity is computed by fusing the quality-aware cosine similarity and plain projection metric, and subspaces are learned by quality-aware subspace learning.
- **QCos+QSub-VPM:** Subspace-to-subspace similarity is computed by fusing the quality-aware cosine similarity and variance-aware projection metric, and subspaces are learned by quality-aware subspace learning.

IJB-B: Figures 9.6 and 9.7 show some examples of our face association results using TFA in IJB-B dataset. Table 9.1 shows the Top-K Accuracy results for IJB-B video protocol. In this dataset, besides the baselines, our method is compared with original results in [5]

Fig. 9.6 Examples of face association results by TFA on IJB-B. The target annotation is in the red box, and the associated faces of the target subject are in magenta-colored boxes

corresponding to different iteration numbers. The results shown are the average of the two galleries. Notice that our proposed system and [5] use the same face association method, but we have different networks and feature representation techniques.

IJB-S: Figure 9.8 shows some examples of our face association results using SORT in IJB-S dataset. Tables 9.2, 9.3 and 9.4 show the results for IJB-S surveillance-to-single protocol, surveillance-to-booking protocol and surveillance-to-surveillance protocol respectively. Notice that under the **with Filtering** configuration, we use the regular top-K average accuracy for evaluation. Under the **without Filtering** configuration, we use the End-to-End Retrieval Rate (EERR) metric proposed in [23] for evaluation. For surveillance-to-surveillance protocol, we show results for two different network configurations as well. We also implement state-of-the-art network ArcFace [11] on IJB-S and compare it with our method. Results from ArcFace are shown with the prefix **Arc-**.

Fig. 9.7 Associated faces by TFA corresponding to examples in Fig. 9.6. Face images are in the order of the confidence of face association

Two recent works [15, 17] have reported results on the IJB-S dataset. These works mainly focused on face recognition and not detection so they built video templates by matching their detections with ground truth bounding boxes provided by the protocols and evaluated their methods using identification accuracy and not EERR metric. Our system focuses on detection, association, and recognition. Therefore after detection, we associate faces across the video frames to build templates without utilizing any ground truth information and evaluate our system using both identification accuracy and EERR metric. Since these two template-building procedures are so different, a direct comparison is not meaningful.

MBGC: The verification results for the MBGC dataset are shown in Table 9.5 and Fig. 9.5. We compare our method with the baseline algorithms, **Hybrid** [60] and [9] using either raw pixels as $\mathbf{DFRV}_{px}$ (reported in their paper) or deep features as $\mathbf{DFRV}_{deep}$ (our implementation). We also report the results of the proposed method applied to the ArcFace features with the prefix **Arc-**. Figure 9.5 does not include all the baselines, for a clearer view. The result of [9] is not in the table because the authors did not provide exact numbers in their paper.

Table 9.1 1:N Search Top-K Average Accuracy and TPIR/FPIR of IJB-B video search protocol

Methods	Rank = 1	Rank = 2	Rank = 5	Rank = 10	Rank = 20	Rank = 50	FPIR = 0.1	FPIR = 0.01
[5] with Iteration 0	55.94%	–	68.40%	72.89%	–	83.71%	44.60%	28.73%
[5] with Iteration 3	61.01%	–	73.39%	77.90%	–	87.62%	49.73%	34.11%
[5] with Iteration 5	61.00%	–	73.46%	77.94%	–	87.69%	49.78%	33.93%
Cos	78.37%	81.35%	84.39%	86.29%	88.30%	90.82%	73.15%	52.19%
QCos	78.43%	81.41%	84.40%	86.33%	88.34%	90.88%	**73.19%**	**52.47%**
Cos+Sub-PM	77.99%	81.45%	84.68%	86.75%	88.96%	91.91%	72.31%	38.44%
QCos+Sub-PM	78.02%	81.46%	84.76%	86.72%	88.97%	91.91%	72.38%	38.88%
QCos+QSub-PM	78.04%	81.47%	84.73%	86.72%	88.97%	91.93%	72.39%	38.91%
QCos+QSub-VPM	**78.93%**	**81.99%**	**84.96%**	**87.03%**	**89.24%**	**92.02%**	71.26%	47.35%

