# Still-to-Video Face Matching Using Multiple Geodesic Flows

Yu Zhu, Yan Li, Guowang Mu, Shiguang Shan, and Guodong Guo, *Senior Member, IEEE*

***Abstract*—Still-to-video (S2V) face recognition has recently attracted attention from researchers because of its great applications in real-world scenarios. In S2V FR, still images are usually of high quality, captured from cooperative users under controlled environment, such as mugshots, while video clips may be acquired with low resolutions and low quality, from non-cooperative users under uncontrolled environment. Because of those significant differences, we interpret the S2V FR as a heterogeneous matching problem, and propose an approach aiming at building multiple "bridges" between those two heterogeneous face modalities. Considering the unbalanced distributions and large diversities between two modalities, we propose to exploit a Grassmann manifold learning method to construct subspaces in between to find connections (or transitions) between the still images and video clips. Multiple geodesic flows are generated connecting the subspace of still images and the clustered subspace centers of videos, which are representative and robust to characterize the relationship between still images and video frames. Extensive experiments are conducted on two large scale benchmark databases, COX-S2V and PaSC, with different recognition tasks: face identification and verification. The experimental results show that the proposed approach outperforms the state-of-the-art methods under the same experimental settings.**

***Index Terms*—Manifold, geodesic flows, clustering, still-to-video, face recognition.**

## I. Introduction

FACE recognition (FR) has attracted some attention in computer vision and biometrics over the last few decades. In addition to the still image based face recognition, the availability of inexpensive cameras and increasing usage of surveillance systems have driven several recent works on video based face recognition [1]–[5] where faces captured by video cameras usually contain more variations caused by illumination, head pose, facial expression, and motion blur. More recently, Still-to-video (S2V) face recognition has attracted some attentions [6]–[11] because of its wide range of real-world applications, such as identifying the criminal suspects in video clips from a large mugshot database, and rapid locating and tracking target subjects with still images from the whole city security surveillance videos. Typically in S2V face recognition, still images are of high quality, captured from cooperative users under controlled environment, such as the mugshots. On the contrary, video clips are often acquired with low resolutions and low quality, from non-cooperative users under uncontrolled environment. In such scenario, the large disparity between still images and video clips poses great challenges for S2V face recognition. To address this issue, we consider the S2V face recognition as a heterogeneous matching problem [12] from different modalities, i.e., still images and videos.

Previous works on S2V face recognition mostly concentrate on learning a metric between still images and videos to build a measurement of data from different modality, but fail to simultaneously utilize the rich information and relationship between them. This prompts us to consider a different way to model the relationship between the heterogeneous modalities more appropriately. In this paper, we mainly focus on finding the "connections" between still images and videos in a natural, unsupervised manner. We explore the transitions from one modality to the other, which is used to generate models for subsequent face representation and matching. Particularly, inspired by the recent advances in subspace learning, we approach the S2V face recognition through building the connection of heterogeneous subspaces lying on the Grassmann manifold [13], characterizing the transition from still images to videos, and reducing the differences between them. In our preliminary work [12], we studied the S2V face recognition based on a manifold learning method, i.e., Geodesic Flow Kernel (GFK) [14]. Our view is that the manifold learning based approaches are promising to address the problem of S2V face recognition.

In this paper, we still consider the S2V face recognition as a heterogeneous matching problem [12], and a novel approach is proposed, as shown in Fig. 1. The approach can address the unbalanced distributions between still images and videos in a robust way by generating multiple "bridges" to connect the still images and video frames.

Our main contributions include:

- We consider the still-to-video face recognition as a heterogeneous face matching problem, which is for the first time [12]. Thus, we deal with the large differences between still images and videos frames *explicitly*.

Manuscript received February 21, 2016; revised June 8, 2016 and July 25, 2016; accepted August 3, 2016. Date of publication August 16, 2016; date of current version October 26, 2016. This work was supported by the Center of Identification Technology Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastien Marcel. *(Corresponding author: Guodong Guo.)*

Y. Zhu and G. Guo are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: yzhu4@mix.wvu.edu; guodong.guo@mail.wvu.edu).

Y. Li and S. Shan are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China (e-mail: yan.li@vipl.ict.ac.cn; sgshan@ict.ac.cn).

G. Mu is with the School of Science, Hebei University of Technology, Tianjin 300401, China (e-mail: guowang.mu@mail.wvu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2016.2601060

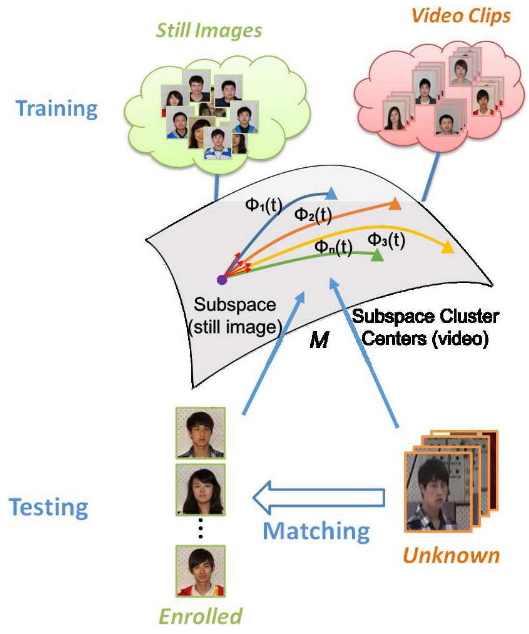


Fig. 1. Illustrate the system of still to video face matching using multiple geodesic flows.

- We propose to bridge the gaps between still images and video frames on the Grassmann manifold by generating multiple "bridges" to characterize the complexity in video frames especially, reducing the gaps between them more effectively.
- We conduct extensive experimental evaluations for S2V face recognition, including both face identification and verification on two large databases. Significant improvements are achieved over the state-of-the-art methods.

The remaining of the paper is organized as follows: A brief review of related works is given in Section II. The proposed heterogeneous matching approach based on the Grassmann manifold is presented in Sections III and IV. Experiments on two large databases are given in Section V. Finally, we draw some conclusions in Section VI.

## II. Related Work

In some earlier works, e.g., [15], [16] a probabilistic model was used to consider each video as a time series state space model, and temporal information is fused to characterize the identity. The still face images in gallery are all of low quality or directly extracted from the videos. However in real world S2V face recognition scenarios, the still images are often captured with high quality under controlled environment, while the videos are usually with low quality or resolutions.

