IEICE TRANS. INF. & SYST., VOL.E97–D, NO.10 OCTOBER 2014

PAPER

Point-Manifold Discriminant Analysis for Still-to-Video Face Recognition

Xue CHEN[†a)], ***Nonmember***, **Chunheng WANG**[†], ***Member***, **Baihua XIAO**[†], ***Nonmember***, ***and*** **Yunxue SHAO**[†], ***Member***

SUMMARY In Still-to-Video (S2V) face recognition, only a few high resolution images are registered for each subject, while the probe is video clips of complex variations. As faces present distinct characteristics under different scenarios, recognition in the original space is obviously inefficient. Thus, in this paper, we propose a novel discriminant analysis method to learn separate mappings for different scenario patterns (still, video), and further pursue a common discriminant space based on these mappings. Concretely, by modeling each video as a manifold and each image as point data, we form the scenario-oriented mapping learning as a Point-Manifold Discriminant Analysis (PMDA) framework. The learning objective is formulated by incorporating the intra-class compactness and inter-class separability for good discrimination. Experiments on the COX-S2V dataset demonstrate the effectiveness of the proposed method.

key words: *face recognition, still-to-video, discriminant analysis, point-manifold distance, scenario-oriented*

1. Introduction

Face recognition from a single still image has been extensively studied for over a decade. Recently, the usage of cameras has contributed to a rapid increase in the availability of video resources. Typically, vast amounts of videos are continuously acquired to monitor government compounds, military installations, commercial sites, and private premises. As a result, video-based face recognition (VFR) applications have become an emerging topic. Based on the type of the gallery set, we can classify VFR into two categories: Video-to-Video (V2V) face recognition and Still-to-Video (S2V) face recognition [1].

In the V2V scenario, video resources are available for both the gallery and probe set. Similar to image(s)-to-image(s) recognition, most existing approaches extract the same type of features for the two sets and then perform recognition by comparing them directly. Wang et al. [2] used a second-order statistic, covariance matrix, to model the video sequences, and exploited Log-Euclidean Distance for explicitly mapping and recognition. Li et al. [3] modeled face dynamics using identity surfaces, and performed recognition by matching the face trajectories constructed on the identity surfaces. Moreover, hidden Markov models have also been applied to model the video information by learning the statistics dynamics over time in each video [4].

Manuscript received February 21, 2014.
Manuscript revised June 6, 2014.

[†]The authors are with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
a) E-mail: xue.chen@ia.ac.cn

DOI: 10.1587/transinf.2014EDP7057

Face recognition from still images is another important category for VFR, which is named as S2V face recognition. Typical application of S2V is the mug-shot matching which includes the recognition of faces in drivers licenses, passports, credit cards etc. In this scenario, very few (single, in many cases) still images per person are enrolled in the gallery while multiple video clips are captured as the probe, as shown in Fig. 1. Generally, the enrolled set are captured under controlled conditions and are of high quality. However, the probe videos, captured on arbitrary locations, are of low resolution and even exhibit considerable blur. S2V face recognition poses a huge challenge due to the great discrepancies of imaging conditions, pose and facial expression between the still scene and video scene. The difference brought by these factors could lead to faces of a certain person even lying in different subspaces, making the S2V recognition very challenging.

Traditionally, the S2V scenario has been formulated in frameworks of the subspace methods [5], [6]. In the subspace framework, a video is represented as a subspace, and a canonical angle between the subspace and a still image is computed as a matching score. Although classic subspace-based methods could obtain representative face features, performance degenerates severely when wide differences exist in the intra-class samples. A natural way to deal with this problem is to learn a common mapping space for the polymorphous samples. Typically, Huang et al. [7] proposed an improved LDA [8] to learn projections by using partial weighting to emphasize cross-scenario images in the discriminant analysis. One shortcoming of this method is its reliance on a single mapping to build a common dis-

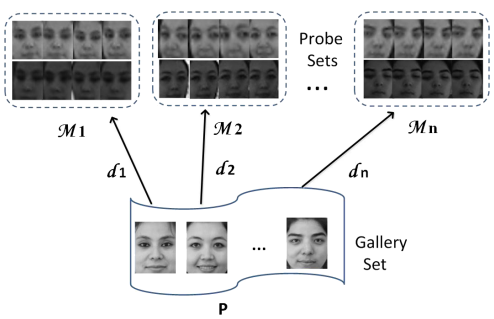


Fig. 1 Face recognition on the S2V scenario. Each subject is enrolled with a single still image in the gallery, and the probe covers multiple video clips under different conditions.

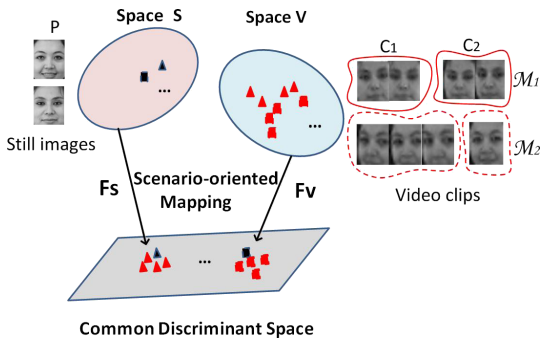


Fig. 2 An overview of the PMDA approach. Manifold $\mathcal{M}_i$ on each video clip V_i is modeled as a set of local models C_i in space V, and still images are denoted as points data P in space S. $\{F_s, F_v\}$ are mapping functions for space S and V. Shapes represent classes, and colors present scenarios.

criminant space for samples in different scenarios. As data present distinctly different characteristics towards specific scenario, it is obviously inefficient to model them in a uniform mode.

In this paper, we propose a novel discriminant analysis method by exploiting separate mappings for different scenario patterns (still, video). Intuitively, scenario-oriented mappings could model the underlying manifold in corresponding modality more effectively. An overview of the proposed method is shown in Fig. 2. By modeling each video clip V_i as a manifold $\mathcal{M}_i$ and still image as data point P, we form the S2V mapping learning as a Point-Manifold Discriminant Analysis (PMDA) framework. Concretely, the manifold is modeled as a collection of local consistent models, e.g., $\{C_1, C_2\}$. The learning objective is formulated by compelling the local models towards images of the same identity, but far from those of distinct identities. Finally, the scenario-oriented mappings $\{F_s, F_v\}$ pursue a common discriminant space where samples from different scenarios have good clustering property. Performance on the COX-S2V dataset [7] demonstrates a remarkable improvement over previous methods.

The rest of this paper is organized as follows. Section 2 details the proposed method. Experiments and results are presented in Sect. 3, while our conclusions are drawn in Sect. 4.