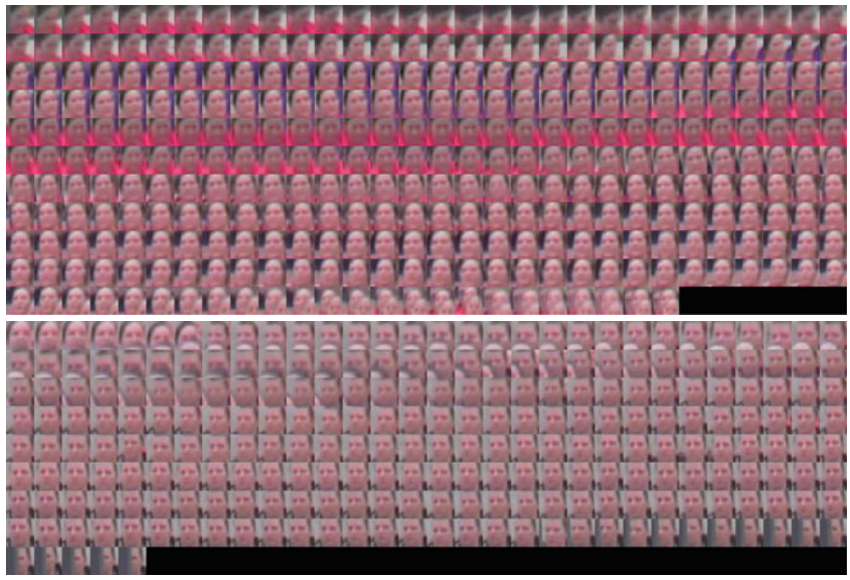

Fig. 9.8 Associated faces using SORT in IJB-S. Face images are in their temporal order. Notice the low-quality faces at the boundaries of tracklets since the tracker cannot reliably track anymore

FOCS: The verification results of FOCS dataset are shown in Table 9.5 and Fig. 9.5. O'Toole et al. [33] evaluated the human performance on this dataset. In the figures, **Human** refers to

Table 9.2 1:N Search results of IJB-S surveillance-to-single protocol. Using both Network D and E for representation

Methods	Top-K average accuracy **with filtering**						EERR metric **without filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos [11]	52.03%	56.83%	63.16%	69.05%	76.13%	88.95%	24.45%	26.54%	29.35%	32.33%	36.38%	44.81%
Arc-QCos+QSub-PM	60.92%	65.06%	70.45%	75.19%	80.69%	90.29%	28.73%	30.44%	32.98%	35.40%	38.70%	45.46%
Cos	64.86%	70.87%	77.09%	81.53%	86.11%	93.24%	29.62%	32.34%	35.60%	38.36%	41.53%	46.78%
QCos	65.42%	71.34%	77.37%	81.78%	86.25%	93.29%	29.94%	32.60%	35.85%	38.52%	41.70%	46.78%
Cos+Sub-PM	69.52%	75.15%	80.41%	84.14%	87.83%	94.27%	32.22%	34.70%	37.66%	39.91%	42.65%	47.54%
QCos+Sub-PM	69.65%	75.26%	80.43%	84.22%	87.81%	94.25%	32.27%	34.73%	37.66%	39.91%	42.67%	47.54%
QCos+QSub-PM	**69.82%**	**75.38%**	**80.54%**	**84.36%**	**87.91%**	**94.34%**	**32.43%**	**34.89%**	**37.74%**	**40.01%**	**42.77%**	**47.60%**
QCos+QSub-VPM	69.43%	75.24%	80.34%	84.14%	87.86%	94.28%	32.19%	34.75%	37.68%	39.88%	42.56%	47.50%