Most recent works on S2V face recognition consider a video clip as a set of individual frames [6], [8]–[10], [17], [18]. Thus the traditional methods for still image-based face recognition can be applied naturally to the S2V scenario. Consequently, the similarity between still images and video clips can be measured based on individual comparisons between each video frame and the still images.

In [6], the authors argued that the S2V face recognition has not been well studied yet, and mentioned several key issues in S2V, such as different resolutions between still images and video frames. A large database is provided for S2V face recognition, with a benchmark of various methods on their S2V database.

To deal with the issue that faces in video are usually of low quality, a quality alignment method for S2V was proposed in [17]. The basic idea is to select the frames of "best quality" from videos assuming these frames can match the quality of still images. They have designed a method that can jointly perform quality, and geometric alignment for face recognition. The experimental results showed that their method obtained a better accuracy than previous works.

In [9], a discriminant analysis based method was proposed to handle the differences between still images and videos. Different from traditional discriminant analysis methods, the authors proposed to exploit separate transformation for still images and video frames, and then pursue a common discriminant space where samples from different scenarios (i.e., sill images and videos) have a good clustering property. They also extended the method to a weighted form in [10] and the S2V matching performance is further improved.

In [19], a Multidimensional Scaling (MDS) based method was proposed for the S2V face recognition. Their method focuses on the issue that low-resolution (LR) probe images and high resolution (HR) gallery images are quite different. They proposed to simultaneously transform the features from the probe and the gallery images to a common space. Then the distances between probe and gallery images can be approximated in the same condition as the gallery images. It is based on MDS and the side information from HR gallery images are used for transformation learning. The SIFT feature is used and the experimental results show a good performance.

Metric learning is another popular methodology for face matching. Different variations have been developed, such as Neighborhood Components Analysis (NCA) [20], Information Theoretic Metric Learning (ITML) [21], Local Fisher Discriminant Analysis (LFDA) [22] and Large Margin Nearest Neighbor (LMNN) [23]. Typically, metric learning intends to learn a transformation (metric) from one Euclidean space to another, by pulling the samples with the same label as close as possible, while pushing the ones with different labels as far as possible. More recently, point-to-set based metric learning methods, such as Point-to-Set Distance Metric Learning (PSDML) [24], locality repulsion projections and sparse reconstruction-based similarity (LRP-SRSM) [25] and Learning Euclidean-to-Riemannian Metric (LERM) [7], have been proposed and applied to the S2V face recognition problem successfully. In these methods, the metric is learned between single samples and the set models, so that the one (still image) vs. multiple (video clip) recognition can be conducted using the minimum point-to-set measurement. These methods are regarded as a more appropriate way for the S2V face recognition task, and the obtained performance [7] exhibits improvement over traditional metric learning based methods.

However, neither the straightforward strategy which treats video as separated frames, nor the point-to-set metric learning methods, have considered the fact that the two modalities are quite heterogeneous. In S2V, the enrolled still images are usually of high quality, while videos are often with low quality.

On the other hand, several methods have been proposed for heterogeneous face recognition, e.g., [26]–[28] for face photo and sketch matching, [29], [30] for recognition between NIR and visible. In [31]–[34] heterogeneous face recognition between visible and beyond-visible spectra have been studied. However, little attention has been paid to the S2V as a heterogeneous matching problem.

Inspired by the recent success on manifold based learning methods for object recognition, we propose an approach to model the transitions between still images and video clips on Grassmann manifold.

In object recognition across domains, it has been successful to learn the domain transitions on the Grassmann manifold [35], where some intermediate representations are derived between two different domains. By sampling the subspaces gradually from one domain to the other, and combine all subspace projections to form a common feature representations. Experimental results on objection recognition have shown the effectiveness of learning on manifold [35]. It can also be extended to a fine-grained manner [36]. That is, a number of generative subspaces are derived by reassembling the data using various proportions of samples from the source and target domains. Instead of using only one path between source and target domain, geodesic paths are computed between those fine-grained subspaces and the connected representations. Inspired by [35], the Geodesic Flow Kernel (GFK) is proposed in [14] to sample an infinite number of subspaces on geodesic flows for multi-view object recognition. It integrates the projections on all subspaces along the geodesic flow.

Different from object recognition [14], [35], [36], we focus on S2V face recognition, where a very limited number of still images are provided in the gallery, e.g., only one photo per subject, while there are a number of video frames in the probe set. To mitigate the unbalance issue, we propose to utilize the geodesic flows on Grassmann manifold, to build "bridges" between still images and video modalities. We construct multiple geodesic flows between the two modalities, to model the complexity in videos frames and improve the S2V face recognition performance. In order to obtain multiple subspace representations for the video frames, we cluster the subspaces for videos on manifold such that the multiple cluster centers on manifold can characterize the complex distributions of the video data. The learned multiple centers force to construct multiple geodesic flows for S2V face recognition.

## III. Heterogeneous Matching Based on Grassmann Manifold

In still-to-video face recognition, for each subject, usually there are a very limited number of still facial images, e.g., 1 with high resolution, frontal view and neutral expression. We denote $S = \{s_1, s_2, \cdots, s_N\}$ with $N$ subjects enrolled in the gallery. There are video clips of low resolutions, and low quality in the probe set, denoted as $V = \{v_1, v_2, \cdots, v_M\}$. The task of still-to-video face identification is formulated

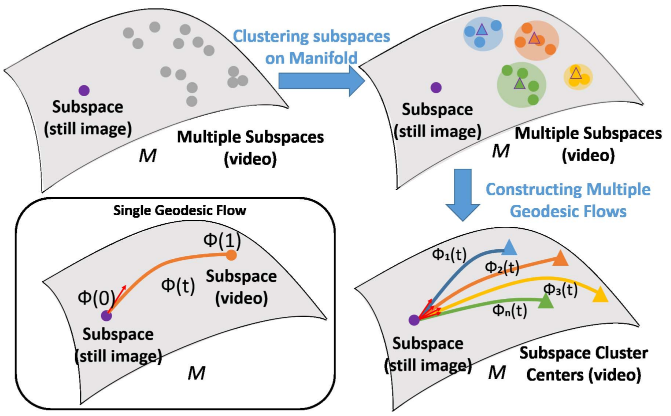


Fig. 2. Schematic illustration of the proposed approach. The bottom left shows the basic idea of using geodesic flow for S2V face recognition. In the training stage, still images and video clips are represented as two points on the Grassmann manifold to learn the "transitions" between them. During testing, the still images and video clips are projected on to a sequence of continuous subspaces before matching. Our approach utilizes multiple geodesic flows that are generated by subspace clustering on the manifold in order to bridge the gaps between still images and video frames.

as follows: given the gallery of still images, for each video clip in the probe set, find the matching $c$ of subject $v_k$, i.e.,

$$c = \arg min_i \, d(s_i, v_k), \tag{1}$$

where $d(s_i, v_k)$ is the distance between still images in gallery and video clips in probe set.