2. Point-Manifold Discriminant Analysis

In this section, we first give a primary formulation of S2V face recognition problem. Then, we describe the PMDA algorithm and how it leads to scenario-oriented discriminative mappings for effective face recognition. At last, face recognition is performed by matching the closest part of two modes in the learned embedding space.

2.1 Problem Formulation

In S2V face recognition scenario, there is generally only one high resolution still image enrolled for the gallery while a set of low resolution video clips are available for probing. Formally, let $S = \{s_1, s_2, \ldots s_{N_S}\}$ be the gallery set containing N_S still images, where s_i is the still image of the i^{th} person. Correspondingly, assume $V = \{V_1, V_2, \ldots, V_{N_V}\}$ be the query set containing N_V video clips, where $V_j = \{v_{j,1}, v_{j,2}, \ldots, v_{j,N_{V_j}}\}$ denotes the j^{th} query video and N_{V_j} is the number of video frames in V_j. In this way, the identity recognition of video V_j from the gallery S performs the following algorithm $\mathcal{A}$:

$$\mathcal{A} : \hat{i} = \arg \min_{i=1,2,\ldots,N_s} d(s_i, V_j), \tag{1}$$

where $d(,)$ defines the distance between image s_i and video clip V_j. Obviously, an effective image-video distance measure is the critical link for S2V recognition.

Typically, the underlying structure of high-dimensional observation data, whose variations are controlled by only a few factors, can be modeled by a low-dimensional manifold [9], [10]. In this paper, we just consider a relative ideal condition, where the facial appearance variations are just caused by common factors of pose, expression or lighting changes. In this situation, it is natural to assume that sequential frames of video V_j, which generally contain relative simple condition changes, distribute on a simplified nonlinear manifold $\mathcal{M}_j$. Moreover, as local linearity property holds everywhere on a global nonlinear manifold, it is rational to model the manifold as a collection of approximate linear local subspaces [10], [11], e.g., $\mathcal{M}_j = \{C_{j,1}, C_{j,2}, \ldots, C_{j,N_{M_j}}\}$. Here, N_{M_j} denotes the number of local models, which is far smaller than the sample number N_{V_j} on manifold $\mathcal{M}_j$, in most cases. Thus, the image-video distance $d(s_i, V_j)$ modeled by the Point-Manifold distance $d(P_i, M_j)$ can be further converted to a concrete form:

$$d(s_i, V_j)_{s_i \to P_i, V_j \to \mathcal{M}_j} = d(P_i, \mathcal{M}_j), \tag{2}$$

where manifold $\mathcal{M}_j$ models the video clip V_j, and point P_i denotes the still image s_i.

With the statement above, the PMDA approach transforms to learning a discriminating embedding space, which can better distinguish the Point-Manifold data clusters with different class labels and enhance the within-class local compactness. Specially, as the within-class data in still scene S and video scene V present distinctly different characteristics, it is very necessary to apply separate mappings for point data in S and manifold data in V respectively, in order to model the data distribution in corresponding modality effectively. In this way, face recognition in the scenario-oriented embedding space can be expressed as:

$$\mathcal{A} : \hat{i} = \arg \min_{i=1,2,\ldots,N_s} d(F_s(P_i), F_v(\mathcal{M}_j))_{s_i \to P_i, V_j \to \mathcal{M}_j}, \tag{3}$$

where F_s and F_v are the mapping functions for scene S and V respectively.

2.2 Constructing Local Consistent Models

The idea of extracting local models from manifold has been

exploited in several methods. They often tend to adopt well-designed clustering algorithms along with complicated constraints to guarantee explicit linearity for local models [11], [12]. Particularly, in a strict linear space, data in a cluster distributes on a plane in the feature space. However, in our algorithm, we just need a set of submodels to model the local statistic characteristics on a nonlinear manifold effectively. For example, for a video sequence covering continuous head pose changes, each of the local models is trained to represent the image frames set with a certain pose. Generally, minor range of deviation is allowed among the internal images. In this setup, we don't need samples in a local model distribute strictly on a plane. Instead, those around the plane are also considered. In this way, the strict linearity is not exactly required for our clustering algorithm. Therefore, in this paper, we adopt the linearity criterion as relaxed constraints for constructing local models .

The basic idea of our linearity constrained clustering (LCC) algorithm is that, in the first level, all samples are initiated as a singleton cluster. Then, in each new level, two seed points are selected by the furthest geodesic distance, and two new local clusters expand from them respectively according to relaxed linearity constraints. Finally, we are able to obtain a series of local models hierarchically, associated with different local characteristics. Following the notations of Sect. 2.1, we give a detailed implementation of the LCC algorithm. For a video clip $V_j = \{v_{j,1}, v_{j,2}, \ldots, v_{j,N_{V_j}}\}$ modeled as manifold $\mathcal{M}_j$, we aim to extract a collection of local models, denoted as:

$$\mathcal{M}_j = \{C_{j,1}, C_{j,2}, \ldots, C_{j,N_{M_j}}\}, \tag{4}$$

$$C_{j,l} = \{v^{(l)}_{j,1}, v^{(l)}_{j,2}, \ldots, v^{(l)}_{j,N_{C_{j,l}}}\}, \quad \sum_{l=1}^{N_{M_j}} N_{C_{j,l}} = N_{V_j}, \tag{5}$$

where N_{M_j} is the number of local models, and $N_{C_{j,l}}$ is the number of samples clustered in model $C_{j,l}$.

Firstly, both Euclidean distance matrix $D_E \in \mathbb{R}^{N_{V_j} \times N_{V_j}}$ and geodesic distance matrix $D_G \in \mathbb{R}^{N_{V_j} \times N_{V_j}}$ are computed for all the pair-wise samples $\{v_m, v_n\}$ in V_j. Specially, the geodesic distance is computed as the Euclidean distance sum of a sequence of neighboring points on the data distribution, which form a path between the two points [9]. Then the ratio matrix R is obtained by:

$$R(v_m, v_n) = D_G(v_m, v_n)/D_E(v_m, v_n). \tag{6}$$

Referring to the definition above, the geodesic distance D_G is computed based on a path on the data distribution. Naturally, as the curvature of the distribution is greater, geodesic distance between the two points gets larger. Yet, Euclidean distance D_E just involves computing the linear distance of two points directly. In this way, greater curvature of the distribution leads larger ratio R. Namely, matrix R could well reflect the non-linearity degree of a cluster. Besides, matrix $H \in \mathbb{R}^{k \times N_{V_j}}$ holding the $k-$NNs' indices of all the samples in each column is also constructed on V_j.