Table 9.3 1:N Search results of IJB-S surveillance-to-booking protocol. Using both Network D and E for representation

Methods	Top-K average accuracy **with filtering**						EERR metric **without filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos [11]	54.59%	59.12%	65.43%	71.05%	77.84%	89.16%	25.38%	27.58%	30.59%	33.42%	37.60%	45.05%
Arc-QCos+QSub-VPM	60.86%	65.36%	71.30%	76.15%	81.63%	90.70%	28.66%	30.64%	33.43%	36.11%	39.57%	45.70%
Cos	66.48%	71.98%	77.80%	82.25%	86.56%	93.41%	30.38%	32.91%	36.15%	38.77%	41.86%	46.79%
QCos	66.94%	72.41%	78.04%	82.37%	86.63%	93.43%	30.66%	33.17%	36.28%	38.84%	41.88%	46.84%
Cos+Sub-PM	69.39%	74.55%	80.06%	83.91%	87.87%	**94.34%**	32.02%	34.42%	37.59%	39.97%	42.64%	**47.58%**
QCos+Sub-PM	69.57%	74.78%	80.06%	83.89%	87.94%	94.33%	32.16%	34.61%	37.62%	39.99%	42.71%	47.57%
QCos+QSub-PM	69.67%	74.85%	80.25%	84.10%	88.04%	94.22%	32.28%	34.77%	37.76%	40.11%	**42.76%**	47.57%
QCos+QSub-VPM	**69.86%**	**75.07%**	**80.36%**	**84.32%**	**88.07%**	94.33%	**32.44%**	**34.93%**	**37.80%**	**40.14%**	42.72%	**47.58%**

human performance with all bodies of target subjects seen, and **Human_Face** refers to the performance that only faces of the target subjects are seen. Here besides baseline algorithms and **Hybrid** [60], we also compare our method with [9] in either raw pixels as $\mathbf{DFRV}_{px}$ (reported in their paper) or deep features as $\mathbf{DFRV}_{deep}$ (our implementation). We also report the results using ArcFace features. Similarly, the results of [9] and human performance are not in the table since they did not provide exact numbers.

Table 9.4 1:N Search results of IJB-S surveillance-to-surveillance protocol. D stands for only using Network D for representation. D+E stands for using both Network D and E for representation

Methods	Top-K average accuracy **with filtering**						EERR metric **without filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos [11]	8.68%	12.58%	18.79%	26.66%	39.22%	68.19%	4.98%	7.17%	10.86%	15.42%	22.34%	37.68%
Arc-QCos+QSub-PM	8.64%	12.57%	18.84%	26.86%	39.78%	**68.21%**	**5.26%**	**7.44%**	**11.31%**	15.90%	**22.68%**	**37.83%**
Cos(D+E)	9.24%	12.51%	19.36%	25.99%	32.95%	52.95%	4.74%	6.62%	10.70%	14.88%	19.29%	30.64%
QCos+QSub-VPM(D+E)	**9.56%**	**13.03%**	19.65%	27.15%	35.39%	56.02%	4.77%	6.78%	10.88%	15.52%	20.51%	32.16%
Cos(D)	8.54%	11.99%	19.60%	28.00%	37.71%	59.44%	4.42%	6.15%	10.84%	15.73%	21.14%	33.21%
QCos(D)	8.62%	12.11%	19.62%	28.14%	37.78%	59.21%	4.46%	6.20%	10.80%	15.81%	21.06%	33.17%
Cos+Sub-PM(D)	8.19%	11.79%	19.56%	28.62%	39.77%	63.15%	4.26%	6.25%	10.79%	16.18%	22.48%	34.82%
QCos+Sub-PM(D)	8.24%	11.82%	19.68%	28.68%	39.68%	62.96%	4.27%	6.25%	10.92%	16.18%	22.39%	34.69%
QCos+QSub-PM(D)	8.33%	11.88%	19.82%	28.65%	39.78%	62.79%	4.33%	6.21%	10.96%	16.19%	22.48%	34.69%
QCos+QSub-VPM(D)	8.66%	12.27%	**19.91%**	**29.03%**	**40.20%**	63.20%	4.30%	6.30%	10.99%	**16.23%**	22.50%	34.76%

