Multi-manifold metric learning for face recognition based on image sets

Huang, Likun; Lu, Jiwen; Tan, Yap Peng

2014

Huang, L., Lu, J., & Tan, Y.-P. (2014). Multi-manifold metric learning for face recognition based on image sets. *Journal of visual communication and image representation*, 25(7), 1774-1783.

<https://hdl.handle.net/10356/103837>

<https://doi.org/10.1016/j.jvcir.2014.08.006>

© Elsevier Inc. This is the author created version of a work that has been peer reviewed and accepted for publication by *Journal of Visual Communication and Image Representation*, Elsevier Inc. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1016/j.jvcir.2014.08.006>].

Downloaded on 26 Apr 2023 11:21:59 SGT

Multi-Manifold Metric Learning for Face Recognition Based on Image Sets

Likun Huang^{a,*}, Jiwen Lu^b, Yap-Peng Tan^a

^a*School of Electrical and Electronics Engineering, Nanyang Technological University,
Singapore, 639798.*

^b*Advanced Digital Sciences Center, Singapore, 138632.*

Abstract

In this paper, we propose a new multi-manifold metric learning (MMML) method for the task of face recognition based on image sets. Different from most existing metric learning algorithms that learn the distance metric for measuring single images, our method aims to learn distance metrics to measure the similarity between manifold pairs. In our method, each image set is modeled as a manifold and then multiple distance metrics among different manifolds are learned. With these distance metrics, the intra-class manifold variations are minimized and inter-class manifold variations are maximized simultaneously. For each person, we learn a distance metric by using such a criterion that all the learned distance metrics are person-specific and thus more discriminative. Our method is extensively evaluated on three widely studied face databases, i.e., Honda/UCSD database, CMU MoBo database and Youtube Celebrities database, and compared to the state-of-the-arts. Experimental results are presented to show the effectiveness of the proposed method.

Keywords:

Face recognition, face image sets, metric learning, manifold-to-manifold distance.

*Corresponding Author

Email addresses: huan0181@e.ntu.edu.sg (Likun Huang), jiwen.lu@adsc.com.sg (Jiwen Lu), eyptan@ntu.edu.sg (Yap-Peng Tan)

1. Introduction

Over the past few decades, face recognition has attracted great research interest in a wide range of applications in security and commerce. A large number of methods of face recognition based on single-shot images have been proposed in the literature [1, 2, 3, 4, 5, 6, 7, 8], which have already achieved satisfying performance under controlled conditions. However, most of these methods can not guarantee a reliable recognition performance in many practical environments, where face images are usually captured under unconstrained conditions with large variations in illumination, expression, pose, occlusion, and etc.

Compared with a single face image, an image set is able to provide more useful information to cover the variations of a person’s facial appearance. Hence, more discriminative information can be exploited from image sets to model the appearances of human faces. Generally, an image set can be a collection of unordered images from different views or frames from a video clip, even the video clip is not continuous. With the rapid development of social networks, a large number of social media sharing websites, such as Facebook, YouTube and Flickr, have emerged. Everyday hundreds of millions of people are posting their personal photos and videos onto these social media networks. With the massive data, video sequences and multiple still images of an individual are much easier to obtain than before. This motivates us to investigate the problem of face recognition based on image sets in this paper, where an input is an image set that needs to be classified to one of the classes in the gallery dataset, in which each reference is also an image set.

In this work, we propose a new multi-manifold metric learning (MMML) method for face recognition based on image sets. Given a face image set, we consider it as a manifold and model it using a collection of affine hulls. Based on this description, the manifold-to-manifold distance is transformed to the pairwise affine hull distance. To make these affine hulls more separable, we learn multiple distance metrics, one for each person, such that the inter-class manifold variations can be maximized and the intra-class manifold variations can be minimized simultaneously. In the classification phase, a probe image set is also modeled as a collection of affine hulls and compared with each gallery set associated with the learned distance metric. For a better illustration, Fig. 1 shows the basic idea of our proposed method. The key features of the proposed MMML are summarized as follows:

Figure 1: An illustration of our proposed multi-manifold metric learning method. For each image set, we first model it as a manifold, which is then approximated by a collection of affine hulls. For each person, we learn a person-specific distance metric to maximize the inter-class manifold variations and minimize the intra-class manifold variations simultaneously, such that more discriminative information can be exploited for recognition. In the testing phase, the probe image set is compared with each gallery image set that is associated with a learned distance metric, and then a label is assigned according to the nearest neighbor rule.

- MMML learns discriminant metrics. These metrics are able to maximize the separability of neighboring manifolds from different classes, and at the same time minimize the compactness of manifolds from the same class.
- MMML learns class-specific metrics. In contrast to existing metric learning and subspace learning algorithms that learn only one single distance metric or a global linear transformation for all the samples from different classes, the proposed MMML learns a collection of discriminative distance metrics, one for each class. The class-specific distance metrics are able to better differentiate the current class from others than a unified metric, since face images from different classes may lie on different manifolds with different intrinsic dimensions.

- MMML learns set-based metrics. Most existing metric learning methods are instance-based methods that can not deal with the comparisons among image sets. In the scenario of classification based on image sets, the proposed distance metric is more suitable to measure the distances between manifolds.