We consider the still-to-video face recognition as a heterogeneous matching problem [12]. We learn the relationship between still images and video clips, which have large differences in quality, i.e., resolution, illumination, pose, expression, and motion blur. By modeling the relationship between them on manifold, still images and videos can be mapped to a new spaces. Intuitively, the proposed approach constructs a "path" between the two modalities, which is learned by exploiting the geometry of the subspaces on Grassmann manifold. The potential intermediate subspaces on this "path" are used to project data from different modalities. Fig. 2 (bottom left) gives the schematic illustration of the proposed approach.

Specifically, let matrix $X$ be the $N$ data samples from the set of still images, where $X = x_i{}_{i=1}^{N}$, $x_i \in \mathbb{R}^D$. $Y = y_i{}_{i=1}^{M}$ denotes the data from the videos,where $y_i \in \mathbb{R}^D$ is the image frame from the video clip. Statistically, the data can be embedded in low-dimensional linear subspaces, where the set of all low-dimensional linear subspaces is termed as the Grassmann manifold [13], denoted by $\mathcal{G}(d, D)$, where $d$ is the dimension of the subspaces. The Principal component analysis (PCA) can be used to generate the subspaces while preserving the data characteristics. Intuitively, by applying PCA on the still images and video clips, respectively, the generated subspaces $S$ and $V$ can be viewed as two points lying on a Grassmann manifold. In our approach, the geometry properties that defined on this Grassmann manifold are used to model the relationship between the still images and video clips, of which the subspaces are two points on the manifold. The minimum length curve connecting these two points on the Grassmann manifold is termed as geodesics. Geodesics can be locally interpreted as curves of the shortest length between subspaces, the transitions from one subspace to the other can be computed, so that the gaps between still images and the video clips can be bridged smoothly. Intermediate subspaces can then be sampled from the geodesic path. The key idea is to utilize the geodesic path between two points on Grassmann manifold, and then utilize the conveyed intermediate subspaces to learn the feature representation considering both still images and video clips.

Formally, let $P_{\text{img}} \in \mathbb{R}^{D\times d}$ be the set of subspaces for the data from still images, and $P_{\text{vid}} \in \mathbb{R}^{D\times d}$ the set of subspaces for the data from video clips. The geodesic flow parameterizes how one modality smoothly change to the other modality.(more mathematical details see [13]). Thus the geodesic flow $\Phi$ is parameterized as $\Phi(i), i \in (0, 1)$, such that $\Phi(0) = P_{\text{img}}$ and $\Phi(1) = \tilde{P_{\text{img}}}$, to compute $\Phi(i)$ [14]:

$$\Phi(i) = P_{\text{img}} U_1 \Gamma_i - \tilde{P_{\text{img}}} U_2 \Sigma_i, \tag{2}$$

where $i$ is the parameter, $\tilde{S_I}$ is defined as the orthogonal complement of $P_{\text{img}}$, i.e., $\tilde{P_{\text{img}}}^T P_{\text{img}} = 0$. $U_1 \in \mathbb{R}^{d\times d}$and $U_2 \in \mathbb{R}^{(D-d)\times d}$ are orthogonal matrices which are given by the following Singular Value Decomposition (SVD):

$$P_{\text{img}}^T P_{\text{vid}} = U_1 \Gamma V^T, \quad \tilde{P_{\text{img}}}^T P_{\text{vid}} = -U_2 \Sigma V^T. \tag{3}$$

$\Gamma$ and $\Sigma$ are $d \times d$ diagonal matrix, in which the diagonal elements are sine and cosine values of the principal angles [13] between $P_{img}$ and $P_{vid}$. More properties about geodesics and Grassmann manifold can be found in [37].

Given the training still image and video data, after this process the parameterized geodesic flow is computed and can characterize the smooth changes from still images to video clips. A series of subspaces $\Phi(t)$, $t \in (0, 1)$ is obtained between these two modalities. Intuitively, if $i$ is close to 0, the subspace is closer to the still images, while if $i$ is closer to 1, the subspace $\Phi(i)$ is more similar to video clips. By projecting a feature vector $x$ onto the intermediate subspaces $\Phi(i)$, using different values of $i$, data from either still images or videos can be transformed into a new representation, which is hopefully not sensitive to the variations between still images and videos.

## IV. Multiple Geodesic Flows on Grassmann Manifold

Based on the basic formulations in the previous section, we propose a novel method called Multiple Geodesic Flows (MGF). In our method, multiple geodesic flows are constructed by using subspaces on manifold. Multiple cluster centers of subspaces on manifold are utilized to handle the complicated distributions of video frames for S2V face recognition. In the following, we present the general description of MGF, and computational details of the method. Then the details on how to apply MGF to S2V face recognition are illustrated.

### *A. Representing Still-to-Video Transitions Using Multiple Geodesic Flows*

In still-to-video face recognition, often there are a small number of still images provided in the gallery [6], while the videos usually consists of a number of frames in the probe set. For example, in one of the popular S2V dataset COX-S2V [6], there are only 1,000 still face images, but about 120,000 video frames collected. In order to model the relationship between still images and videos, as mentioned in Section III, intermediate subspaces between the two points on Grassmann manifold are generated. It is based on the two "end" points lying on the Grassmann manifold, which are derived from the still images and videos. When only one subspace is used for all videos, it could raise two issues: (1) The large number of video frames usually contain very large variations, i.e., different resolutions, pose, expression, and motions blurs, only one linear subspace might not be sufficient to represent the whole distribution of videos. (2) The subspaces from still image $P_{\text{img}} \in \mathbb{R}^{D\times d}$ and video $P_{\text{vid}} \in \mathbb{R}^{D\times d}$ lie on the same Grassmann manifold $\mathcal{G}(d, D)$, thus the dimensionality of the subspaces is the same. However, as mentioned above, the different distributions of controlled still images vs. uncontrolled videos indicate that the dimension of subspace is small for still images, while large for video clips. To handle these issues, we propose to construct multiple geodesic flows between the still images and video frames, using Fig. 2 to illustrate this idea.