To preform LCC, we first select two furthest seed points $\{v_{j,L}, v_{j,R}\}$ from V_j by geodesic distance matrix D_G, and then initialize two new clusters $\{C_{j,L}, C_{j,R}\}$ with them. So, V_j is left exclusive of $\{v_{j,L}, v_{j,R}\}$. Secondly, as neighboring samples have similar statistic property, we collect the K-NN set $\{v_{j,c}\}_{c=1}^{K}$ for each point in $C_{j,L}$ as candidate. Next, we only have to check the candidates by the linearity constraint for constructing new local models, without traversing all the samples in the training set. Concretely, if $v_{j,c}$ satisfies simultaneously:

$$v_{j,c} \in V_j, \tag{7}$$

$$R(v_{j,c}, v_{j,k}) < \theta, \quad \forall\, v_{j,k} \in C_{j,L}, \tag{8}$$

$v_{j,c}$ will be added into $C_{j,L}$. The second step is performed repeatedly for the gradually expanding $C_{j,L}$ until no candidates can be added. Thus, V_j is left exclusive of those points added into $C_{j,L}$. Next, if V_j is not empty, same operations will be executed on the $C_{j,R}$; else, $C_{j,R}$ with only the seed point merges with $C_{j,L}$ as a holistic cluster, and clustering ends. Furthermore, if the clustering doesn't end, we select two furthest seed points from the left V_j by geodesic distance matrix again, and then repeat the series of operations above, until V_j is empty. Clearly, as samples satisfying the linearity constraint lie approximatively on a plane and so have similar statistic property, the local clusters constructed hold good characteristic consistency. Further, they can also be used to model the local statistic characteristics on a nonlinear manifold effectively.

Note that, the threshold parameter θ in Eq. (6) reflects the linear perturbation degree of local models. A small θ implies more local models and better linear preserving in each model, and vice versa. As we just use linearity constraints as an auxiliary for the $k-$NN criterion to improve the local consistency of our clustering algorithm, strict linearity is not exactly required. Therefore, we just use the relaxed linearity constraint ($R < \theta$) to construct local models for applications, where θ is a constant that is a little larger than 1.

Moreover, referring to the implementation of LCC, we can figure out that, compared with Euclidean distance based clustering methods such as k-means [1], [13], the LCC could better guarantee the local consistency in the aspect of region locality and characteristics consistency, and further make more meaningful clusters to model the local statistic property of the manifold, in most cases. An auxiliary illustration is shown in Fig. 3. Samples of local models from k-means may be on different manifold subspaces and present diversiform characteristics, while results from the LCC are more aligned with the fact.

2.3 Learning Scenario-Oriented Discriminant Mappings

Classical LDA supposes data of each class are generated from a single normal distribution and seeks a uniform mapping for all the classes [14]. However, in the S2V scenario, each class contains two types of data, low resolution video frames and high resolution still images. To model the data distribution in corresponding modality effectively, we exploit separate mappings for the still scene and video scene

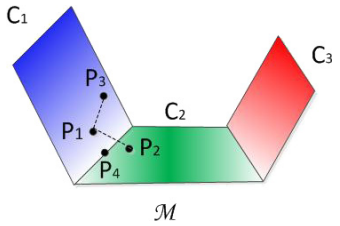


Fig. 3 Constructing local models. Manifold $\mathcal{M}$ covers three local subspaces $\{C_1, C_2, C_3\}$. Data points $\{P_1, P_3, P_4\}$ lie on C_1, and $\{P_2, P_4\}$ on C_2. P_2, P_3 are equidistant from P_1. Generally, P_2, P_3 on different subspaces are clustered as one local model with P_1 by k-means. However, as geodesic distance $d(P_1, P_4, P_2)$ is obviously larger than Euclidean distance $d(P_1, P_2)$, LCC will cluster P_1 into C_1 but P_2 into C_2 respectively.

respectively. Concretely, by modeling each video as a manifold, we formulate the scenario-oriented mapping learning as a Point-Manifold Discriminant Analysis framework. The optimization objective is formulated by incorporating the intra-class compactness and inter-class dispersion for good discrimination.

2.3.1 Separability-Compactness Based Constraints

PMDA shares similar motivation with traditional discriminant learning methods in characterizing the within-class compactness and between-class separability for optimization. Differently, instead of operating directly on samples, this algorithm constructs the constraints upon manifold local models. Based on the introduction of local models above, we formulate the manifold based discriminant analysis involving compactness and separability constraints as follows.

Assume the training set be $T = \{s_i \in S \cup V_i \in V\}, 1 \leq i \leq N_S$, where $S = \{s_i \in \mathbb{R}^{d_S}\}_{i=1}^{N_S}$ is the still images set as denoted in Sect. 2.1, and $V = \{V_1, V_2, \ldots, V_{N_S}\}$ holds the corresponding video clip $V_i = \{v_{i,k} \in \mathbb{R}^{d_V}\}_{k=1}^{N_{V_i}}$ for each person i in S. d_S and d_V are the sample dimensions. Given the manifold $\mathcal{M}_i = \{C_{i,l}\}_{l=1}^{N_{M_i}}$, $C_{i,l} = \{v_{i,m}^{(l)}\}_{m=1}^{N_{C_{i,l}}}$ on video clip V_i and point data P_i on still image s_i, we can simply define the distance between local model $C_{i,l}$ and point P_i by their sample average $\overline{\mathbf{v}}_{i,l}$, $\overline{\mathbf{s}}_i$:

$$d(C_{i,l}, P_i) = \|\overline{\mathbf{v}}_{i,l} - \overline{\mathbf{s}}_i\|, \tag{9}$$

$$\overline{\mathbf{v}}_{i,l} = \frac{1}{N_{C_{i,l}}} \sum_{m=1}^{N_{C_{i,l}}} v_{i,m}^{(l)}, \ \textit{and} \quad \overline{\mathbf{s}}_i = s_i. \tag{10}$$

Referring to the property of linear space, samples can be reconstructed linearly by other samples from the same linear subspace. As some local parts of faces of varying poses or expressions change severely, it's hard to use images of one pose (expression) to reconstruct linearly that of another pose (expression). Yet, face images of similar poses (expressions) have similar appearance. Referring to the principle of image reconstruction, they can be used to reconstruct faces under the same condition linearly. Based on this, we can suppose that face images lying on a linear plane have similar poses (expressions). Namely, the local models obtained by the LCC hold good local characteristic consistency. In this situation, the sample average could capture the data property of the local model effectively. For a special case, face images can be largely changed on a linear space by global illumination changes, but they are also clustered as a local model by the LCC. In this condition, the sample average may not precisely represent the appearance characteristic of the data cluster. However, we think that the less accurate model approximation is just enough for the proposed method. Since faces with severe illumination changes lie on a linear space and could reconstruct each other linearly, the facial texture and structure information of these images are very similar. Just the image intensity changes linearly with a lighting offset. In this situation, the sample average is still able to capture the typical facial texture and structure information in these faces. The following discriminant training then make still images close to video frames, which hold similar facial texture and structure information with the sample average, and the learning results can still obtain good intra-class compactness and inter-class discrimination. Above all, it is suitable to use the sample average to represent the local models form the LCC in the proposed method. Coincidentally, similar model approximation is also adopted in [12].