The rest of this paper is organized as follows. In Section II, we review the related work on the problem of face recognition based on image sets. Then we propose a Multi-Manifold Metric Learning (MMML) method in Section III, which utilizes the discriminative property of the metric learning step to enhance the recognition performance of multi-model based methods. For better illustrations and comparisons, our method is extensively evaluated on three popular and widely used face databases and compared to the state of the arts. Experimental results are presented in Section IV to show the effectiveness of the proposed method. Finally, Section V concludes the paper.

2. Related Works

In this section, we briefly review the existing work on two related topics: 1) face recognition based on image sets, and 2) metric learning.

2.1. Face Recognition Based on Image Sets

Recently, a number of methods have been proposed for face recognition based on image sets [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], which can be mainly classified into three categories: model-free methods, parametric methods and nonparametric methods.

For the model-free methods, there is usually no mathematical model used to describe a face image set. To compute the set-to-set distance, these methods compare two image sets by directly calculating the distances between pair-wise samples across different sets [15] [32] or adopt the minimum linear reconstruction error as the distance between sets [28]. All of these methods suffer a limit that the computational complexity rises with the increasing size of each image set.

For parametric methods, an image set is usually modeled as a predefined parametric distribution [11, 12, 13], such as a multivariate Gaussian distribution or Gaussian Mixture Models (GMM), and then a probability similarity measure such as the Kullback-Leibler divergence is utilized to measure the similarity between two such models. However, the multivariate Gaussian

distribution is generally insufficient to represent an image set containing the complex variations [14]. Thus a more principled model known as GMM was proposed to model such complex sets. Unfortunately, the GMM model usually requires a very high computational complexity. One main shortcoming of parametric methods is that they need to address a challenging problem of parameter estimation, which may fail to work properly when there are weak statistical relationships between the training and testing image sets. Furthermore, the classification performance could be affected if the practical data sets do not follow the predefined densities. Due to these limitations, nonparametric methods have attracted more and more research attention in face recognition based on image set in recent years.

In the last decade, many nonparametric methods have been proposed to measure the similarity between two image sets [18, 19, 10, 20, 23, 25, 26, 9, 30, 31]. Representative nonparametric models include linear subspace, affine subspace, convex hull, regularized hull, etc. Among these models, the linear subspace model that requires the lowest computational complexity was adopted in [18, 19, 10, 20]. However, a linear subspace is a rather loose representation compared to a convex/regularized hull, which was utilized in [9, 30]. These linear models can fill-in the missing data for each image set and show robustness to the outliers. However, most existing nonparametric methods only used certain similarity measurements to compute the distance between two image sets, whereas a discriminative learning procedure was absent. Hence, some discriminative information may not be effectively exploited in the classification phase. More recently, Wang and Chen [33] proposed a manifold discriminant analysis method to learn a linear subspace such that the image sets from different classes can be separated. Unfortunately, only one single subspace was learned in their method, which may not be discriminative enough because face images from different classes may lie on different manifolds with different intrinsic dimensions. Another way to exploit discriminative information is by use of sparse representations [29, 31]. However, the improvements on recognition performance come with a significant increase in the computational complexities.

Motivated by the above work, in this paper we propose a multi-manifold metric learning method by learning multiple distance metrics such that the person-specific and discriminative features can be exploited for recognition.

2.2. Metric Learning

A number of distance metric learning methods have been proposed in recent years [34, 35, 36, 37, 38, 39], which can be mainly classified into two categories: unsupervised and supervised. Generally, supervised methods are more popular for classification because more discriminative information can be exploited in the learned distance metric. While a number of distance functions have been developed for different metric learning methods, the Mahalanobis metric is the most popular one because it has shown good generalization ability.

However, most existing metric learning methods only learn the distance metric from single image instances and cannot deal with image sets directly. A possible solution is to make use of each sample within an image set independently to learn a distance metric. However, the relations between these samples within an image set are ignored and some discriminative information fails to be exploited. Motivated by this, we first model each image set as a manifold and then learn distance metrics from these manifolds, rather than from single image instances [34, 35, 36, 37, 38, 39]. Hence, our proposed method is complementary to the existing metric learning methods, and it is more suitable for the problem of face recognition based on image sets.

3. Proposed Method

3.1. Multi-Affine Hulls Model

Consider N training sets from C classes. For each set, the face images are assumed to lie on or near an underlying nonlinear and intrinsically low-dimensional manifold $\mathbf{M}$. To better capture the characteristic of each manifold, we adopt multiple local models rather than a single model. Existing multi-model based methods [33, 40] first divide a nonlinear manifold $\mathbf{M}$ into several subsets $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$ and each subset $\mathbf{s}_i$, $i \in \{1, 2, \dots, m\}$, is then described by a linear subspace, such that a collection of linear subspaces is used to model the manifold $\mathbf{M}$. However, considering that the linear subspace is a rather loose approximation of a subset, we use more compact representations, i.e., a collection of affine hulls to characterize a manifold $\mathbf{M} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m\}$, where $\mathbf{h}_i$, $i \in \{1, 2, \dots, m\}$, is an affine hull. Since the mean information is important to characterize an affine hull $\mathbf{h}_i$ because it reflects the averaged positions of face images within a set, we divide an image set into multiple subsets by using the spectral clustering algorithms [41] [42]

such that a stable and representative mean vector can be obtained for each affine hull.