Inspired by the bagging [38] scheme, one can generate multiple subspaces from the video data in a similar manner. The basic idea of bagging is to sample the training data multiple times, so that classifiers are trained on those multiple sets, and then the final prediction is formed by aggregating the individual predictions. In our MGF, one may use the subsets which are randomly selected from the original data each time, to generate multiple subspaces on the manifold.

Intuitively, there will be a number of points lying on Grassmann manifold for the video data. Multiple geodesic flows can then be computed connecting the single point of still images to every points from the videos. However, using those randomly generated subspaces not only brings large burden of computational cost, but also may produce unstable performance [38]. To avoid this problem, we propose to cluster the set of subspaces to find the representative ones, in order to construct multiple geodesic flows between still images and videos.

### *B. Clustering Subspaces on Grassmann Manifold*

We cluster the subspaces on Grassmann manifold using the K-means algorithm. By doing so, a large set of randomly generated subspaces on Grassmann manifold can be grouped into a few cluster centers, of which the center subspaces are more representative and robust, indicating the intrinsic structure of the video data distribution.

Unlike K-means clustering in Euclidean space, clustering subspaces on the Grassmann manifold relies on the distance between two points on the manifold. And also, the means or averages of subspaces should be computed other than the Euclidean means of feature points.

The distance between two points on a Grassmann manifold can be measured by the length of its geodesic path, which is defined as the shortest path between the two points.

**Algorithm 1** Algorithm to Compute Karcher Mean

```
Input: A set of $P$ points on manifold $\{X_i\}_{i=1}^{P} \in \mathcal{G}(d, D)$
Output: Karcher mean $\mu_K$
1 Set an initial estimate of Karcher mean $\mu_K = X_i$ by
  randomly picking one point in $\{X_i\}_{i=1}^{P}$
2 Compute the average tangent vector
  $A = \frac{1}{P} \Sigma_{i=1}^{P} \log_{\mu_K}(X_i)$
3 If $||A|| < \epsilon$, then return $\mu_K$, stop. Else, go to Step 4
4 Move $\mu_K$in average tangent direction $\mu_K = exp_{\mu_K}(\alpha A)$,
  where $\alpha > 0$ is a parameter of step size. Go to Step 2,
  until $\mu_K$ meets the termination conditions (reaching the
  max iterations, or other convergence conditions)
```

The geodesic distance, in other words, the length of the minimal curve connecting two points, is based on the intrinsic geometry of Grassmann manifold, which can be expressed with principle angles, a.k.a. canonical angles.

Formally, the principle angles, $\Theta(P_{\text{img}}, P_{\text{vid}}) = [\theta_1, \cdots, \theta_d]$, where $\theta_k \in [0, \pi/2$ between two $d$ dimensional subspace $P_{\text{img}} \in \mathbb{R}^{D \times d}$ and $P_{\text{vid}} \in \mathbb{R}^{D \times d}$ are defined by:

$$\cos(\theta_k) = s_k^T v_k = \max_{s \in P_{\text{img}}} \max_{v \in P_{\text{vid}}} s^T v \tag{4}$$

subject to $||s|| = ||v|| = 1$, $s^T s_i = 0$, $v^T v_i = 0$, $i = 1, \cdots, k-1$. $|| \cdot ||$ denotes the standard Euclidean norm.

Alternatively, based on the singular value decomposition (SVD), the cosines of the principle angles can be computed as the singular values of $P_{\text{img}}^T P_{\text{vid}}$:

$$P_{\text{img}}^T P_{\text{vid}} = U(\cos(\Theta))V^T, \tag{5}$$

where $U = [u_1 \cdots u_d]$, $V = [v_1 \cdots v_d]$. And $\cos(\Theta)$ is the diagonal matrix defined by:

$$\cos(\Theta) = diag(\cos\theta_1, \cdots, \cos\theta_d) \tag{6}$$

The geodesic distance, which can be derived from principle angles, is given by:

$$d\left(P_{\text{img}}, P_{\text{vid}}\right) = ||\Theta||_2, \tag{7}$$

where $|| \cdot ||$ indicates $l_2$-norm.

Once we have the distance measure between subspaces on Grassmann manifold, the next step is to define the subspace averages. One of the methods for computing averages of subspaces is Karcher mean [39], a.k.a. the Riemannian center of mass. More details of various subspace means can be found in [40]. Particularly, Karcher mean is a kind of geometric mean of several matrices, which minimizes the sum of squared geodesic distances on manifolds. In our work, we apply subspace K-means clustering based on the Karcher mean on Grassmann manifolds.

The procedure for computing the Karcher mean is given by Algorithm 1. Given the geodesic distance between the subspaces on manifold, and the Karcher mean of the subspaces, the set of subspaces on Grassmann manifold can be partitioned into $K$ clusters in which each subspace belongs to a cluster center with the nearest mean. This can be implemented by either K-means clustering or hierarchical clustering. The cluster centers are considered more representative for the video frames, comparing to randomly selected subspaces on the Grassmann manifold. Moreover, the cluster centers should be dissimilar to each other. As a result, the set of cluster centers are more robust to represent the complex and large number of video frames.

### *C. Multiple Geodesic Flows for S2V Face Recognition*

By clustering the subspaces, multiple geodesic flows can be computed (as described in Section III) through a *one-to-many* connections, i.e., one subspace for still images, while multiple subspaces for videos. For each one of the geodesic flows, still images and videos can be mapped onto a common set of subspaces (according to $\phi(t)$), so that the differences between them can be reduced. Then, matching can be applied based on each geodesic flow.

Now we present the framework of our MGF based S2V face recognition. Particularly, in the training phase, given the still images that form the gallery set $\mathcal{S}$, PCA is applied on $\mathcal{S}$ to generate the subspace $S_I \in \mathbb{R}^{D \times d}.S_I$ can be viewed as one point on the Grassmann Manifold $\mathcal{G}(d, D)$. On the other hand, for video data $\mathcal{V}$, the first step is to generate a collection of subspaces $\left[S_V^1, \cdots S_V^m\right]$, $S_V^i \in \mathbb{R}^{D \times d}$, $i = 1 \ldots m$, using the scheme described in Section IV-A. The subspaces is considered to be lying on the same Grassmann manifold $\mathcal{G}(d, D)$. In the next step, subspace K-means clustering on Grassmann manifold is applied (as shown in Section IV-B) on the set of subspaces generated from the video frames. The derived cluster centers are denoted as $\left[\bar{S}_V^1, \cdots \bar{S}_V^n\right]$, $\bar{S}_V^i \in \mathbb{R}^{D \times d}$, $i = 1 \ldots n$. For each center subspace $\bar{S}_V^i$, $i = 1 \ldots n$, a geodesic flow between $S_I$ and $\bar{S}_V^i$ is computed as described in Section III. The collection of geodesic flows is denoted as $[\Phi_1(t), \cdots, \Phi_n(t)]$, $t \in (0, 1)$. Next, for each geodesic flow $\Phi_i(t)$,a number of intermediate subspaces are generated according to different values of $t$, e.g., [0, 1] with interval 0.1. Given a still image or a video frame, MGF feature representation is obtained by concatenating all feature vectors that are projected on the intermediate subspaces of $\Phi_i$, $i = 1 \ldots n$.