Next, we denote the transforms for still scene S and video scene V by $F_s(\theta_s) \in \mathbb{R}^{d' \times d_S}$ and $F_v(\theta_v) \in \mathbb{R}^{d' \times d_V}$ respectively, where θ_s and θ_v are the mapping parameters, and d' is the mapping dimension of transform matrixes. Then, the intra-class compactness term J_w and inter-class separability term J_b in the new space are computed as :

$$J_w(\theta_s, \theta_v) = \frac{1}{N_w} \sum_{i=1}^{N_s} \sum_{l=1}^{N_{M_i}} \|F_v \overline{\mathbf{v}}_{i,l} - F_s s_i\|^2, \tag{11}$$

$$J_b(\theta_s, \theta_v) = \frac{1}{N_b} \sum_{i=1}^{N_s} \sum_{j=1,2,..,N_s; j \neq i} \sum_{l=1}^{N_{M_j}} \|F_v \overline{\mathbf{v}}_{j,l} - F_s s_i\|^2, \tag{12}$$

where N_w and N_b are the number of pairs from the same class and different classes respectively. Here, we use the sample centers $\overline{V}_i = \{\overline{\mathbf{v}}_{i,l}\}_{l=1}^{N_{M_i}}$ of the local model set $\{C_{i,l}\}_{l=1}^{N_{M_i}}$ to represent the video clip V_i, and compare them with still images in the transformed space.

A common problem of applying discriminant learning in the S2V scenario is that the number of still images in the training set is much smaller than that of video frames in total, and the discordance would cause serious bias for the following scenario-oriented training. Similar problem is also discussed in [7], [15]. From Eqs. (11), (12), we can see that the local model number is much smaller than that of the total frames in each video clip ($N_{M_j} \ll N_{V_j}$), so our algorithm could effectively alleviate the number discordance between still scene and video scene, by modeling the video frames set as a set of local models.

To better discriminate the samples from different classes, we should compel the video frames towards the still images with the same identity, but far from those of distinct identities. Based on this principle, the objective function of PMDA arrives at the following optimization criterion:

$$\min_{\theta_s,\theta_v} J = J_w(\theta_s,\theta_v) - \alpha * J_b(\theta_s,\theta_v), \tag{13}$$

where α indicates the nonnegative tradeoff parameter.

2.3.2 Solving the Optimization Model

To solve the problem in Eq. (13) with a simply matrix derivation, we reform it in the following way. Let $S = [s_1, s_2, \ldots s_{N_S}] \in \mathbb{R}^{d_S \times N_S}$ collect all the still images for N_S person, and $\overline{V}_i = [\overline{\mathbf{v}}_{i,1}, \overline{\mathbf{v}}_{i,2}, \ldots, \overline{\mathbf{v}}_{i,N_{M_i}}] \in \mathbb{R}^{d_V \times N_{M_i}}$ collect the N_{M_i} local model centers of video V_i for person i, $1 \le i \le N_S$, then $\overline{V} = [\overline{V}_1, \overline{V}_2, \ldots, \overline{V}_{N_S}] \in \mathbb{R}^{d_V \times N_m}$ represent the whole local model matrix for all the N_S person. In addition, we set:

$$\bar{S} = [\bar{s}_1, \bar{s}_2, .., \bar{s}_{N_s}],\ \bar{s}_i = [s_i, .., s_i] \in \mathbb{R}^{d_S \times N_{M_i}}, \tag{14}$$

$$\bar{V}^\dagger = [\bar{V}_1^\dagger, \bar{V}_2^\dagger, .., \bar{V}_{N_S}^\dagger],\ \bar{V}_i^\dagger = \bar{V}/\bar{V}_i \in \mathbb{R}^{d_V \times (N_m - N_{M_i})}, \tag{15}$$

$$S^\dagger = [s_1^\dagger, s_2^\dagger, .., s_{N_S}^\dagger],\ s_i^\dagger = [s_i, .., s_i] \in \mathbb{R}^{d_S \times (N_m - N_{M_i})}, \tag{16}$$

where $\bar{V}/\bar{V}_i$ indicates a residue matrix of removing the i^{th} submatrix $\bar{V}_i$ from matrix $\bar{V}$. Then, we cast the objective function in Eq. (13) into a simplified form:

$$\min_{F_s,F_v} J = \frac{1}{N_w}\|F_v\overline{V} - F_s\bar{S}\|_F^2 - \frac{\alpha}{N_b}\|F_v\bar{V}^\dagger - F_sS^\dagger\|_F^2, \tag{17}$$

where $\|\square\|_F^2$ stands for the Frobenius norm of matrix $\square$.

We adopt the gradient descend algorithm for optimizing the model above. According the matrix theory, the derivation of function $J(\theta_s, \theta_v)$ with respect to parameters $\{\theta_s, \theta_v\}$ can be computed as:

$$\begin{aligned}\partial J/\partial F_s =\ & \frac{2}{N_w}(F_sSN^\dagger S^\top - F_v\overline{V}\bar{S}^\top) \\ & -\frac{2\alpha}{N_b}(F_sSN^\ddagger S^\top - F_v\bar{V}^\dagger S^{\dagger\top}),\end{aligned} \tag{18}$$

$$\begin{aligned}\partial J/\partial F_v =\ & \frac{2}{N_w}(F_v\overline{V} - F_s\bar{S})\overline{V}^\top \\ & -\frac{2\alpha}{N_b}(F_v\bar{V}^\dagger - F_sS^\dagger)\bar{V}^{\dagger\top},\end{aligned} \tag{19}$$

where $\square^\top$ denote the transposition of matrix $\square$, and $N^\dagger \in \mathbb{R}^{N_S \times N_S}$ and $N^\ddagger \in \mathbb{R}^{N_S \times N_S}$ are diagonal matrixes with the i^{th} diagonal element as N_{M_i} and $(N_m - N_{M_i})$ respectively. Finally, the parameters $\{F_s, F_v\}$ are updated according to:

$$\begin{aligned}F_s &= F_s - \eta \times (\partial J/\partial F_s), \\ F_v &= F_v - \eta \times (\partial J/\partial F_v),\end{aligned} \tag{20}$$

where η is the learning rate, namely the step size of parameter updating at each iteration. A too large rate leads the updating unstable, while a too small rate makes the convergence too slow. So, choosing a proper learning rate is vital for gradient descent. In our experiments, we adjust the learning rate η according to the gradient value $\partial J/\partial F$, so that the increment of gradient $\eta \times (\partial J/\partial F)$ is comparative to the original value F and the parameter updating is stable.