The spectral clustering: The spectral clustering algorithms [41] [42] cluster the data points based on the first k eigenvectors of a normalized Laplacian matrix $\mathbf{L}$ of $\mathbf{S}$, which is an adjacency matrix storing the similarities between pairwise points of a data set $\mathbf{X}$. Then the objective function aims to find a collection of orthogonal embeddings of the data set denoted as $\mathbf{V}$, which consists of the approximations of the cluster indicator vectors:

$$\max_{\mathbf{V} \in \mathbf{R}^{n \times k}} \text{tr}(\mathbf{V}^T \mathbf{L} \mathbf{V}) \quad s.t. \quad \mathbf{V}^T \mathbf{V} = \mathbf{I}, \quad (1)$$

where the optimal $\mathbf{V}$ contains the first k eigenvectors of the symmetric normalized Laplacian matrix $\mathbf{L}$ as columns. Then each row vector of $\mathbf{V}$ can be normalized and used to assign a data point to one of the k clusters through the k -means algorithm [43]. The spectral clustering algorithm has many nice properties such as well-defined mathematic framework, more discriminative nature for clustering, good performance on clusters with arbitrary shapes, and etc.

The affine hull model for each subset: For a subset $\mathbf{s}_i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_{n_i}^i]$, its affine hull approximation $\mathbf{h}_i$ is defined as

$$\mathbf{h}_i = \{\mathbf{x}_i = \mu_i + \mathbf{U}_i \mathbf{v}_i | \mathbf{v}_i \in \mathbf{R}^D\}, \quad (2)$$

where sample $\mathbf{x}_i \in \mathbf{R}^D$, μ_i is the mean vector of the subset $\mathbf{s}_i$, $\mathbf{U}_i$ is the orthonormal basis that spans the whole affine hull and $\mathbf{v}_i$ is a vector of free parameters of $\mathbf{x}_i$ within the affine hull. $\mathbf{U}_i$ is obtained from the singular vectors of Singular Value Decomposition (SVD) on $[\mathbf{x}_1^i - \mu_i, \mathbf{x}_2^i - \mu_i, \dots, \mathbf{x}_{n_i}^i - \mu_i]$. By using an affine hull, a subset can be compactly represented since the affine subspace is able to reduce more redundant dimensions than a linear subspace that is forced to go through the origin.

Differences from the AHISD method [9]: The method that utilizes affine hulls or convex hulls to represent image sets was first proposed in [9]. In this AHISD method, an image set is modeled by using a single affine hull or convex hull. In contrast, we first cluster an image set into multiple subsets through a spectral clustering method, and model each subset as an affine hull. Then the whole image set is described by using a collection of affine hulls, i.e., $\mathbf{M} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m\}$. Thus, our proposed method has a stronger capability to characterize an image set with nonlinear distributions.

3.2. Affine Hull-Based Distance Metric

Most existing work such as [10], [40] design the manifold-to-manifold distance measure as the dissimilarity between the most similar parts of two image sets, i.e.,

$$d(\mathbf{M}_p, \mathbf{M}_q) = \min_{\mathbf{h}_i \in \mathbf{M}_p} \min_{\mathbf{h}_j \in \mathbf{M}_q} d(\mathbf{h}_i, \mathbf{h}_j), \quad (3)$$

where $d(\mathbf{M}_p, \mathbf{M}_q)$ denotes the distance between manifolds $\mathbf{M}_p$ and $\mathbf{M}_q$, and $d(\mathbf{h}_i, \mathbf{h}_j)$ denotes the distance between affine hulls $\mathbf{h}_i$ and $\mathbf{h}_j$.

Instead of directly comparing the two affine hulls, we compare them by using a learned distance metric. The conventional metric learning algorithms learn a Mahalanobis distance metric that is an instance-based metric. However, in the considered scenario, an affine hull-based distance metric is needed. Hence, we define a distance metric over affine hulls as:

$$d^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{A}) = \alpha \frac{(\mu_i^T \mathbf{A} \mu_j)^2}{(\mu_i^T \mathbf{A} \mu_i)(\mu_j^T \mathbf{A} \mu_j)} + (1 - \alpha) \text{tr}(\Lambda), \quad (4)$$

where α is a scalar parameter and Λ is obtained from the SVD of $(\tilde{\mathbf{U}}_i^T \mathbf{A} \tilde{\mathbf{U}}_j)$, i.e., $(\tilde{\mathbf{U}}_i^T \mathbf{A} \tilde{\mathbf{U}}_j) = \mathbf{Q}_{ij} \Lambda \mathbf{Q}_{ji}^T$. $\mathbf{U}_i$ and $\mathbf{U}_j$ are orthonormal bases that span two affine hulls respectively. Before we explain $\tilde{\mathbf{U}}_i$ and $\tilde{\mathbf{U}}_j$ that are transformed from $\mathbf{U}_i$ and $\mathbf{U}_j$, we first introduce a decomposition of $\mathbf{A}$, where $\mathbf{A} = \mathbf{W} \mathbf{W}^T$ and $\mathbf{W} \in \mathbf{R}^{D \times d}$. This makes the parameter matrix $\mathbf{A}$ be a $D \times D$ symmetric and positive semidefinite matrix such that the affine hull-based metric is a valid metric that satisfies non-negativity, symmetry and triangle inequality. Then we use $\mathbf{W} \mathbf{W}^T$ instead of $\mathbf{A}$ such that Eq. (4) can be rewritten as