### *D. Face Matching Scheme*

In testing, face matching is conducted according to each geodesic flow, respectively. Each time, when matching a still image from the gallery with a probe video clip, MGF features are computed by projecting the still image or video frames on each geodesic flow $\Phi_i$, $i = 1 \ldots n$. LDA (Linear Discriminant Analysis) is then applied to increase the discriminative power of the feature representations. The obtained features along each geodesic flow can be used to compute the consine distance for measuring the similarity. For face identification, a majority voting scheme over the frames is used to determine the final identification result. For face verification, the matching score is obtained by averaging the frame to image similarity score for each video frame-image pair on all the geodesic flows, according to 1. Fig. 1 shows a brief illustration of the matching scheme.

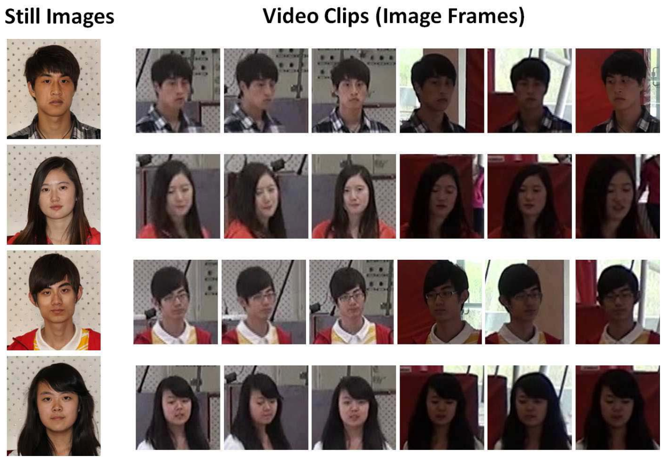


Fig. 3. Example still images (first column) and video frames from the COX-S2V dataset, where still images are with high quality, while video clips have relatively low qualities.

## V. Experiments

In this section, we conduct face recognition experiments on two large databases. To validate the proposed method, both face identification and verification are presented. Firstly, we introduce the two still-to-video face databases. Then the experimental settings are described. Finally, the experimental results are presented with comparisons to the state-of-the-art approaches.

### A. Databases

*1) COX-S2V Dataset:* COX-S2V dataset [6] consists of both still images that were collected by a high resolution camera with cooperative users under controlled conditions, and uncontrolled video clips collected via video cameras. In total, there are 1,000 subjects in this dataset. For each subject, there is one high resolution still image, and three video clips, denoted as CAM1, CAM2 and CAM3, respectively, corresponding to three different installation locations. Faces in video clips contains large differences in illumination condition, also different head poses, and motion blurs are observed in this database. Some example images and video frames of the COX-S2V dataset are shown in Fig. 3.

*2) PaSC Dataset:* PaSC dataset [8] was collected for the point and shoot face recognition challenge, which aims to recognize facial images from inexpensive “point and shoot” cameras. This dataset includes 9,376 still images and 2,802 video clips, collected from 293 subjects using different sensors. Faces in this dataset also have different variations such as head pose, background locations, motion blur and poor focus, for both still images and video clips. Fig. 4 shows some face examples images from the PaSC dataset.

### B. Experimental Settings

To validate the performance of the proposed approach for S2V face recognition, experiments are conducted on both COX-S2V dataset and PaSC dataset. According to the previous works on the two datasets [7], [8], face identification is performed on COX-S2V, and face verification on PaSC.

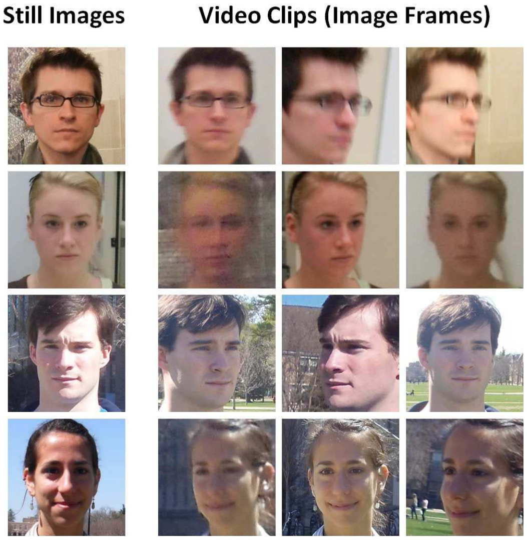


Fig. 4. Example face images from the PaSC dataset. Face photos show many variations in video frames.

On COX-S2V database, we follow the same experimental protocol as in [7]. The video clips are split according camera settings, i.e., CAM1, CAM2, and CAM3, respectively. In training, there are 300 still images and 300 corresponding video clips from each camera setting. In testing, there are 700 still images served as the gallery and 700 video clips from each camera setting as the probe set. There is no subject overlap between the training and test sets. The experiments ran 10 times with randomly generated gallery/probe combinations. The recognition rates are used as the performance measure on this dataset.

On the PaSC dataset, we apply the same protocol for face verification experiments according to [8]. In the training phase, we only use the data that is provided by PaSC dataset without using any external data.Specifically, 2872 still images of 484 subjects and 280 video clips of 170 subjects for training, 4688 still images of 293 subjects (target set) and 1401 handheld video clips of 265 subjects (query set) are used for testing. The verification is conducted based on the similarity matrix with size 4688 × 1401. ROC curves are used to show the performances.

