Still to Video Face Recognition Using a Heterogeneous Matching Approach

Yu Zhu[1], Zhenzhu Zheng[1], Yan Li[1,2], Guowang Mu[1], Shiguang Shan[2], and Guodong Guo[1]

[1]Lane Department of CSEE, West Virginia University, Morgantown WV 26506, USA,
[2]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
yzhu4@mix.wvu.edu, zezheng@mix.wvu.edu, yan.li@vipl.ict.ac.cn, gumu@mail.wvu.edu, sgshan@ict.ac.cn, Guodong.Guo@mail.wvu.edu (corresponding author)

Abstract

In this paper, we address the problem of still-to-video (S2V) face recognition. Still images usually have high qualities, captured from cooperative users under controlled environment, such as the mugshot photos. On the contrary, video clips may be acquired with low resolutions and low qualities, from non-cooperative users under uncontrolled environment. Because of these significant differences, we consider the S2V as a heterogeneous matching problem, and propose to develop a method to bridge the gap between the two heterogeneous modalities. A Grassmann manifold learning method is developed to construct subspaces for the purpose of bridging the gap between the two face modalities smoothly. We conduct extensive experiments 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.

1. Introduction

Face recognition is one of the most important topics in biometric and computer vision. In addition to the traditional 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 [23, 15, 2, 27, 6], where faces captured by video cameras usually contain more variations caused by illumination, head pose, expression, and those due to the motion blur. More recently, the problem of Still-to-Video (S2V) face recognition has attracted attentions [13, 14, 1, 3, 4, 26, 21], since it has a wide range of usage for many practical applications, such as identifying the criminal suspect with his video clip among a huge mug-shot still image database, and rapid locating and tracking target subject with his one still image among the whole city security surveillance video data. However, the great disparity brought by still images and video clips poses a huge challenge to the S2V face recognition task. In this paper, we mainly focus on the S2V face recognition problem that each subject is enrolled with only few high resolution still images, while the query is represented as a video clip which contains a sequence of uncontrolled image frames

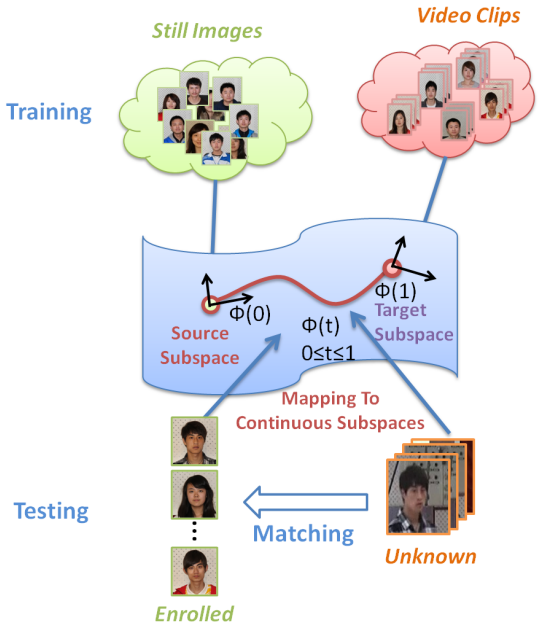

Figure 1. Illustration of the proposed method. In the training stage, still images and video clips are represented as two points on the Grassmann manifold to learn the "transitions" between them. In testing, the still images and video clips are projected on to a sequence of continuous subspaces before matching between them.

with complex variations.

One of the methodologies of solving the S2V face recognition considers video clip as a set of individual frames [13] [1], and methods on still image face recognition can be naturally applied to the S2V scenario. In this way, similarity between still image and video clip can be expressed by the calculations over the individual comparisons between frames.