$$\begin{aligned} & d^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{W}) \\ &= \alpha \frac{(\mu_i^T \mathbf{W} \mathbf{W}^T \mu_j)^2}{(\mu_i^T \mathbf{W} \mathbf{W}^T \mu_i)(\mu_j^T \mathbf{W} \mathbf{W}^T \mu_j)} + (1 - \alpha) \text{tr}(\Lambda) \\ &= \alpha \frac{\left[(\mathbf{W}^T \mu_i)^T (\mathbf{W}^T \mu_j) \right]^2}{\left[(\mathbf{W}^T \mu_i)^T (\mathbf{W}^T \mu_i) \right] \left[(\mathbf{W}^T \mu_j)^T (\mathbf{W}^T \mu_j) \right]} \\ &+ (1 - \alpha) \text{tr}(\Lambda) \end{aligned} \quad (5)$$

where Λ is obtained from the SVD: $(\tilde{\mathbf{U}}_i^T (\mathbf{W} \mathbf{W}^T) \tilde{\mathbf{U}}_j) = (\mathbf{W}^T \tilde{\mathbf{U}}_i)^T (\mathbf{W}^T \tilde{\mathbf{U}}_j) = \mathbf{Q}_{ij} \Lambda \mathbf{Q}_{ji}^T$. It is observed that Λ contains the principal angles between ma-

trices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$ when the two matrices are orthonormal basis matrices of subspaces. Although $\mathbf{U}_i$ and $\mathbf{U}_j$ are orthonormal matrices, in the subspace $\mathbf{W}$, $(\mathbf{W}^T \mathbf{U}_i)$ and $(\mathbf{W}^T \mathbf{U}_j)$ are not generally orthonormal basis matrices. Thus, to obtain the orthogonalized matrices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$, we orthogonalize $(\mathbf{W}^T \mathbf{U}_i)$ and $(\mathbf{W}^T \mathbf{U}_j)$ by applying the QR decomposition

$$\begin{aligned}\mathbf{W}^T \mathbf{U}_i &= \Phi_i \mathbf{R}_i \\ \mathbf{W}^T \mathbf{U}_j &= \Phi_j \mathbf{R}_j\end{aligned}\tag{6}$$

where Φ_i and Φ_j are orthogonal matrices and $\mathbf{R}_i$ and $\mathbf{R}_j$ are invertable upper-triangular matrices. The orthogonalized matrices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$ can be obtained as follows

$$\begin{aligned}\Phi_i &= (\mathbf{W}^T \mathbf{U}_i) \cdot \mathbf{R}_i^{-1} = \mathbf{W}^T \cdot (\mathbf{U}_i \mathbf{R}_i^{-1}) = \mathbf{W}^T \tilde{\mathbf{U}}_i \\ \Phi_j &= (\mathbf{W}^T \mathbf{U}_j) \cdot \mathbf{R}_j^{-1} = \mathbf{W}^T \cdot (\mathbf{U}_j \mathbf{R}_j^{-1}) = \mathbf{W}^T \tilde{\mathbf{U}}_j\end{aligned}\tag{7}$$

where

$$\begin{aligned}\tilde{\mathbf{U}}_i &= \mathbf{U}_i \mathbf{R}_i^{-1} \\ \tilde{\mathbf{U}}_j &= \mathbf{U}_j \mathbf{R}_j^{-1}\end{aligned}\tag{8}$$

In Eq. (4) and Eq. (5), $tr(\Lambda)$ describes the correlation between two orthonormal basis matrices. Furthermore, $tr(\Lambda) \geq 0$ due to the fact that the principal angles are defined in the range of $[0, \frac{\pi}{2}]$. Since $\mathbf{U}_i$ and $\mathbf{U}_j$ are two different coordinate systems with respect to μ_i and μ_j respectively, the first term in Eq. (4) and Eq. (5) denotes how far away the origins of the two coordinate systems $\mathbf{U}_i$ and $\mathbf{U}_j$ are from each other in the subspace $\mathbf{W}$. Whereas the second term in Eq. (4) and Eq. (5) indicates that if the two coordinate systems are both shifted to the same origin, how relevant the two subspaces are. Based on Eq. (5), we observe that learning a good distance metric is equivalent to learn the linear transformations $\mathbf{W}$ for image sets in the original space.

3.3. Parameter Matrix Learning for Distance Metric

With N image sets from C classes, N intrinsically low dimensional manifolds $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_N$ are embedded in the high dimensional ambient space. In order to give a more compact description of a manifold meanwhile avoid the curse of dimensionality, a nonlinear dimension reduction procedure [4,

Figure 2: An illustration of the basic idea of the learning procedure for the i th class. For class i , we learn a distance metric $\mathbf{W}_i$ that pushes the manifolds of class i away from neighboring manifolds of other classes meanwhile pulls affine hulls inside the same manifold closer.