In order to extract feature representations from face images on the two databases, face detection [41] and alignment (according to eye locations) are applied first to obtain face regions for both still images and videos. The cropped face data are then resized to 60 × 48 (which is the same as [7]) and 80 × 60 (heuristically)for COX-S2V and PaSC, respectively. Histogram equalization is then applied to each face image. Next, PCA is used to reduce the dimensionality. The projection on PCA basis are served as the face feature representation. We empirically chose 600 as the reduced dimension for COX-S2V. On PaSC, the reduced dimension is set to about 800 in our experiments. After applying the proposed method, LDA is used on MGF feature to increase the discriminative power. Therefore the dimensionality of facial features is $nClass - 1$, where $nClass$ is the number of classes. The resulting features are fed into the classification for COX-S2V,

TABLE I

FACE IDENTIFICATION ACCURACIES ON COX-S2V USING SINGLE AND MULTIPLE GEODESIC FLOWS. SGF INDICATES SINGLE GEODESIC FLOW. "K" INDICATES THE NUMBER OF SUBSPACES THAT OBTAINED USING SUBSPACE CLUSTERING ON GRASSMANN MANIFOLD

| Method | CAM1 | CAM2 | CAM3 |
|---|---|---|---|
| SGF | $49.01 \pm 1.12$ | $41.76 \pm 1.66$ | $69.28 \pm 0.91$ |
| MGF (K=3) | $49.64 \pm 1.34$ | $43.13 \pm 1.27$ | $70.31 \pm 1.39$ |
| MGF (K=5) | $51.76 \pm 1.80$ | $43.37 \pm 0.86$ | $71.84 \pm 1.32$ |
| MGF (K=7) | $52.82 \pm 1.04$ | $43.24 \pm 1.67$ | $70.90 \pm 1.56$ |
| MGF (K=9) | $49.96 \pm 0.94$ | $42.94 \pm 1.61$ | $71.37 \pm 1.31$ |

TABLE II

RECOGNITION ACCURACIES USING MULTIPLE GEODESIC FLOWS, COMPARING TO THE RESULTS USING EACH SINGLE CLUSTER CENTER

| | Accuracy (%) | | | | |
|---|---|---|---|---|---|
| **CAM1** | **K = 5** | | | | |
| Individual | 51.28 | 52.79 | 50.56 | 50.44 | 49.91 |
| Fusion | **51.76** | | | | |
| **CAM2** | **K = 5** | | | | |
| Individual | 41.37 | 41.52 | 42.01 | 41.58 | 42.04 |
| Fusion | **43.37** | | | | |
| **CAM3** | **K = 5** | | | | |
| Individual | 66.47 | 66.46 | 66.43 | 66.47 | 66.94 |
| Fusion | **71.84** | | | | |

and verification for PaSC, according to the experimental settings described above. The running time of the MFG is approximately 7.5 seconds for training, and the time for testing is less than 0.5 second. All the experiments are performed on the 64 bit Windows 7 platform with Intel Core i7 3.4GHz CPU, and 12GB memory.

### *C. Experimental Results on COX-S2V Dataset*

Face identification experiments are conducted on COX-S2V dataset. Firstly, we test the performance of using single geodesic flow (single GF) for S2V face recognition on COX-S2V. In this experiment, we compute the subspace using all the training still images. Then all the video clips in the training set are used to compute a single subspace, as the representation of the video frames, which together with the image subspace can be viewed as two points on the Grassmann manifold. Next, the geodesic flow is calculated using the approach described in Section III, to obtain the "transition" between still images and video frames. By projecting and concatenating the features on the geodesic flow, the newly achieved features from both still images and videos are fed into PCA for dimension reduction and then LDA for discriminative mapping. Finally still-to-video face classification is applied using the same settings as in [7]. The classification accuracies (average classification accuracies and the standard deviations) are shown in Table I (1st Row). The accuracies of CAM1, CAM2 and CAM3 are $49.01\% \pm 1.12\%$, $41.76\% \pm 1.66\%$, and $69.28\% \pm 0.91\%$, respectively. From the results one can see that, the performance of CAM3 is higher than that in CAM1 and CAM2 settings, the reason might be that the video quality for CAM3 is better than the other two camera settings, in terms of the head pose, occlusion, and motion blur for the captured face frames.

We also conducted experiments using the proposed MGF method. At first, one subspace is calculated using all the training still images. According to the proposed method (Section IV), multiple subspaces are generated from the training videos. Then subspace clustering is applied to achieve a certain number of subspace centers. Then, multiple geodesic flows are computed between still image subspace and all the video subspaces in a one-to-many manner. Finally, matching is applied on each geodesic flow (same as single GF case), the final recognition rate or accuracy is achieved by a score level fusion of each individual geodesic flow.

In Table I, we also report the recognition rates using different number of cluster centers. From the table one can see that: (1) For all three camera settings, the recognition accuracies are improved using the proposed method. The highest accuracy obtained for CAM1 is $52.82\% \pm 1.04\%$, when K is set to 7 for subspace clustering, while on CAM2 and CAM3, the best recognition accuracies are $43.37\% \pm 0.86\%$ and $71.84\% \pm 1.32\%$, respectively, when K is set to 5. These results demonstrates the effectiveness of the proposed method. Using multiple geodesic flows can model the relationship between still images and video frames better, and therefore improve the recognition rates. (2) Different numbers of cluster centers can have different effects on the recognition accuracy. In our experiments, better performances are obtained using 5 or 7 clusters centers. Also from a practical perspective, a larger number of cluster centers is not recommended, since it will bring high computational cost on both training and testing phases.

Then we have designed experiments to test the recognition accuracy on each single geodesic flow based on subspace cluster centers. The results are shown in Table II. From the table one can see that, for each subspace cluster center the accuracy is comparable to the case where single GF is applied for all the three cameras. This shows that individual subspace generated by subspace clustering is capable of representing the faces to a certain degree and the performance is higher (CAM2 and CAM3) or comparable to the single GF. On the other hand, the overall accuracy can be significantly improved by using multiple geodesic flows based on subspace cluster centers. This indicates that using multiple geodesic flows can have significant improvements on S2V face recognition over the single geodesic flow. This also illustrates that each individual subspace cluster centers can model different aspects between face images and video frames, and the multiple centers are truly needed.