Table 9.5 Verification results on MBGC and FOCS datasets

Methods	MBGC						FOCS					
	WW		AW		AA		WW		AW		AA	
	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1
Arc-Cos [11]	84.40%	92.20%	53.88%	75.00%	32.47%	66.49%	98.18%	**99.09%**	48.61%	69.44%	48.36%	78.87%
Arc-QCos+QSub-PM	**85.32%**	92.20%	**55.58%**	75.00%	32.99%	64.43%	**98.64%**	**99.09%**	52.31%	74.07%	50.23%	79.81%
DFRV$_{deep}$ [9]	78.90%	**95.87%**	43.69%	71.36%	33.51%	64.95%	87.73%	96.36%	42.13%	78.70%	56.81%	84.51%
Hybrid [60]	77.06%	94.04%	48.06%	**79.37%**	42.53%	71.39%	95.00%	97.73%	47.69%	79.63%	50.23%	80.75%
Cos	77.52%	92.66%	45.87%	76.94%	**43.30%**	71.65%	94.09%	96.36%	50.46%	81.48%	57.75%	83.57%
QCos	77.52%	92.66%	47.57%	76.94%	**43.30%**	71.13%	95.91%	**99.09%**	**53.70%**	80.09%	**58.22%**	83.57%
Cos+Sub-PM	77.98%	94.95%	47.57%	79.13%	41.24%	72.68%	91.82%	97.27%	49.07%	**83.33%**	54.93%	85.45%
QCos+Sub-PM	77.98%	94.95%	48.30%	78.64%	41.75%	**73.71%**	95.91%	98.64%	52.78%	82.87%	55.40%	**85.92%**
QCos+QSub-PM	77.52%	94.95%	48.54%	78.64%	41.75%	73.20%	95.91%	**99.09%**	52.31%	81.02%	55.87%	**85.92%**
QCos+QSub-VPM	77.06%	94.95%	48.06%	78.16%	41.24%	72.68%	95.91%	**99.09%**	**53.70%**	81.94%	56.34%	**85.92%**

9.4.4 Cross-Spectral Video Face Verification

In this section, we present some results on the IARPA JANUS Benchmark Multi-domain Face (IJB-MDF) [22] dataset. The domains in the IJB-MDF dataset are labeled as below:

- (0) visible enrollment
- (1) visible surveillance

- (2) visible gopro
- (3) visible 500m
- (4) visible 400m
- (5) visible 300m
- (6) visible 500m 400m walking
- (11) swir enrollment nofilter
- (12) swir enrollment 1150
- (13) swir enrollment 1350
- (14) swir enrollment 1550
- (15) swir 15m
- (16) swir 30m

There are a total of 251 subjects. Domains 1, 2, 3, 4, 5, 6, 15, and 16 consist of videos only, while the enrollment domains (0, 11, 12, 13, and 14) consist of still images taken in a constrained setting. Instead of performing an end-to-end evaluation, we are more interested in observing how well a feature extractor (trained on visible images) adapts to these new domains. As such, to simplify the task, we use the ground truth provided with the dataset to obtain the start and end time stamps for non-empty frames in the videos and extract all the relevant frames. The videos are captured at a frame rate of 20fps. Table 9.6 shows the distribution of the frames with respect to various domains.

We select domains 3, 4, 5, and 6 for the task of cross-spectral face recognition of remote faces. The quality of faces is sub-par with a lot of blur and lack of detail in the face. We employ the SCRFD [18] algorithm to detect faces from the video frames. The recall at a score-threshold of 0.5 is about 95%.