Metric learning is another recent popular methodology and manifests in various forms, such as Neighborhood Components Analysis (NCA) [10], Information Theoretic Metric Learning (ITML) [7], Local Fisher Discriminant Analysis (LFDA) [25] and Large Margin Nearest Neighbor (LMNN) [28]. More specifically, metric learning usually intends to learn a transformation (metric) from one Euclidean space to another, by pulling the samples with the same label as closer 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) [31] and Learning Euclidean-to-Riemannian Metric (LERM) [14], have been proposed and successfully applied to the S2V face recognition problem. In these methods, 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 [14] exhibits the improvement over traditional metric learning methods.

However, neither the former straightforward strategy which treats video as separated frames nor the point-to-set metric learning methods, considers the fact that the two modalities are quite heterogeneous, especially in the S2V scenario that enrolled still images are always of high quality yet videos often suffer from relatively low quality. In previous approaches different heterogeneous face problems have been studied, e.g., in [30] [29] [18] for face photo and sketch problem, [16] [19] for face recognition in NIR and visible. And in [22] [17] and [20] heterogeneous face recognition between visible and beyond-visible domains have been studied. More recently, a review of heterogeneous face recognition approaches is conducted in [12]. However, little attention has been devoted to the S2V as a heterogeneous matching problem, attempting to reduce the differences between still images and video clips.

Different from the metric learning perspective, we mainly focus on generating the connections between heterogeneous modalities, i.e., still images and video clips, in a natural unsupervised manner. Specifically, we attempt to explore the transition (connection) from one modality to the other, so that the relationship between them can be modeled and used for the subsequent modality representation and distance matching. Particularly, to make full use of the recent advantages in subspace learning, we approach the S2V face recognition by exploring the connection of heterogeneous subspaces lying on Grassmann manifold [8], thus characterizing the transition from still images to video clips, while reducing the differences between them. Inspired by the recent success of Geodesic Flow Kernel (GFK) [11] on object recognition, which is one type of Grassmann manifold learning methods, we explore the performance for S2V face recognition. We think that the Grassmann manifold based learning could be an appropriate approach to build the relationship between still face images and video clips.

The rest of the paper is organized as following: we introduce the proposed heterogeneous approach based on Grassmann manifold in Section 2. We conduct experiments on two databases and the results are shown in Section 3. Finally, conclusions are drawn in Section 4.

2. Heterogeneous Approach based on Geodesic Flow Kernel

In the problem of still-to-video face recognition, for each subject, generally there are very limited number (usually only 1) of high resolution still images $X = \{x_1, x_2, \cdots, x_N\}$ which are enrolled as the gallery set. And there are low resolution video clips $Y = \{y_1, y_2, \cdots, y_M\}$ forming the probe set. The task of still-to-video face identification is formulated as: given the gallery set of still images, for each video clip in the probe set, find the matching label (subject identity) of y_k, which can be inferred as:

$$c = \arg \mathit{min}_i d(x_i, y_k), \tag{1}$$

where $d(x_i, y_k)$ is the distance between still image from gallery and video clip from probe.

In this paper, we consider the Still-to-Video face recognition problem as a heterogeneous matching problem. The approach that we propose aims to explore the relationship between still image and video clips, which have vast differences in quality. By modeling the relationship between those two data fields using Grassmann manifold [8], video clips and still images can be mapped to common subspaces for the matching task. Intuitively, we have the data from both still images and the video clips, the proposed approach constructs a "path" between the two modalities, which is learned by exploiting the geometry of their subspaces on Grassmann manifold. Figure 1 shows the schematic illustration of the proposed method.