However, the gradient descent algorithm is likely to converge to a local minimum for many optimization problems. To discuss the rationality of using the gradient descent method to solve our problem in Eq. (17), we introduce another optimization method, the eigenvalue decomposition method [15], [16] for auxiliary illustration. The eigenvalue method proposed by [15] is derived analytically and obtains the accurate analytical solution of Eq. (13) by a two-stage eigen-decomposition algorithm. It involves no iterative updating operation. Moreover, the analysis and demonstrations in [16] provide detailed proofs ensuring that the eigenvalue method efficiently achieves the global optimal solution of problems like Eq. (13). Here, Eq. (17) is the matrix form of Eq. (13), which are equivalent to each other. Namely, the eigenvalue method ensures to get the global optima of Eq. (17) equivalently. Therefore, we introduce the eigenvalue method as the standard reference to check the convergence performance of the gradient descent method towards Eq. (17). Specially, our experimental results show that the recognition rates from the eigenvalue method are almost the same with that of the gradient descent method. This result, to some extent, indicates that the gradient descent method converges to the global minimum of Eq. (17) and it is reasonable to use the gradient descent method for optimizing in this paper.

2.4 Recognition Algorithm by PMDA

According to the description in Sect. 2.1, S2V face recognition comes down to calculating the similarity of a point and a set of local models in the mapping space. Given the scenario-oriented mapping matrixes $\{F_s, F_v\}$, we can obtain the following projection expressions for still image s_i and local model centers $\overline{V}_j = \{\overline{\mathbf{v}}_{j,l}\}_{l=1}^{N_{M_j}}$ of video clip V_j:

$$y_i = F_s \times s_i, \quad i = 1, 2, \ldots, N_S, \tag{21}$$

$$r_{j,l} = F_v \times \overline{\mathbf{v}}_{j,l}, \quad l = 1, 2, \ldots, N_{M_j}, \tag{22}$$

where $R_j = \{r_{j,1}, r_{j,2}, \ldots, r_{j,N_{M_i}}\}$ is the model based projection of V_j. Typically, when two sets with the same identity contain images taken from different conditions but with a certain overlap, to match them as the same class, the most effective solution is to measure the similarity of their most common parts [17]. Therefore, we define the S2V distance by matching the closest parts of two modes as:

$$d(y_i, R_j) = \min_{l=1,2,\ldots,N_{M_j}} d(y_i, r_{j,l}) \tag{23}$$

Finally, the recognition of video clip V_j in Eq. (3) is performed in a tractable way:

$$\mathcal{A} : \hat{i} = \arg \min_{i=1,2,\ldots,N_s} d(y_i, R_j)_{s_i \to y_i, V_j \to \mathcal{R}_j}, \tag{24}$$

3. Experiment

To evaluate the proposed PMDA, we perform face recognition experiments on the COX-S2V database [7]. The introduction of the database and experiment setting is shown

Table 1 Environment setting of the four videos. For Viewpoint, Illumination and Expression, "√" means the setting is fixed; "×" indicates the setting is varying. For Resolution, "√" means without degradation; "×" means with degradation.

	Video1	Video2	Video3	Video4
Viewpoint	√	√	×	×
Illumination	×	×	×	×
Expression	×	×	×	×
Resolution	×	√	×	√

in Sect. 3.1. The model parameters are analyzed detailedly in Sect. 3.2. At last, we give the evaluation in Sect. 3.3 and comparison with other methods in Sect. 3.4.

3.1 Database and Experiment Setting

COX-S2V is a dataset designed for the real-world Still-to-Video face recognition research, released by [7]. The dataset consists of high resolution still images and four low resolution videos of 1000 subjects. Table 1 gives the shooting condition of the videos. In particular, video 1 and 3 are more blur than video 2 and 4 for further shooting distance. To clearly show the characteristics of different videos, we provide some specific frames for each type of video in Fig. 4. According to the protocol in [7], we use the still images and video clips of 300 persons for training, and that of the rest 700 persons for testing. During testing stage, the still images serve as the gallery, and videos serve as the probe. The rank-1 recognition rate is used to test the performance.

All images are scaled to 96×120 pixels firstly. The pixel descriptor is used for the baseline assessment. To explore the potentiality of the proposed PMDA, we use local phase quantization (LPQ) and Gabor magnitude [18] to compose the complementary phase-magnitude descriptor for face representation. For LPQ, we set the local window size and the low frequency coefficient as $7 \times 7, 1/7$ respectively. For Gabor, we use 40 Gabor wavelets with 5 scales and 8 orientations. The Gabor kernel's size, the frequency parameters k_{max}, f^v, and the parameter σ are set to $31 \times 31, 1.0$, $\sqrt{2}$, and 2 respectively. Before training, we apply Principal Components Analysis (PCA) [19] on all the descriptors, and the target dimension is set as 1400 compromising between discrimination preserving and noises compression.

Parameters selection is a key issue. For the model constructing, the number of kNN candidates is set to 5. For the discriminant training, we set learning rate η=0.01 when using LPQ descriptor, and η=0.0001 when using Gobor descriptor. The matrix pair $\{F_s, F_v\}$ are initialized using unit matrix. Besides, the trade-off parameter α is set to 0.5 for equal weights of compactness and separability constraints.

3.2 Analysis of Model Parameters

The important parameters of PMDA include: linearity degree threshold θ, in the model construction step; the dimension of mapping space d', in the discriminant learning step. To provide a better understanding of the proposed approach, we give a detailed analysis of these parameters. Experiments are performed on video 2 and 4 as a representative.

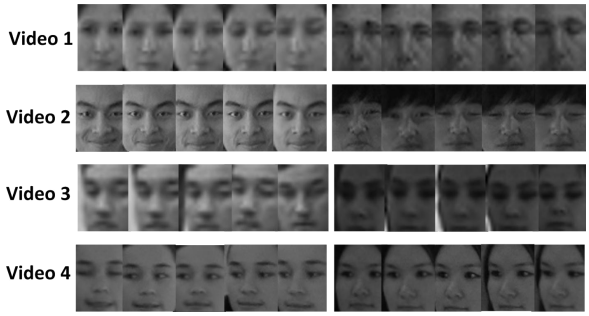


Fig. 4 Examples of the video frames in the COX-S2V dataset. Each row corresponds to the video frames from Video 1, 2, 3, 4 respectively. We present two clips of frame sequences for each type of video.