5, 6] is applied on a manifold. Based on the proof [3] that the Laplacian of a graph can be considered as a good approximation of Laplace-Beltrami operator defined on the manifold, many graph embedding algorithms have been proposed to linearly map a manifold from a high dimensional Euclidean space to a low dimensional subspace, such that the embedding is defined everywhere. However, these methods [3], [33] treated all the training samples from different classes with the same embedding and ignored that manifolds across classes might distribute very differently, i.e., there are large variations in the intrinsic structures across different classes. In view of this, we learn a class-specified distance metric $d_c^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{W}_c)$ for each class c . Then the proposed method aims to learn a collection of linear transformations $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$ from multiple manifolds.

There are N manifolds $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_N$ that constitute the whole training dataset $\mathbf{X} = [\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \dots, \mathbf{x}_{1,l_1}, \dots, \mathbf{x}_{N,1}, \mathbf{x}_{N,2}, \dots, \mathbf{x}_{N,l_N}]$. For nota-

tion simplicity, $\mathbf{X}$ is denoted as $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l]$ where $l = \sum_{i=1}^N l_i$. From the classification perspective, we expect that the learned transformations $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$ are able to maximize the scatterness between neighboring manifolds but from different classes, and minimize the compactness of manifolds from the same class simultaneously. Fig. 2 shows the basic idea of the learning procedure for the i th class. To achieve this goal, we maximize the discriminability of training data by solving the following objective function for the c th class:

$$\max_{\mathbf{W}_c} f_b(\mathbf{W}_c), \quad s.t. \quad f_w(\mathbf{W}_c) = 1, \quad (9)$$

where

$$f_b(\mathbf{W}_c) = \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_b^{(c)}(i, j), \quad (10)$$

$$f_w(\mathbf{W}_c) = \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_w^{(c)}(i, j), \quad (11)$$

respectively.

For the c^{th} class, the penalty graph $G_b^{(c)}$ that describes the between-class scatterness introduces a penalty of $\|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ is near to each other but from different classes, whereas the intrinsic graph $G_w^{(c)}$ encourages $\mathbf{W}_c^T \mathbf{x}_i$ and $\mathbf{W}_c^T \mathbf{x}_j$ to get closer if $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same class. The penalty graph $G_b^{(c)}$ and intrinsic graph $G_w^{(c)}$ are defined as:

$$G_b^{(c)}(i, j) = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|}{t_b}}, & \text{if } \mathbf{M}(\mathbf{x}_j) \in N_k(\mathbf{M}(\mathbf{x}_i)) \text{ where only } \mathbf{x}_i \in \text{class } c, \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

$$G_w^{(c)}(i, j) = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|}{t_w}}, & \text{if both } \mathbf{x}_i, \mathbf{x}_j \in \text{class } c, \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

where t_b and t_w are parameters of the heat kernels that follow an Gaussian distribution approximately in a local area. $\mathbf{M}(\mathbf{x}_i)$ denotes the manifold to which $\mathbf{x}_i$ belongs, and $N_k(\mathbf{M}(\mathbf{x}_i))$ denotes the manifolds that are the k nearest neighbors of manifold $\mathbf{M}(\mathbf{x}_i)$. In Eq. (13), when the value of t_w approximates zero (e.g., $t_w \leq 10^{-6}$), the heat kernel in intrinsic graph becomes increasingly localized and tends to the Dirac delta function, which makes that the intrinsic graph $G_w^{(c)}$ tends to an identity matrix. In this situation, the constraint matrix $G_w^{(c)}$ is not a Laplacian of the penalty graph but an identity matrix

for scale normalization.

For ease of analysis, we simplify $f_b(\mathbf{W}_c)$ in Eq. (10) as:

$$\begin{aligned}
f_b(\mathbf{W}_c) &= \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_b^{(c)}(i,j) \\
&= \text{tr} \left\{ \mathbf{W}_c^T \left[\sum_{i,j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_b^{(c)}(i,j) \right] \mathbf{W}_c \right\} \\
&= \text{tr} \left\{ \mathbf{W}_c^T \left[\sum_{i,j} \mathbf{x}_i G_b^{(c)}(i,j) \mathbf{x}_i^T - \sum_{i,j} \mathbf{x}_j G_b^{(c)}(i,j) \mathbf{x}_i^T \right. \right. \\
&\quad \left. \left. - \sum_{i,j} \mathbf{x}_i G_b^{(c)}(i,j) \mathbf{x}_j^T + \sum_{i,j} \mathbf{x}_j G_b^{(c)}(i,j) \mathbf{x}_j^T \right] \mathbf{W}_c \right\} \\
&= \text{tr} \left\{ \mathbf{W}_c^T \left[\sum_i \mathbf{x}_i D_b^{(c)}(i,i) \mathbf{x}_i^T - \mathbf{X} \mathbf{G}_b^{(c)} \mathbf{X}^T \right. \right. \\
&\quad \left. \left. - \mathbf{X} \mathbf{G}_b^{(c)} \mathbf{X}^T + \sum_j \mathbf{x}_j D_b^{(c)}(j,j) \mathbf{x}_j^T \right] \mathbf{W}_c \right\} \\
&= 2\text{tr} \left(\mathbf{W}_c^T \mathbf{X} \mathbf{D}_b^{(c)} \mathbf{X}^T \mathbf{W}_c - \mathbf{W}_c^T \mathbf{X} \mathbf{G}_b^{(c)} \mathbf{X}^T \mathbf{W}_c \right) \\
&= 2\text{tr} \left(\mathbf{W}_c^T \mathbf{X} \mathbf{L}_b^{(c)} \mathbf{X}^T \mathbf{W}_c \right), \tag{14}
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with $D_b^{(c)}(i,i) = \sum_j G_b^{(c)}(i,j)$ and $\mathbf{L}_b^{(c)} = \mathbf{D}_b^{(c)} - \mathbf{G}_b^{(c)}$ is the Laplacian of the penalty graph $\mathbf{G}_b^{(c)}$.