Specifically, we denote matrix X as 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, those data usually can be embedded into low-dimensional linear subspaces, where the set of all low-dimensional linear

subspaces is termed the Grassmann manifold [8], denoted by $\mathcal{G}(d, D)$, where d is the dimension of the subspaces. Principal component analysis (PCA) is one of the methods to generate the subspaces while preserving the data characteristics. Intuitively, the subspaces S_I and S_V which are generated from the PCA on the still images and video clips respectively, can be viewed as two points on a Grassmann manifold. In our approach, the geometry properties that defined on this Grassmann manifold is used to model the relationship between the still images and the video clips, of which the subspaces are two points on the Grassmann manifold. The minimum length curve connecting two points on the manifold is termed as geodesics. Inspired by the fact that geodesics can be locally interpreted as curves of the shortest length between subspaces, we want to find the transition from one subspace to the other, so that the still images and the video clips can be bridged smoothly. The key idea is to utilize the geodesic path between two points on Grassmann manifold, and then utilize the intermediate subspaces to learn the feature from the one modality, i.e., still images to the other modality i.e., video clips.

Formally, let $S_I \in \mathbb{R}^{D\times d}$ denote the set of subspaces for the data from still images, and $S_V \in \mathbb{R}^{D\times d}$ the set of subspaces for the data from video clips. The geodesic flow Φ is constructed through the canonical metric on the Grassmann manifold, which is induced by Frobenius norm on the tangent space [8]. Thus the geodesic flow Φ is parameterized as $\Phi(i), i \in (0, 1)$, such that $\Phi(0) = S_I$ and $\Phi(1) = \tilde{S}_I$, to compute $\Phi(i)$ [11]:

$$\Phi(i) = S_I U_1 \Gamma_i - \tilde{S}_I U_2 \Sigma_i, \tag{2}$$

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

$$S_I^T S_V = U_1 \Gamma V^T, S_I^{\tilde{T}} S_V = -U_2 \Sigma V^T. \tag{3}$$

Γ and Σ are $d \times d$ diagonal matrix, in which the diagonal elements are sine and cosine value of the principal angles [8] between S_I and S_V. More mathematical details about geodesics and Grassmann manifold can be found in references [9].

After the learning process mentioned above, the parameterized geodesic flow 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 domains. Intuitively if i is close to 0, the subspace is more likely to the still images, while if i is close to 1, the subspace $\Phi(i)$ is more like coming from the video clip. By projecting a feature vector x or y onto the subspace $\Phi(i)$, the data from either still image or videos can be transited into a new representation, which is hopefully insensitive to the varieties between still images and videos and therefore can be used for further matching tasks.

Motivated by [11], all of the subspaces along the geodesic path is utilized for the projection so that the projected features are robust to the variations that leans to the two different modalities. Specifically, for two original feature vectors x_i and x_j, we need to compute their projected feature vector x_i' and x_j' on $\Phi(t)$ continuously from 0 to 1. The geodesic-flow kernel (GFK) [11] $G \in \mathbb{R}^{D\times D}$ is then defined as the inner product between them:

$$\int_0^1 \left(\Phi(t)^T x_i\right)^T \left(\Phi(t)^T x_j\right) dt = x_i^T G x_j. \tag{4}$$

The matrix G is computed by using the matrices in the above:

$$\begin{aligned} G &= \int_0^1 \left(\Phi(t)^T\right)^T \left(\Phi(t)^T\right) dt \\ &= \begin{bmatrix} S_I U_1 & \tilde{S}_I U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2 & \Lambda_3 \end{bmatrix} \begin{bmatrix} U_1^T S_I^T \\ U_2^T \tilde{S}_I^T \end{bmatrix} \end{aligned} \tag{5}$$

The elements of diagonal matrices Λ_1, Λ_2 and Λ_3 are:

$$\begin{aligned} &\lambda_{1i} = \frac{2\theta_i + \sin(2\theta_i)}{2\theta_i}, \quad \lambda_{2i} = \frac{\cos(2\theta_i) - 2\theta_i}{2\theta_i}, \\ &\lambda_{3i} = \frac{2\theta_i - \sin(2\theta_i)}{2\theta_i}. \end{aligned} \tag{6}$$

In this way, still images and video clips can be mapped onto common set of subspaces so that the varieties between them can be reduced. It has utilized the geodesic flow kernel since it is based on the integral of all the subspaces on geodesic path other than discrete selection of subspaces, and the only free parameter is the dimension of subspaces. To increase the discriminative power after using GFK, LDA [24] is then applied to the features for the recognition task.