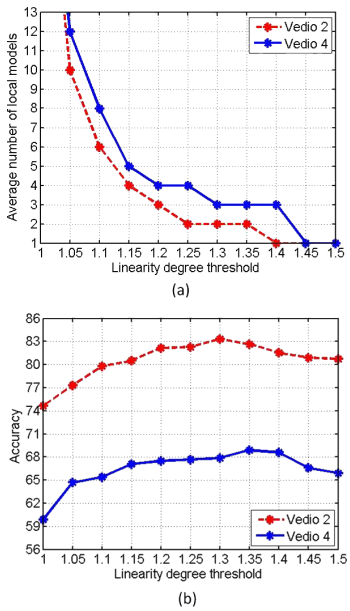


Fig. 5 Analysis of the linearity degree threshold. (a) Average number of local models on different linearity degrees. (b) Performance with different linearity degrees.

3.2.1 Linearity Degree Threshold

In a strict linear space, the nonlinearity degree defined in Eq. (6) results in 1 for all the pair-wise samples. As we just use linearity constraints as a auxiliary for the distance-based clustering algorithm to improve the consistency of local models, strict linearity is not exactly required. Following, we experiment to search for a relatively relaxed value for linearity degree threshold θ, which is beneficial for subsequent discriminant training meanwhile. The threshold varies from 1 to 1.5, with an interval of 0.05. Experimental results on the LPQ descriptor are shown in Fig. 5, as a representative.

Subgraph (a) reflects how the threshold influences the results of local model constructing. Generally, the model number of video 4 is slightly bigger than that of video 2,

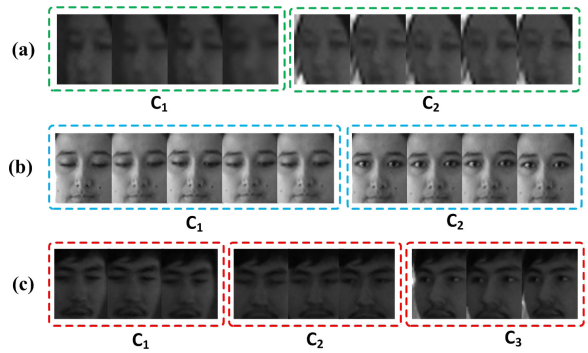


Fig. 6 Some results of the learned local manifolds. The images circled in a dotted box present a local model, denoted by $C_i, i = 1, 2, 3$. (a) The local models from a varying-lights video: C_1 clusters the frames of dark light; C_2 clusters the frames of bright light. (b) The local models from a varying-expressions video: C_1 clusters the frames of close eyes; C_2 clusters the frames of open eyes. (c) The local models from a varying-views video: C_1 clusters the frames of rightward view; C_2 clusters the frames of front view; C_3 clusters the frames of leftward view.

for more complicated shooting condition. Similar variations appear on both of the videos. Just as θ is around 1.3, the results reach a stable level at 2 or 3. As the linearity constraints strengthen (θ gets smaller), the algorithm produces more local models. Particularly, when θ reduces close to 1, the model number approaches the sample number. However, as θ grows over a certain value, all the samples are clustered as a single model. Subgraph (b) shows the influence to the global performance. As seen, just as θ is round 1.3, the recognition rate reaches the peak. According to the analysis in Sect. 2.3, a large number of local models would lead to badly discordance of the sample number in still scene and video scene, and further serious model bias for subsequent discriminant learning. Therefore, the accuracy decreases severely as θ approaches 1. Contrastively, as θ gets bigger, the algorithm tend to use the average of the video frame set to represent each video clip. Avoiding the model bias, performance degenerates slightly for losing the diversity of samples in the simply average operation. Above all, we set θ at the value corresponding the model number as 2 and 3 for video V2 (V1) and V4(V3) respectively, namely θ=1.25 for V2 (V1), and θ=1.3 for V4(V3).

Moreover, we also present some results of the learned local manifolds in Fig. 6 to claim the effectiveness of the proposed linearity constrained clustering (LCC) method. In Fig. 6, each row shows the learned local models from a video. The images circled in a dotted box present a local model. In this paper, we assume that the frames of a video distribute on a nonlinear manifold and develop the LCC algorithm to learn the local models on a video. As described in Sect. 2.2, the LCC clusters the frames sequence with similar statistic property as a local model. The local model we obtain is supposed to be a collection of similar frames (with similar light, expression or view) in a video. From Fig. 6, we can see that for each video, the frames with similar setting (light, expression or view) are just separately clustered as a local model in our experiments, which indicates the effectiveness of the proposed clustering method. Specially, we just show several typical frames for each model in this figure for simplicity.

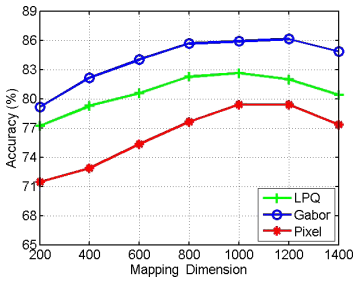


Fig. 7 Performance with different mapping dimensions.

Table 2 Accuracy of the proposed PMDA(%).

	Video1	Video2	Video3	Video4
Pixel	63.23	79.40	19.42	54.31
LPQ	61.25	82.26	27.82	67.86
Gabor	69.35	85.66	29.43	64.35
LPQ+Gabor	**79.02**	**89.53**	**38.22**	**76.64**

3.2.2 Dimension of the Mapping Space

The original feature has been reduced to 1400 dimension by PCA before training. In this section, we vary the mapping dimension from 200 to 1400 by a step of 200 to test which is more suitable for the S2V facial description. Experimental results on video 2 are drawn in Fig. 7. As seen, the accuracy benefits from increasing the mapping dimension. With too low dimension, performance drops for losing much discriminative information in the mapping operation. However, as the dimension exceeds 800, the accuracy increment is inapparent, less than 0.5% for LPQ and Gabor. The same situation also appears on the pixel descriptor towards the dimension 1000. Moreover, continued growth even leads to a downward trend of performance. The phenomenon may arise from that a high dimension makes the model much too complicated for the current problem, leading the model overfitting on the small training set while generalizing badly on the test set. Above all, considering increasing the feature dimension makes training much more complex, we set the mapping dimension as 800 for LPQ and Gabor, and 1000 for the pixel descriptor.