Finally, we compare the recognition rates with the state-of-the-art methods on COX-S2V dataset shown in Table III. Specifically, we show the baseline algorithm, i.e., Nearest-Neighbor Classifier (NNC) [42] along with the state-of-the-art metric learning methods NCA [20], LMNN [23], and the state-of-the-art point-to-set method LERM [7]. From the table one can see that, in all three video camera settings, our proposed method achieves high accuracies than all other methods listed in the table. Note that for a fair comparison, the feature representation are kept the same (pixel values with histogram

TABLE III

EXPERIMENTAL RESULTS OF STILL-TO-VIDEO FACE RECOGNITION: RECOGNITION RATE (%) ON COX-S2V DATASET

| Method | COX-S2V | | |
|---|---|---|---|
| | Still-Video1 | Still-Video2 | Still-Video3 |
| NNC [42] | $9.96 \pm 0.61$ | $7.14 \pm 0.68$ | $17.37 \pm 6.16$ |
| NCA [20] | $39.14 \pm 1.33$ | $31.57 \pm 1.56$ | $57.57 \pm 2.03$ |
| LMNN [23] | $34.44 \pm 1.02$ | $30.03 \pm 1.36$ | $58.06 \pm 1.35$ |
| LERM [7] | $45.71 \pm 2.05$ | $42.80 \pm 1.86$ | $58.37 \pm 3.31$ |
| Single GF [35] | $42.00 \pm 1.38$ | $31.43 \pm 0.96$ | $60.29 \pm 1.28$ |
| Single GF [35] + LDA | $49.01 \pm 1.12$ | $41.76 \pm 1.66$ | $69.28 \pm 0.91$ |
| GFK + LDA [12] | $48.96 \pm 1.22$ | $42.99 \pm 2.17$ | $69.81 \pm 1.72$ |
| Ours | $\mathbf{52.80 \pm 0.71}$ | $\mathbf{43.37 \pm 0.86}$ | $\mathbf{71.84 \pm 1.32}$ |

TABLE IV

EXPERIMENTAL RESULTS OF STILL-TO-VIDEO FACE VERIFICATION ON PaSC. VERIFICATION RATE IS COMPUTED WHEN THE FAR EQUALS 0.01

| Method | Verification Rate |
|---|---|
| Geodesic Flow (Single) | 0.24 |
| MGF (K=3) | 0.25 |
| MGF (K=5) | **0.26** |
| MGF (K=7) | 0.26 |
| MGF (K=9) | 0.26 |

TABLE V

THE EXPERIMENTAL RESULTS (VERIFICATION RATE) FOR STILL-TO-VIDEO FACE VERIFICATION ON PaSC, WHEN THE FAR EQUALS 0.01

| **Method** | **Verification Rate** |
|---|---|
| NNC [42] | 0.07 |
| NCA [20] | 0.17 |
| LMNN [23] | 0.17 |
| LERM [7] | 0.16 |
| GFK [12] | 0.22 |
| Ours | **0.262** |

equalization and PCA for dimensionality reduction) for all the listed methods. The comparisons demonstrate that our MGF method is very appropriate for still-to-video face matching. The recognition accuracies are better than the state-of-the-art methods.

### *D. Experimental Results on PaSC Dataset*

We also conduct experiments for S2V face verification. The same experimental settings are adopted according to [8] on PaSC dataset. Firstly, we report experimental results using MGF method with different numbers of cluster centers, shown in Table IV. When a single geodesic flow is applied, the verification rate is 0.237, which illustrates that geodesic flow between still images and video frames can be built for face matching. When using the proposed MGF method, the verification rate can be improved further. The best verification rate is 0.262, obtained when the number of cluster centers is set to 5.

Next we compared the proposed method with the state-of-the-art methods. Those methods, i.e., NCA, LMNN, and LERM, have shown good performance on the COX-S2V dataset in previous works [7]. Note that all the listed methods are using the same gray level features as mentioned in Section V-B. The verification rates are shown in Table V. From the results one can see that the performance of our proposed method achieves 0.262 (the Verification Rate (VR) when False Acceptance Rate (FAR) is 0.01), which is significantly better than the listed methods. This further demonstrates the robustness and effectiveness of our method for still to video face recognition.

TABLE VI

COMPARISON WITH THE PARTICIPANT METHODS [43] AND A COMMERCIAL SYSTEM (REPORTED FROM [8]) ON THE PaSC COMPETITION

| **Method** | **Verification Rate** |
|---|---|
| PittPatt (Commercial) [8] | 0.42 |
| PLDA-WPCA-LLR | 0.26 |
| Eigen-PEP | 0.24 |
| LBP-SIFT-WPCA-SILD | 0.23 |
| ISV-GMM | 0.11 |
| LRPCA | 0.10 |
| Ours | 0.26 |

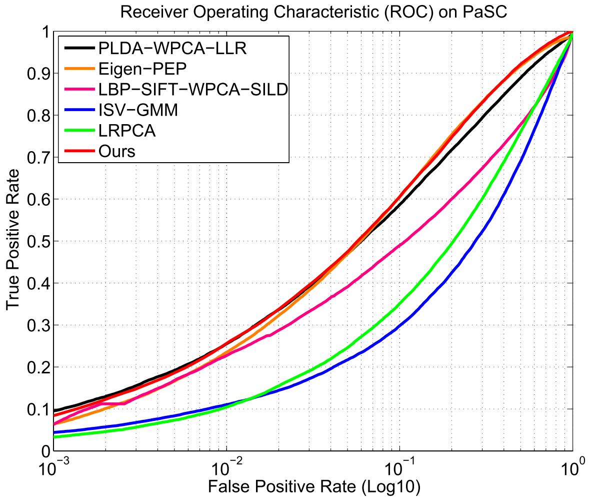


Fig. 5. ROC curve on PaSC for S2V face verification.

Since the PaSC dataset has also served for a face recognition competition, we compare the proposed method with the results of the competition participants. The verification rates (when the FAR equals 0.01) are shown in Table VI. From the table one can see that our method achieve the accuracy of 0.26, which is considerably better than LRPCA (0.10), ISV-GMM (0.11), Eigen-PEP (0.24), LPB-SIFT-WPCA-SILD (0.23) methods, and comparable to the PLDA-WPCA-LLR. Note that the PittPatt result reported in [8] is a commercial system, was trained on external data, and they had much more bandwidth to refine their systems. Although the performance of our method is similar to other competing results, ours has some advantages: (1) The feature representation in our method is the gray level values, while other methods applied more advanced and complex features, e.g., SIFT, Gabor, 2D-DCT, and LPQ features. (2) Only the provided training data on PaSC are used for learning in our methods, while the above listed methods used external data in the training process. Following the competition protocol, the ROC curves are shown in Fig. 5, where our method is significantly higher than many others, and is comparable to the best competition result.