3. Experiment

In this section, we first give a brief introduction of the databases used in the experiment. Next we perform an extensive evaluation on two practical classification tasks, i.e., S2V face identification on COX-S2V [13] dataset, and S2V face verification on PaSC [1] dataset. Results are shown and compared with other competing approaches on these two databases, to demonstrate the capability of the proposed approach.

3.1. Databases

3.1.1 COX-S2V Dataset

COX-S2V dataset [13] consists of both still images that collected by SLR camera with cooperative user under controlled conditions, and uncontrolled video clips collected

via video cameras. Totally 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 huge appearance variations, such as illumination, head pose, and motion blurs. Some example images and video frames of COX-S2V dataset are shown in Figure 2.

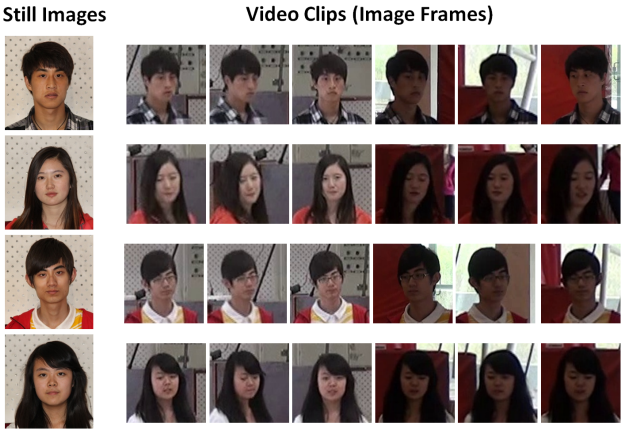

Figure 2. Example still images (first column) and video frames of COX-S2V dataset, where still images are always with high quality and video clips has relatively low quality.

3.1.2 Point and Shoot Face Recognition Challenge Dataset (PaSC)

PaSC dataset [1] is 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. Figure 3 shows some face example images from the PaSC dataset.

3.2. 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 databases, face identification is performed on COX-S2V dataset, and face verification is conducted on the PaSC dataset.

Following the original protocol on COX-S2V dataset [14], 300 subjects' still images and the corresponding three video clips of each subject are used for training, and the remaining 700 subjects' still images and the video clips are used for testing. In the testing phase, the set of 700 still images server as the gallery set, where the corresponding video clips (i.e., Cam1-3) form the probe set.

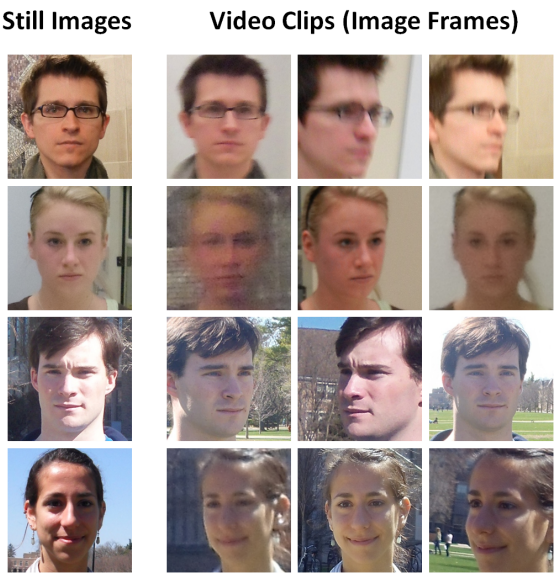

Figure 3. Example face images from the PaSC dataset. Faces show great varieties in video clips.

The experiments are run 10 times with randomly collected gallery/probe combinations and the recognition rates are used as the performance measurement on this dataset.