Table 9.6 IJB-MDF data distribution

Domain	Num videos	Num frames	Approx size of frame	Name of domain
1	358	191,971	19MB	Visible Surveillance
2	24	39,263	7 MB	Visible GoPro
3	31	56,024	6 MB	Visible 500m
4	34	61,446	4 MB	Visible 400m
5	34	61,442	1 MB	Visible 300m
6	26	24,194	7 MB	Visible 500m 400m walking
15	42	56,406	250 KB	SWIR 15m
16	42	50,368	350 KB	SWIR 30m

Table 9.7 Verification performance with SCRFD, AdaptiveWingLoss and ArcFace loss

Domain	Rank 1	Rank 2	Rank 5	Rank 10
(3) Visible 500m	20.8%	25.5%	33.2%	41.8%
(4) Visible 400m	95.0%	97.1%	98.6%	99.1%
(5) Visible 300m	98.5%	99.3%	99.7%	99.9%
(6) Visible 500m 400m walking	57.8%	64.4%	71.6%	77.5%
(3, 4, 5, 6) together	76.9%	79.5%	82.5%	85.1%

We use the AdaptiveWingLoss [53] algorithm on the cropped faces to detect the face key points. Then we perform face alignment and use the resulting images for feature extraction. For these experiments, we use a model trained on visible data (using ArcFace loss [11]) to extract features from the remote frames and evaluate face verification performance between the remote frames (probe set) and the visible enrollment images (gallery set).

Using only the frames that match the ground truth frames (removing false positives), the verification performance is shown in Table 9.7.

We can see from the results that the model adapts well to videos at 300m and 400m, but there is a definite drop in performance as we go from 400m to 500m.

9.4.5 Discussions

For the IJB-B dataset, we can see that the proposed system performs consistently better than all the results in [5] and the baseline **Cos** on identification accuracy. For open-set metric TPIR/FPIR, the proposed quality-aware cosine similarity achieves better results, but the proposed subspace similarity metric still performs better than [5] with a large margin. For the IJB-S dataset, we have similar observations: the proposed system with subspace-to-subspace similarity metric performs better than **Cos** on surveillance-to-single and surveillance-to-booking protocols, by a relatively large margin. It also achieves better accuracy than **Cos** on the surveillance-to-surveillance protocol. We notice that the fusion of Network D and E does not work well on surveillance-to-surveillance protocol, especially at higher rank accuracy. Such observations are consistent under both tracklets filtering configurations and their corresponding metrics: **with Filtering** with Top-K average accuracy and **without Filtering** with the EERR metric. The proposed system also outperforms ArcFace with a larger margin in surveillance-to-single and surveillance-to-booking protocols of IJB-S. For MBGC and FOCS datasets, from the tables and plots we can see that in general, the proposed approach performs better than **Cos** baseline, $\mathbf{DFRV}_{deep}$, $\mathbf{DFRV}_{px}$ and **Hybrid**.

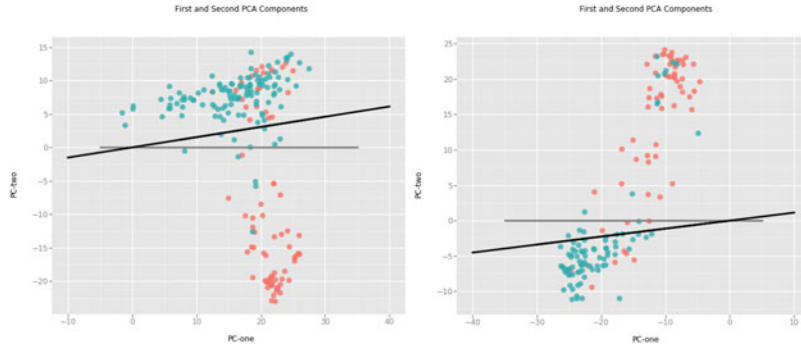

Fig. 9.9 Visualization of example templates in IJB-S. Each sample is a dot in the plot with its first two principal components as the coordinates. Samples with $d_i \geq 0.7$ are in **blue** dots and the rest samples are in **red** dots. **Gray** line and **black** line are the projection of the first subspace basis learned by **Sub** and **QSub** respectively