Similarly, $f_w(\mathbf{W}_c)$ in Eq. (11) can be simplified as:

$$\begin{aligned}
f_w(\mathbf{W}_c) &= \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_w^{(c)}(i,j) \\
&= 2\text{tr} \left(\mathbf{W}_c^T \mathbf{X} \mathbf{L}_w^{(c)} \mathbf{X}^T \mathbf{W}_c \right). \tag{15}
\end{aligned}$$

Thus, the optimal transformation matrix $\mathbf{W}_c$ can be derived by solving a generalized eigen-decomposition of matrices $\mathbf{L}_w^{(c)}$ and $\mathbf{L}_b^{(c)}$:

$$\mathbf{X} \mathbf{L}_b^{(c)} \mathbf{X}^T \mathbf{w} = \lambda \mathbf{X} \mathbf{L}_w^{(c)} \mathbf{X}^T \mathbf{w}. \tag{16}$$

The largest d eigenvectors $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$ that correspond to the largest d eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$, are chosen as the transformation matrix $\mathbf{W}_c$. On this basis, in the learning phase we compute $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$ for each class respectively.

3.4. Recognition

Having obtained a collection of discriminant parameter matrices $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$, the recognition procedure is conducted on a probe image set, whose manifold is denoted as $\mathbf{M}_p$. The probe set is first divided into m_p clusters by using a spectral clustering algorithm and then characterized as a set of affine hulls $\mathbf{M}_p = [\mathbf{h}_1^{(p)}, \mathbf{h}_2^{(p)}, \dots, \mathbf{h}_{m_p}^{(p)}]$. Similarly, a gallery set from class c is also described as $\mathbf{M}_g^{(c)} = [\mathbf{h}_1^{(c,g)}, \mathbf{h}_2^{(c,g)}, \dots, \mathbf{h}_{m_g}^{(c,g)}]$. To compare the probe manifold to the gallery manifold, we directly match them using the manifold-to-manifold distance metric associated with the parameter matrix $\mathbf{W}_c$:

$$d^{\text{mf}}(\mathbf{M}_p, \mathbf{M}_g) = \min_{\mathbf{h}_i^{(p)} \in \mathbf{M}_p} \min_{\mathbf{h}_j^{(c,g)} \in \mathbf{M}_g} d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c). \quad (17)$$

Then the probe manifold is assigned to a class that has the nearest distance to the probe.

To speed up the proposed algorithm, we let $\alpha = 1$ in Eq. (5). Thus the affine hull distance metric becomes:

$$d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c \mathbf{W}_c^T) = \frac{\mu_i^{(p)T} \mathbf{W}_c \mathbf{W}_c^T \mu_j^{(c,g)}}{(\mu_i^{(p)T} \mathbf{W}_c \mathbf{W}_c^T \mu_i^{(p)}) (\mu_j^{(c,g)T} \mathbf{W}_c \mathbf{W}_c^T \mu_j^{(c,g)})}, \quad (18)$$

where the mean vectors $\mu_j^{(c,g)}$ and $\mu_i^{(p)}$ are used as descriptors of affine hulls $\mathbf{h}_j^{(c,g)}$ and $\mathbf{h}_i^{(p)}$ respectively. Under such circumstances, the Euclidean distance between two mean vectors $\mu_i^{(p)}$ and $\mu_j^{(c,g)}$ can be used as an alternative affine hull distance metric:

$$d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c \mathbf{W}_c^T) = \sqrt{(\mu_i^{(p)} - \mu_j^{(c,g)})^T \mathbf{W}_c \mathbf{W}_c^T (\mu_i^{(p)} - \mu_j^{(c,g)})}. \quad (19)$$

4. Experiments

In this section, we evaluate the proposed MMML method on the task of face recognition based on image sets. Given a query face image set, its

nearest gallery set is determined by using a collection of parameter matrices $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$ and the nearest neighbor classifier.

4.1. Experimental Setup

Goals: Through the extensive experiments, we attempt to evaluate the performance of the proposed MMML method from different aspects to verify its effectiveness.

- In experiment 1, in order to demonstrate the effectiveness of using affine hull models, we compare the performance of the methods based on the affine hull model to that based on the linear subspace model (PCA model).
- In experiment 2, in order to show the effectiveness of using multiple class-specific subspaces, we test the performance of the methods based on multiple class-specific subspaces and one unified subspace respectively.
- In experiment 3, we illustrate the effect of the number of affine hulls on the recognition performance of the proposed MMML method.
- In experiment 4, we show the effect of the number of dimensions of each class-specific subspace on the recognition performance of the proposed MMML method.
- In experiment 5, we compare the proposed MMML method with five state of the arts on three widely studied face video databases to demonstrate the effectiveness of the proposed method.