On the PaSC dataset, according to [1], we apply the same protocol for face verification experiments. In the training phase, we only use the data that provided by PaSC dataset. Specifically, 2872 still images of 484 subjects and 280 video clips of 170 subjects for training, and 4688 still images of 293 subjects (target set) and 1401 handheld video clips of 265 subjects (query set) for testing. The comparison is conducted based on the computed similarity matrix with size 4688×1401.

In order to extract face representations from face images on these databases, face detection and alignment is firstly applied to get the cropped face images for both still images and video clips. All the face data are then resized to 60×48 for both COX-S2V and PaSC. Histogram equalization is then applied to data matrix which column data represents the gray values of each image. Finally PCA is used to reduce the dimensionality. The resulting vectors obtained are served as the face feature representations. We empirically choose 600 as the reduced dimension for COX-S2V in the proposed approach, and on PaSC dataset the reduced dimension is set to about 250. For the proposed approach, subspaces are generated using PCA with the PCA ratio of 0.95.

3.3. Experimental Results

Table 1 shows the experimental results of face recognition rate on COX-S2V dataset. In this table we also show the baseline algorithm, which is Nearest Neighbor Classifier (NNC) [5], along with the state-of-the-art metric learn-

ing methods NCA [10], LMNN [28] and the state-of-the-art point-to-set method LERM [14]. From the table one can see that the best result is achieved by the proposed approach, the recognition performance is significantly improved and the results on all the three video sets are better than other listed methods in Table 1. Note that for fair comparison, the feature representation of the face images and videos are kept the same (histogram equalization and PCA for dimensionality reduction) for all the methods. This experimental results shows the capability of matching faces from still image and video clips by learning the relationship between the two modalities, and recognition performance is better than the other state-of-the-art methods.

Table 1. The experimental results of Still-to-Video face recognition rate (%) on COX-S2V dataset.

Method	COX-S2V		
	Still-Video1	Still-Video2	Still-Video3
NNC	9.96 ± 0.61	7.14 ± 0.68	17.37 ± 6.16
NCA	39.14 ± 1.33	31.57 ± 1.56	57.57 ± 2.03
LMNN	34.44 ± 1.02	30.03 ± 1.36	58.06 ± 1.35
LERM	45.71 ± 2.05	42.80 ± 1.86	58.37 ± 3.31
Ours	$\mathbf{48.96 \pm 1.22}$	$\mathbf{42.99 \pm 2.17}$	$\mathbf{69.81 \pm 1.72}$

Next we conduct experiments for the other scenario: S2V face verification on PaSC dataset. The same experimental settings according to [1] are adopted and the results are shown in Table 2. Firstly we compare our results with the state-of-the-art metric learning methods that has been shown on the COX-S2V dataset. From the results one can see that the performance of the proposed method outperforms the other methods, i.e., NNC, NCA, LMNN, and LERM. Note that the results using those methods are based on same gray features as mentioned in Section 3.2. We also compared the computational time that used for training using different methods and the results are shown in Table 3. From the table we can see that our proposed method takes least time, comparing with other methods, which further illustrate the efficiency and effectiveness of the proposed method.

Table 2. The experimental results (verification rate) for Still-to-Video face verification on PaSC dataset, when FAR equals to 0.01.

Method	**Verification Rate**
NNC [5]	0.05
NCA [10]	0.16
LMNN [28]	0.17
LERM [14]	0.17
Ours	**0.22**

Since PaSC dataset is also served as a face recognition competition database, we now compare our proposed method with the results of the competition participants. The verification rate when FAR equals to 0.01 are shown in Table 4. From the table one can see that our proposed method achieved the performance 0.22, which is better than LRPCA (0.10) and ISV-GMM (0.11) methods. Comparing with PLDA-WPCA-LLR, Eigen-PEP, and LPB-SIFT-WPCA-SILD, although the performance of our method is slightly lower than these methods, our proposed approach has several advantages: (1) The feature representation in our method is the gray level feature vector, while the above three methods used more advanced features, e.g., SIFT, Gabor, 2D-DCT, and LPQ features. (2) Only the provided training data on PaSC are used for training in our methods, while the above three methods used external data for the training process.