Figure 9.9 shows the visualization of two templates in IJB-S dataset in PCA-subspace, which illustrates the advantage of the proposed subspace learning method. In the plot, each dot corresponds to a sample in the template, where x- and y-axes correspond to the first two principal components of the samples, learned from each template respectively. Relatively high-quality detections with detection scores greater than or equal to 0.7 are represented by blue dots. Relatively low-quality detections with detection scores less than 0.7 are represented by red dots. The projections of the first subspace bases learned by **Sub** and the proposed **QSub** onto the PCA-subspace are gray and black straight lines in the plot, respectively. From the plot, we can see that, with quality-aware subspace learning, the subspaces learned by the proposed method put more weight on the high-quality sample. It fits the high-quality samples better than the low-quality ones. But the plain PCA takes each sample into account equally, which is harmful to the representation of the template.

We also compare our system with other baseline methods as part of an ablation study, from baseline cosine similarity **Cos** to the proposed quality-aware subspace-to-subspace similarity **QCos+QSub-VPM**. As we gradually modify the method by including quality-aware cosine similarity **QCos**, quality-aware subspace learning **QSub**, and variance-aware projection metric **VPM**, we can see the performance also gradually improves, especially for IJB-B and IJB-S datasets.

From the results above, we observe the following:

- The proposed system performs the best in general, which shows the effectiveness of (1) learning subspace as template representation, (2) matching video pairs using the subspace-to-subspace similarity metric and (3) utilizing quality and variance information to compute exemplars, learn subspaces and measure similarity.

- **QCos** generally performs better than **Cos**, which shows that quality-aware exemplars weigh the samples according to their quality and better represent the image sets than plain average exemplars.
- In most of the cases, **Cos+Sub-PM** achieve higher performance than **Cos**. It implies that a subspace can utilize the correlation information between samples and is a good complementary representation of exemplars as global information.
- **QCos+QSub-PM** performs better than **QCos+Sub-PM** in general. It shows that similar to **QCos**, we can learn more representative subspaces based on the quality of samples.
- **QCos+QSub-VPM** works better than **QCos+QSub-PM** in most of the experiments. It implies that by considering the variances of bases in the subspaces, **VPM** similarity is more robust to variations in the image sets.
- The improvement of the proposed system over the compared algorithms is consistent under both **with filtering** and **without filering** configurations on the IJB-S dataset. It shows that our method is effective for both high-quality and low-quality tracklets in surveillance videos.
- For IJB-S, the performance on surveillance-to-surveillance protocol is in general lower than the performance on other protocols. This is because the gallery templates of this protocol are constructed from low-quality surveillance videos, while the remaining two protocols have galleries from high-resolution still images.
- The fusion of Network D and E does not perform as well as single Network D on surveillance-to-surveillance protocol, especially at higher rank accuracy. It is probably because of the low-quality galleries in this protocol which Network E cannot represent well.
- On IJB-S, the proposed method performs better than state-of-the-art network ArcFace [11] in general, especially on surveillance-to-single and surveillance-to-booking protocols, which shows the discriminative power of the features from the proposed networks. ArcFace still performs better on surveillance-to-surveillance protocol. But the results also show that using the quality-aware subspace-to-subspace similarity improves the performance for ArcFace features as well.
- On MBGC and FOCS, ArcFace performs better in the walking-vs-walking protocol but Network D outperforms ArcFace on more challenging protocols like activity-vs-activity. Also, by applying the proposed subspace-to-subspace similarity on both features, the performance consistently improves, which shows its effectiveness on different datasets and using different features.
- For the FOCS dataset, the performance of our system surpasses the human performance, which again demonstrates the effectiveness of the proposed system.