Databases: Three widely studied face video databases, i.e., the Honda/UCSD database [44], the CMU MoBo database [45] and the Youtube Celebrities database [46], are considered in our experiments, where each video sequence is broken into frames and saved as an image set.

- **The Honda/UCSD Database:** The Honda/UCSD database [44] was built for evaluating algorithms on the video based face tracking and face recognition tasks. All the video sequences in this database were captured in indoor environment where the noises and the variations in lighting conditions were limited. However, the person in each video rotates his/her head and makes different expressions, which may

Figure 3: Some exemplar face images cropped from videos of three persons from the Honda/UCSD database.

lead to large variations in pose, expression, occlusion and scaling. The Honda/UCSD database consists of 59 videos from 20 persons. The original resolution of the frame of each video sequence is 640×480 pixels. We apply the Viola-Jones algorithm [47] to detect faces in each image set and resize each face image to 20×20 pixels as in [40, 48]. Some exemplar face images from the Honda/UCSD database are shown in Fig. 3.

- **CMU MoBo database:** The CMU Motion of body (MoBo) database [45] was created for biometric identification of humans based on the characteristics of subjects, which contains 96 video sequences from 24 different individuals and the number of frames for each video sequence is 300. For each person, there are 4 video clips corresponding to four walk patterns, i.e., slow walk, fast walk, incline walk and walk with a ball. This results in a large number of variations in the pose and expression of face images. By applying the Viola-Jones algorithm [47], the gray scale face images can be detected and cropped to the size of 40×40 pixels as described in [48]. Some exemplar face images from the CMU MoBo database are shown in Fig. 4. In our experiments, each histogram equalized face image is divided into patches of size 8×8 pixels and then a uniform Local Binary Pattern feature can be extracted from each patch as the local descriptor.
- **The Youtube Celebrities Database:** The YouTube Celebrities database [46] was built for evaluating algorithms on the tasks of face tracking and recognition. From the Youtube website, 1910 video sequences

Figure 4: Some exemplar face images cropped from videos of five persons from the CMU MoBo database.

were collected from 47 celebrities and each video sequence contained hundreds of frames. These videos were captured from famous persons including actors/actresses and politicians. Compared to Honda/UCSD and CMU MoBo databases where the videos were captured under controlled conditions, the videos in Youtube Celebrities database were obtained in real life environments where there exist very large variations in face appearances, poses and expressions, together with low quality, misalignments and occlusions. Thus the existing face detection algorithms often fail to work properly on this database. To overcome this problem, we apply an incremental visual tracking algorithm [49] to track faces in each frame of the Youtube Celebrities database, as mentioned in [48]. The first face cropped from each video sequence is provided by [46], and each gray scale face image is cropped and resized to 50×50 pixels. Some exemplar face images are illustrated in Fig. 5. To minimize the effect of illuminations, the histogram equalization is performed as a preprocessing step.

4.2. Experiment 1: Affine Hulls vs Linear Subspaces

In the proposed MMML method, we adopt the affine hulls instead of linear subspaces to represent a nonlinear manifold. In general, an affine hull is a more compact representation than a linear subspace from the geometric viewpoint. Thus the affine hull-based methods are expected to perform better than the linear subspace-based methods in terms of the recognition performance. To verify the advantage of the proposed MMML method, we

Figure 5: Some exemplar face images cropped from videos of four persons from the YouTube Celebrities database.

Table 1: Experimental results of experiment 1 on the Honda/UCSD database.

Methods	Average Results
Linear subspace-based method	87.2% $\pm$ 4.1%
Affine hull-based method	95.4% $\pm$ 3.3%

design experiment 1 as follows. From the Honda/UCSD database, we select 20 video sequences as gallery sets, with the remaining 39 videos as probe sets. For fair comparisons between the affine hull-based method and the linear subspace-based method, we adopt the same algorithm in each step except the modeling part, where each subset is modeled using the affine hulls and linear subspaces respectively. It is noted that this experiment does not involve any learning stage for reducing the computational complexity. Under such conditions, we randomly repeat the experiment 30 times. Then the averaged recognition results are given in Table 1 for the two methods respectively.

In Table 1, it is obvious that the affine hull-based method, which achieves a recognition rate of 95.4%, significantly outperforms the linear subspace-based method by around 8.2%.

4.3. Experiment 2: Multiple Class-Specific Subspaces vs One Unified Subspace

Different from the existing methods that only learn a unified subspace for all the training samples, the proposed MMML method learns a collection of class-specific subspaces $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_C$ based on the observation that the

Table 2: Experimental results of experiment 2 on the CMU MoBo database.