Table 3. The computational time that used for training in our experiments on COX-S2V and PaSC datasets.

Method	COX-S2V	PaSC
NCA [10]	11 hours	6 hours
LMNN [28]	372 sec	172 sec
LERM [14]	73 sec	52 sec
Ours	33 sec	28 sec

Table 4. The verification rate for Still-to-Video face verification compared with the PaSC face recognition challenge, when FAR equals to 0.01.

Method	**Verification Rate**
PLDA-WPCA-LLR [1]	0.26
Eigen-PEP [1]	0.24
LPB-SIFT-WPCA-SILD [1]	0.23
ISV-GMM [1]	0.11
LRPCA [1]	0.10
Our Method	0.22

4. Conclusions

In this paper, a heterogeneous matching approach is proposed to deal with the still-to-video face recognition problem. Different from other works e.g., metric learning and point-to-set schemes, which learns the distance metrics between image-to-video, the proposed approach aims at learning the "transitions" that can smoothly connect between the two modalities, i.e., still image and video clips. This approach is based on utilizing the subspaces between two different modalities on Grassmann manifold, and projecting the data from different domains onto common subspaces. A kernel function is applied so that all the continuous subspaces on geodesic flow can be used to model the relationship between still images and video clips. Experimental results on two large databases demonstrate that the proposed approach performs better than the state-of-the-art methods.

Acknowledgments

This paper was partially supported by a NSF CITeR grant.

References

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

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

[3] X. Chen, C. Wang, B. Xiao, and X. Cai. Scenario oriented discriminant analysis for still-to-video face recognition. In *ICIP*, pages 738–742. IEEE, 2014.

[4] X. Chen, C. Wang, B. Xiao, and C. Zhang. Still-to-video face recognition via weighted scenario oriented discriminant analysis. In *IJCB*, pages 1–6. IEEE, 2014.

[5] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[6] Z. Cui, H. Chang, S. Shan, B. Ma, and X. Chen. Joint sparse representation for video-based face recognition. *Neurocomputing*, 135:306–312, 2014.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[8] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[9] K. A. Gallivan, A. Srivastava, X. Liu, and P. Van Dooren. Efficient algorithms for inferences on grassmann manifolds. In *Workshop on Statistical Signal Processing*, pages 315–318. IEEE, 2003.

[10] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, pages 513–520, 2005.

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE CVPR*, pages 2066–2073. IEEE, 2012.

[12] G. Guo. Heterogeneous face recognition: An emerging topic in biometrics. *Intel® Technology Journal*, 18(4), 2014.

[13] 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 *ACCV*, pages 589–600. Springer, 2013.

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

[15] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[16] B. Klare and A. K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *ICPR*, pages 1513–1516. IEEE, 2010.

[17] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410–1422, 2013.

[18] B. F. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[19] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*, pages 1123–1128. IEEE, 2009.

[20] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li. Coupled discriminant analysis for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 7(6):1707–1716, 2012.

[21] 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 *IEEE CVPR*, pages 4758–4767, 2015.

[22] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, pages 13–26. Springer, 2006.

[23] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *CVPR*, volume 1, pages I–340. IEEE, 2003.

[24] B. Scholkopft and K.-R. Mullert. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1:1, 1999.

[25] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

[26] H. Wang, C. Liu, and X. Ding. Still-to-video face recognition in unconstrained environments. In *IS&T/SPIE Electronic Imaging*, pages 94050O–94050O. International Society for Optics and Photonics, 2015.

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

[28] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[29] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, pages 420–433. Springer, 2010.

[30] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520. IEEE, 2011.

[31] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: Extend the learning of distance metrics. In *IEEE ICCV*, pages 2664–2671, 2013.