9.5 Concluding Remarks

In this chapter, we proposed an automatic face recognition system for unconstrained video-based face recognition tasks. The proposed system learns subspaces to represent video faces and matches video pairs by subspace-to-subspace similarity metrics. We evaluated our system on four video datasets and the experimental results demonstrate the superior performance of the proposed system.

References

1. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 581–588. IEEE (2005)
2. Bansal, A., Nanduri, A., Castillo, C.D., Ranjan, R., Chellappa, R.: Umdfaces: An annotated face dataset for training deep networks. In: IEEE International Joint Conference on Biometrics (IJCB) (2017)
3. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(2), 218–233 (2003)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP, pp. 3464–3468 (2016)
5. Chen, C.H., Chen, J.C., Castillo, C.D., Chellappa, R.: Video-based face association and identification. In: 12th FG, pp. 149–156 (2017)
6. Chen, J.C., Lin, W.A., Zheng, J., Chellappa, R.: A real-time multi-task single shot face detector. In: ICIP (2018)
7. Chen, J.C., Patel, V.M., Chellappa, R.: Unconstrained face verification using deep CNN features. In: WACV (2016)
8. Chen, J.C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C.H., Patel, V.M., Castillo, C.D., Chellappa, R.: Unconstrained still/video-based face verification with deep convolutional neural networks. IJCV **126**(2), 272–291 (2018)
9. Chen, Y.C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face recognition from video. In: ECCV (2012)
10. Chen, Y.C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face and person recognition from unconstrained video. IEEE Access **3**, 1783–1798 (2015)
11. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
12. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. CoRR http://arxiv.org/abs/1607.05427 (2016)
13. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998)
14. Gong, S., Shi, Y., Jain, A.: Low quality video face recognition: Multi-mode aggregation recurrent network (marn). In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
15. Gong, S., Shi, Y., Jain, A.K., Kalka, N.D.: Recurrent embedding aggregation network for video face recognition. CoRR http://arxiv.org/abs/1904.12019 (2019)

16. Gong, S., Shi, Y., Kalka, N.D., Jain, A.K.: Video face recognition: Component-wise feature aggregation network (c-fan). In: 2019 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2019)
17. Gong, S., Shi, Y., Kalka, N.D., Jain, A.K.: Video face recognition: Component-wise feature aggregation network (C-FAN). CoRR http://arxiv.org/abs/1902.07327 (2019)
18. Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. ArXiv preprint arXiv:2105.04714 (2021)
19. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: ECCV (2016)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. ArXiv preprint arXiv:1506.01497 (2015)
21. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Tech. rep (2007)
22. Kalka, N.D., Duncan, J.A., Dawson, J., Otto, C.: Iarpa janus benchmark multi-domain face. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9. IEEE (2019)
23. Kalka, N.D., Maze, B., Duncan, J.A., O'Connor, K.J., Elliott, S., Hebert, K., Bryan, J., Jain, A.K.: IJB-S : IARPA Janus Surveillance Video Benchmark (2018)
24. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In: CVPR (2015)
25. Li, Y., Gong, S., Liddell, H.: Video-based online face recognition using identity surfaces. In: Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 40–46. IEEE (2001)
26. Liu, X., Chen, T., Thornton, S.M.: Eigenspace updating for non-stationary process and its application to face recognition. Pattern Recogn. **36**(9), 1945–1959 (2003)
27. Liu, X., Kumar, B., Yang, C., Tang, Q., You, J.: Dependency-aware attention control for unconstrained face recognition with image sets. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 548–565 (2018)
28. Liu, Y., Junjie, Y., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR (2017)
29. Liu, Z., Hu, H., Bai, J., Li, S., Lian, S.: Feature aggregation network for video face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
30. Information Technology Laboratory, NIST.: Multiple Biomertic Grand Challenge http://www.nist.gov/itl/iad/ig/mbgc.cfm
31. Mei, T., Yang, B., Yang, S.Q., Hua, X.S.: Video collage: presenting a video sequence using a single image. Vis. Comput. **25**(1), 39–51 (2009)
32. O'Toole, A.J., Harms, J., Snow, S.L., Hurst, D.R., Pappas, M.R., Ayyad, J.H., Abdi, H.: A video database of moving faces and people. TPAMI **27**(5) (2005)
33. O'Toole, A.J., Phillips, P.J., Weimer, S., Roark, D.A., Ayyad, J., Barwick, R., Dunlop, J.: Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. Vis. Res. **51**(1) (2011)
34. Parchami, M., Bashbaghi, S., Granger, E., Sayed, S.: Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
35. Park, U., Jain, A.K., Ross, A.: Face recognition in video: Adaptive fusion of multiple matchers. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
36. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)