3.3 Evaluation of the PMDA Approach

At last, we give a global evaluation of the PMDA approach based on the discussions above. Experimental results on the COX-S2V dataset are shown in Table 2. Accuracy on the pixel descriptor gives the baseline performance of the proposed approach, which is quite satisfactory on the difficult identifying environment. The LPQ and Gabor promote the performance significantly comparing with the baseline, suggesting that the two descriptors are quite effective for capturing information on the uncontrolled environment. The last line of Table 2 gives the accuracy derived from fusing the similarities of LPQ and Gabor with a simple average op-

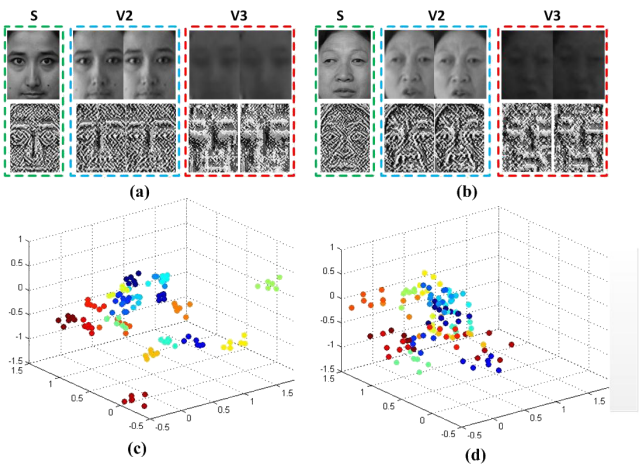


Fig. 8 An illustration of the extracted features from the blurred videos. (a)-(b) The first row presents three types of images: the high-resolution still images (S); the frames from the clear Video 2 (V2); the frames from the blurred Video 3 (V3). The second row presents the LPQ feature maps on the corresponding images. (c) The feature space of the clear frames in Video 2 (V2). (d) The feature space of the blurred frames in Video 3 (V3). The LPQ features of 20 subjects, each of which has 6 images randomly selected from the corresponding video clip, are projected into a three-dimension space. Images of the same subject are visualized in the same color.

eration. The fusing results show remarkable improvement over that of single descriptor. Actually, LPQ emphasizes the facial texture information, while Gabor magnitude emphasizes the structure information of faces [18]. Accuracy can be obviously improved by combining the two features, for their strong complementation to each other.

Moreover, from Table 2, we can find that video 1 and 2 generally achieve better accuracy than video 3 and 4 on all the descriptors. Specially, performance of video 1 is worse than video 2, and performance of video 3 is further worse than video 4. Referring to Table 1, video 1 and 2 are captured under fixed pose, varying lighting and expression, while video 3 and 4 are under varying pose, lighting and expression. Video 1 and 3 are more blur than video 2 and 4 for further shooting distance. Based on the condition differences on the four videos, the performance comparisons above indicate that changes in pose and the resolution degradation are two main factors affecting the recognition rate. Actually, referring to Fig. 6, local models from varying-lighting (expression) videos look similar and just the local image changes a little, while models from varying-pose videos correspond to different visible parts of a face and show large differences. In this situation, it is hard to model the complex variations in the varying-pose local models effectively for the proposed method by using a single linear mapping. On the other hand, the resolution degradation makes the content in an image hard to identify. In this situation, it is difficult to acquire discriminative information for different classes. An illustration of the extracted features from the blurred videos is shown in Fig. 8. We take the LPQ feature on Video 3 as an example. Figure 8 (a) and (b) present the LPQ feature maps of the high-resolution still images (S), the clear frames (V2) and the blurred frames (V3) for two subjects respectively. We observe that the feature maps in V2 is similar with that of the still images (S), which both precisely encode the facial texture information. However, the feature maps in V3 tend to consist of scattered response points, which can't describe the facial texture effectively. Moreover, the comparisons between Fig. 8 (a) and (b) further show that for two different subjects, the feature maps in V3 look rather similar. In this situation, it is hard to distinguish different subjects via the features of the blurred images. To statistically illustrate this issue, we further present the feature spaces for the clear frames (V2) and the blurred frames (V3) in Fig. 8 (c) and (d) respectively. As seen, the features of the same classes mostly cluster together and the features from different classes hold certain separability in V2, while the features from different classes mix together and the inter-class borders are hard to identify in V3. This result shows that the extracted features from the blurred images are inadequate to compose the descriptors to provide the inter-class discriminative information effectively.

Based on the discussion above, changes in pose and the resolution degradation are two main factors degenerating the learning ability of the proposed method. Referring to Table 1, we can figure that video 3 is captured under the worst condition (varying pose and of low resolution). In this way, learning on video 3 is much less effective, and video 3 finally gets the worst performance.

Moreover, to evaluate the optimization algorithm in our method, we also experiment by introducing the eigenvalue decomposition method [15] to optimize our model. The recognition performance is shown in Table 3. Comparing the results of Table 2 and Table 3, we can get that the recognition rates of the gradient descent method (Table 2) are almost the same with that of the eigenvalue method on all the four videos. Since the eigenvalue method ensures to get the global optima of Eq. (17) as analyzed in Sect. 2.3.2, the result above, to some extent, indicates that the gradient descent method converges to the global minimum of our objective function in Eq. (17). Therefore, it is adequate to use the gradient descent method for optimizing in this paper. Specially, since the gradient descent method performs iterates to approach the optimal solution and we apply just a few iterations (around ten) in our experiments, the performance of the gradient descent method in Table 2 is slightly inferior to the optimal performance in Table 3.

Table 3 Accuracy of the PMDA optimizing with the eigenvalue decomposition method [15](%).

	Video1	Video2	Video3	Video4
Pixel	63.29	79.43	19.57	54.42
LPQ	61.28	82.28	27.85	67.86
Gabor	69.42	85.71	29.43	64.42
LPQ+Gabor	79.14	89.57	38.28	76.71

3.4 Comparison Results with Other Methods

We compare the PMDA to state-of-art methods in Table 4. Here, we include the results in [7] from the database

Table 4 Performance comparison on the COX-S2V dataset (%).

Methods	Video1	Video2	Video3	Video4
LDA+Pixel [7]	47.57	68.28	20.00	49.85
LPP+Pixel [7]	47.43	68.57	20.12	49.14
LFDA+Pixel [7]	21.86	44.00	3.29	16.14
CDEF+Pixel [7]	8.14	12.99	6.57	5.00
M-CLAFIC+Pixel [6]	6.36	10.72	3.89	4.27
PaLo-LDA +Pixel [7]	52.43	73.00	22.00	56.71
PMDA+Pixel	**63.23**	**79.40**	**19.42**	**54.31**
PaLo-LDA+LPQ+Gabor	65.74	88.25	23.57	72.71
PMDA+LPQ+Gabor	**79.02**	**89.53**	**38.22**	**76.64**

releaser, including methods such as LDA [8], LPP [20], LFDA [21]. As the subspace method is closely related in terms of S2V recognition problem, we also experiment with the classical modified CLAFIC method [6]. Performence of conventional discriminant analysis methods, which apply a single mapping for all the data from different scenarios, are generally not so good. The classical subspace method (M-CLAFIC) tends to project the input data (still images) to the reference subspace build on the videos for classification. Due to the great discrepancies of cross-scenario data, the direct projection process may not capture the essential features of intra-class samples in this situation, and the performance also degenerates to some extent. The best results reported are from the PaLo-LDA upon the pixel descriptor. With the pixel descriptor, our approach performs better on video 1 and 2, but worse on video 3 and 4. In complex conditions, the performance of PMDA may be disturbed by the low-level expression of the original pixels. Actually, with the phase-magnitude descriptor, our approach performs significantly better than the PaLo-LDA on all the four videos, with a large increment of 13.28%, 1.28%, 14.65%, 3.93% correspondingly. As seen, advantages on the videos of complex conditions (video 1,3,4) are pretty obvious.