## VI. Discussion and Conclusions

We have presented a new approach to still-to-video face recognition, where each subject is enrolled with a limited number of images, while a large amount of video clips are used to define a probe set. We interpret the S2V face recognition as a heterogeneous face matching problem, and build the relationship between the unbalanced distributions of still images and video clips of different quality. The proposed approach aims at learning multiple "transitions" from still images to video frames. A subspace clustering algorithm has been applied to generate multiple representative subspaces for video clips, so that multiple geodesic flows can be constructed. Extensive experiments on two large databases have shown significant improvements over the state-of-the-art methods.

## Acknowledgment

Co-authors Guowang Mu and Yan Li worked on the project during their visit to Prof. Guo's Computer Vision Lab at WVU.

## References

[1] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden Markov models," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2003, pp. 340–345.

[2] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[3] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2567–2573.

[4] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2496–2503.

[5] Z. Cui, H. Chang, S. Shan, B. Ma, and X. Chen, "Joint sparse representation for video-based face recognition," *Neurocomputing*, vol. 135, no. 5, pp. 306–312, Jul. 2014.

[6] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset," in *Proc. Asian Conf. Comput. Vis.*, Daejeon, Korea, 2013, pp. 589–600.

[7] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1677–1684.

[8] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2013, pp. 1–8.

[9] X. Chen, C. Wang, B. Xiao, and X. Cai, "Scenario oriented discriminant analysis for still-to-video face recognition," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 738–742.

[10] X. Chen, C. Wang, B. Xiao, and C. Zhang, "Still-to-video face recognition via weighted scenario oriented discriminant analysis," in *Proc. IEEE Int. Joint Conf. Biometrics*, Oct. 2014, pp. 1–6.

[11] H. Wang, C. Liu, and X. Ding, "Still-to-video face recognition in unconstrained environments," *Proc. SPIE*, vol. 9405, p. 94050O, Feb. 2015.

[12] Y. Zhu, Z. Zheng, Y. Li, G. Mu, S. Shan, and G. Guo, "Still to video face recognition using a heterogeneous matching approach," in *Proc. IEEE Biometrics, Theory, Appl. Syst.*, Sep. 2015, pp. 1–6.

[13] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[14] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2012, pp. 2066–2073.

[15] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 214–245, Jul./Aug. 2003.

[16] S. K. Zhou and R. Chellappa, "Beyond one still image: Face recognition from multiple still images or a video sequence," in *Face Processing: Advanced Modeling Methods*. 2005, pp. 547–567.

[17] Z. Huang, X. Zhao, S. Shan, R. Wang, and X. Chen, "Coupling alignments with recognition for still-to-video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3296–3303.

[18] Y. Li, R. Wang, Z. Huang, S. Shan, and X. Chen, "Face video retrieval with image query via hashing across Euclidean space and Riemannian manifold," in *Proc. CVPR*, Jun. 2015, pp. 4758–4767.

[19] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3037–3049, Dec. 2013.

[20] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 513–520.

[21] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[22] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[23] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1473–1480.

[24] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2664–2671.

[25] J. Lu, Y.-P. Tan, G. Wang, and G. Yang, "Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 1070–1080, Jun. 2013.

[26] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 513–520.

[27] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Proc. ECCV*, 2010, pp. 420–433.

[28] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[29] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 1513–1516.

[30] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1123–1128.

[31] D. Lin and X. Tang, "Inter-modality face recognition," in *Computer Vision*. Graz, Austria: Springer, 2006, pp. 13–26.

[32] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[33] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1707–1716, Dec. 2012.

[34] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[35] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 999–1006.

[36] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, Nov. 2014.

[37] K. A. Gallivan, A. Srivastava, X. Liu, and P. V. Dooren, "Efficient algorithms for inferences on Grassmann manifolds," in *Proc. IEEE Workshop Statist. Signal Process.*, Oct. 2003, pp. 315–318.

[38] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

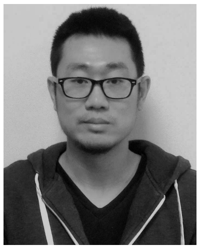
[39] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Commun. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–541, Sep. 1977.

[40] T. Marrinan, B. Draper, J. R. Beveridge, M. Kirby, and C. Peterson, "Finding the subspace mean or median to fit your need," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1082–1089.

[41] *Face++ API*, accessed on Dec. 1, 2014. [Online]. Available: http://www.faceplusplus.com

[42] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[43] J. R. Beveridge *et al.*, “The IJCB 2014 PaSC video face and person recognition competition,” in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–8.

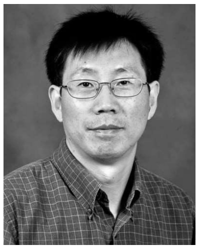


**Yu Zhu** received the B.E. degree in software engineering from Northwestern Polytechnical University in 2010, and the Ph.D. degree in computer science from the Lane Department of Computer Science and Electrical Engineering, West Virginia University, in 2016. His research interests lie in computer vision and pattern recognition, in particular, in the area of action recognition, face recognition, age estimation, and biometrics.

**Yan Li** received the B.S. degree in computer science and technology from Nankai University, Tianjin, China, in 2010. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, pattern recognition, image processing, and, in particular, face recognition and binary code learning.

**Guowang Mu** received the B.S. and M.S. degrees in computational mathematics from Jilin University, Changchun, China, and the Ph.D. degree from Beihang University, Beijing, China. He is currently a Professor with the Hebei University of Technology, Tianjin, China. He was a Visiting Scholar with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, in 2015, and as a Postdoc with the Department of Computer Science, North Carolina Central University from 2008 to 2010. His research interests include computer vision, image processing, and computer-aided design.

**Shiguang Shan**, photograph and biography not available at the time of publication.



**Guodong Guo** (M’07–SM’07) received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, China, and the Ph.D. degree in computer science from the University of Wisconsin-Madison, Madison, WI, USA. He visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; Microsoft Research, Beijing, China; and North Carolina Central University. He is an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, WV, USA. He authored a book, *Face, Expression, and Iris Recognition Using Learning-Based Approaches* (2008), co-edited a book, *Support Vector Machines Applications* (2014), and published about 100 technical papers. His research interests include computer vision, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, Outstanding Researcher (2013–2014) at CEMR, WVU, and New Researcher of the Year (2010–2011) at CEMR, WVU. He was selected the “People’s Hero of the Week” by BSJB under the Minority Media and Telecommunications Council in 2013. Two of his papers were selected as “The Best of FG’13” and “The Best of FG’15,” respectively.