37. Ranjan, R., Bansal, A., Xu, H., Sankaranarayanan, S., Chen, J., Castillo, C.D., Chellappa, R.: Crystal loss and quality pooling for unconstrained face verification and recognition. CoRR http://arxiv.org/abs/1804.01159 (2018)
38. Ranjan, R., Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B., Nanduri, A., Chen, J.C., Castillo, C., Chellappa, R.: A fast and accurate system for face detection, identification, and verification. CoRR http://arxiv.org/abs/1809.07586 (2018)
39. Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J.C., Patel, V.M., Castillo, C.D., Chellappa, R.: Deep learning for understanding faces: Machines may be just as good, or better, than humans. IEEE Sig. Process. Mag. **35**, 66–83 (2018)
40. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 12th IEEE FG, vol. 00, pp. 17–24 (2017)
41. Rao, Y., Lin, J., Lu, J., Zhou, J.: Learning discriminative aggregation network for video-based face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3781–3790 (2017)
42. Rao, Y., Lu, J., Zhou, J.: Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3931–3940 (2017)
43. Sankaranarayanan, S., Alavi, A., Castillo, C.D., Chellappa, R.: Triplet probabilistic embedding for face verification and clustering. CoRR http://arxiv.org/abs/1604.05417 (2016)
44. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
45. Shakhnarovich, G., Fisher, J.W., Darrell, T.: Face recognition from long-term observations. In: European Conference on Computer Vision, pp. 851–865. Springer (2002)
46. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS (2014)
47. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: CVPR (2015)
48. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
49. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014)
50. Wang, R., Chen, X.: Manifold discriminant analysis. In: CVPRW, pp. 429–436 (2009)
51. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR, pp. 2496–2503 (2012)
52. Wang, R., Shan, S., Chen, X., Dai, Q., Gao, W.: Manifold-manifold distance and its application to face recognition with image sets. IEEE TIP **21**(10) (2012)
53. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6971–6981 (2019)
54. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K.: IARPA Janus Benchmark-B face dataset. In: CVPRW (2017)
55. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR, pp. 529–534 (2011)
56. Xu, Y., Roy-Chowdhury, A., Patel, K.: Pose and illumination invariant face recognition in video. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE (2007)
57. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: CVPR, pp. 4362–4371 (2017)
58. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4362–4371 (2017)

59. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR http://arxiv.org/abs/1411.7923 (2014)
60. Zheng, J., Chen, J.C., Patel, V.M., Castillo, C.D., Chellappa, R.: Hybrid dictionary learning and matching for video-based face verification. In: BTAS (2019)
61. Zheng, J., Ranjan, R., Chen, C.H., Chen, J.C., Castillo, C.D., Chellappa, R.: An automatic system for unconstrained video-based face recognition. CoRR http://arxiv.org/abs/1812.04058 (2018)