The PaLo-LDA is an improved version from classical LDA, which uses partial weighting and local weighting to take the cross-resolution and other variations (e.g., pose, illumination, lighting etc.) into account for discriminative learning. This method involves weights calculation of all the pair-wise samples in the training set and a generalized eigenvalue problem on the scatter matrixes, which are both time-consuming tasks as the sample number gets large. Yet, in the PMDA, instead of turning to the weighting skills, we seek scenario-oriented mappings for the cross-scenario problem. Comparing with the single-mapping based PaLo-LDA, our approach could model the underlying data manifold in corresponding scenario more effectively, and thus identifies faces in varying scenarios much better. Moreover, our iterative optimization approach effectively avoids the complicated matrix decomposition of eigen-solvers.

As for the complexity, the PaLo-LDA mainly involves weight matrix calculation and eigenvalue decomposition. Therefore, the two-part computational cost is $O(DN^2 + D^3)$, where N is the sample number including all the still images and video frames, and D is the feature dimension. In the PMDA, learning includes two stages: local model construction and model based discriminant training. Detailedly, the clustering based model construction costs $O(N_S N_0 N_C)$, where N_0 and N_C are the video frames number and the model number in each video clip, and N_S is the category number enrolled. As each category of samples are clustered independently, this part could be executed in parallel very efficiently. Using the denotation in Sect. 2.3, the cost of gradient computation during discriminant learning results in $O(D^2 N_S N_m (d_S + d_V))$. In our application, the mapping dimensions d_S, d_V are rather small ($d_S, d_V \ll D$), and the category number N_S and model number N_m are generally much smaller than the sample number N. From this, the gradient computation in Eqs. (18), (19) is rather efficient. In fact, it just needs dozens of iterations (around ten) before the PMDA converges. Moreover, through experiments, we find that by choosing a proper learning rate, the objective function value decreases steadily, and the iterative algorithm is guaranteed to get a optimal solution finally.

4. Conclusions

We have developed a Point-Manifold Discriminant Analysis approach for Still-to-Video face recognition. The algorithm models a video clip as a manifold and a still image as a data point and learns separate mappings for samples in different scenario patterns (still, video). The optimization of mappings is based upon separability-compactness constraints. Comparative experiments indicate the proposal's high accuracy and robustness in the S2V scenario. A limitation of the PMDA is applying simple linear projections to model the rich nonlinear variations in facial appearance. Our future work will focus on the extension to nonlinear mapping learning and the exploration of local fiducial features.

References

[1] A. Hadid and M. Pietikainen, "From still image to video-based face recognition: an experimental analysis," Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. pp.813–818, 2004.

[2] R. Wang, H. Guo, L.S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2496–2503, 2012.

[3] Y. Li, S. Gong, and H. Liddell, "Recognising trajectories of facial identities using kernel discriminant analysis," Image Vis. Comput., vol.21, no.13, pp.1077–1086, 2003.

[4] X. Liu and T. Cheng, "Video-based face recognition using adaptive hidden markov models," 2003. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.I–340, 2003.

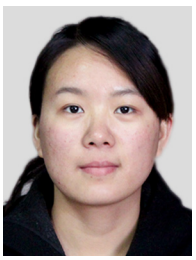
[5] E. Oja, Subspace Methods for Pattern Recognition, Research Study Press, 1983.

[6] Y. Ariki and W. Ishikawa, "Integration of face and speaker recognition by subspace method," Proc. 13th International Conference on Pattern Recognition, 1996, pp.456–460, 1996.

[7] 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," Computer Vision C ACCV 2012, pp.589–600, Springer Berlin Heidelberg, 2012.

[8] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,"

IEEE Trans. , Pattern Anal. Mach. Intell., vol.19, no.7, pp.711–720, 1997.
[9] J.B. Tenenbaum, V. De Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol.290, no.5500, pp.2319–2323, 2000.
[10] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.
[11] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp.1–8, 2008.
[12] R. Wang and X. Chen, "Manifold discriminant analysis," IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. pp.429–436, 2009.
[13] K.C. Lee, J. Ho, M.H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. pp.I–313, 2003.
[14] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," Neural Networks for Signal Processing IX, 1999. Proc. 1999 IEEE Signal Processing Society Workshop., pp.41–48, 1999.
[15] D. Lin and X. Tang, "Inter-modality face recognition," Computer Vision–ECCV 2006, pp.13–26, 2006.
[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," NIPS, pp.585–591, 2001.
[17] 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, 2007.
[18] Y. Li, S. Shan, H. Zhang, S. Lao, and X. Chen, "Fusing magnitude and phase features for robust face recognition," in Computer Vision - ACCV 2012, Springer Berlin Heidelberg, 2012.
[19] I.T. Jolliffe, Principal component analysis, Springer verlag, 2002.
[20] X. Niyogi, "Locality preserving projections," Neural Information Processing Systems, p.153, 2004.
[21] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," J. Machine Learning Research, vol.8, pp.1027–1061, 2007.



Xue Chen received the B.S degree in the school of automation from Huazhong University of Science and Technology, Wuhan, China in 2010. She is currently pursuing the PH.D. degree in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include face recognition, machine learning, and computer vision.

Chunheng Wang received the B.Eng and M.Eng degree from the Dalian university of technology, and the Ph.D degree from the Institute of Automation, Chinese Academy of Sciences. He is currently a professor of State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interest includes pattern recognition, image processing, face recognition, and artificial intelligence. He has published over 40 refereed research papers. He is a member of IEICE.

Baihua Xiao received the B.Eng degree from the automatic control department of the Northwestern Polytechnical University, Xi'an, China, and the Ph.D degree from the Institute of Automation, Chinese Academy of Sciences. He is currently a professor of State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research includes pattern recognition and intelligent system, computer vision, multi-media information processing and retrieval. He has published over 30 conference or journal papers.

Yunxue Shao received the B.S degree in the department of computer science and technology from HeHai University, Nanjing, China in 2008. He is currently pursuing the PH.D. degree in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include character recognition, and machine learning, and pattern recognition. He is a student member of IEICE.