Methods	Average Results
Unified based method	95.8% $\pm$ 1.3%
Class-specific based method	98.6% $\pm$ 2.3%

manifolds in different classes distribute differently. To illustrate the effectiveness of the proposed method that utilizes multiple class-specific subspaces, we design experiment 2 as follows. From the CMU MoBo database, one image set per person is randomly selected for training, with the remaining three image sets used for testing. For fair comparisons between the class-specific subspace-based method and the unified subspace-based method, we adopt the same algorithm in each step except the learning part, where multiple class-specific subspaces and one unified subspace are learned respectively. With these settings, we repeat the experiment 30 times for each method. Then the averaged results of experiment 2 on the CMU MoBo database are listed in Table 2.

We can observe from Table 2 that the recognition performance of the proposed method based on multiple class-specific subspaces reaches 98.6%, which outperforms that of the method based on one unified subspace.

4.4. Experiment 3: The Effect of the Number of Affine Hulls

In the proposed MMML method, the spectral clustering algorithm is adopted to divide an image set into multiple clusters. The number of affine hulls, which is denoted as k , equals to the number of clusters. To investigate how the number of affine hulls affects the recognition performance, we conduct a series of experiments on Honda/UCSD database where 50 frames of each video are used. Since the number of frames of each video varies from 29 to 50, we test the MMML method with varying values of k (from 1 to 29). For each k , we repeat the experiment 30 times and the averaged recognition rates are recorded. Then the averaged recognition rate with regard to k is shown in Fig. 6.

From Fig. 6, it is observed that the performance of the proposed MMML method is robust to the choice of k . When $k=1$, the recognition rate of MMML method only reaches 70% since the whole image set is directly put into the discriminative learning stage without division. With an increase in k , the recognition rate of MMML first rises and then stabilizes around 89% when k varies from 4 to 29. Thus it is apparent that the proposed MMML

Figure 6: The recognition performance of the proposed MMML method with respect to the number of affine hulls k .

method is insensitive to k . Hence, there is a large space for us to choose a suitable number of affine hulls for the proposed MMML method.

4.5. Experiment 4: The Effect of the Dimensions of Each Subspace

For a better illustration of the robustness of the proposed MMML method, we evaluate its performance with an increase in the dimensions of each subspace in experiment 4. Similar to experiment 2, 96 face image sets from the CMU MoBo database are used. One image set per individual is randomly selected for learning, with the remaining three image sets per person used as the probe sets. We repeat the random selection 30 times for each subspace of fixed dimensions, and then increase the dimensions with step size 1. Then the averaged recognition rates are shown in Fig. 7 and Fig. 8, with respect to the number of dimensions of each discriminative subspace. Fig. 7 shows the recognition rates when the number of dimensions varies from 1 to 60. To better illustrate the effect of the dimension on the performance of the proposed method, Fig. 8 demonstrates the recognition rates when the number of dimensions varies from 1 to 12.

From Fig. 7 and Fig. 8, we observe that the performance of the proposed MMML method improves with an increase in the number of dimensions of each discriminative subspace. In Fig. 8, the recognition rate is only around

Figure 7: The recognition performance of the proposed MMML method with respect to the number of dimensions (from 1 to 60) of each discriminative subspace.

Figure 8: The recognition performance of the proposed MMML method with respect to the number of dimensions (from 1 to 12) of each discriminative subspace.

9.72% when the number of dimensions equals to one and quickly rises up to 37.5%, 72.22%, 81.94% and 84.72% with dimensions 2, 3, 4 and 5, respectively. However, when the number of dimensions keeps increasing beyond 12, the improvement in the recognition performance of the proposed MMML

method is limited, as shown in Fig. 7. This demonstrates a robustness of the proposed MMML method to the dimension of each subspace. Thus there is a large space to choose a suitable number of dimensions of each subspace for the proposed MMML method to work properly.

4.6. Experiment 5: Comparisons with Five State of the Arts

We compare the proposed MMML method with five existing image classification methods based on image sets, i.e., Affine hull-based Image Set Distance (AHISD) [9], Discriminant Canonical Correlation Analysis (DCC) [10], Manifold-to-Manifold Distance (MMD) [40], Manifold Discriminant Analysis (MDA) [33] and Sparse Approximated Nearest Points (SANP) [48]. The source codes of these five methods are provided by respective authors. For fair comparisons, the key parameters of each method are tuned and optimized as follows:

- 1) For DCC, PCA is applied to preserve 90% energy of each subspace such that the dimension of each subspace is 10 and the dimension of the embedding subspace is set to 150.
- 2) For MMD and MDA, 12 nearest neighbors are used to compute the geodesic distance and the ratio between geodesic distance and Euclidean distance is tuned to 1.1 to achieve the optimal performance. The other parameters are chosen following the experimental settings in the respective papers.
- 3) For MMD, the principal angles that are used in the manifold-to-manifold distance are the first 10 angles.
- 4) For SANP, the same weights are chosen as in [48] for convex optimization.
- 5) For the proposed MMML method, we adopt the affine hull-based distance metric defined in Eq. (18) and (19) to speed up the algorithm, where the affine hull distance is computed by using their mean vectors. For all these methods, we use the nearest neighbor classifier for recognition.

Experiments on the Honda/UCSD Database: We adopt the same configurations as in [9, 44, 48] where 20 video sequences are selected for training, with the remaining 39 videos for testing the classification performance. Similar to [48], we evaluate the proposed MMML and the existing methods in the following cases: 50 frames selected from each video sequence, 100 frames selected from each video sequence, and full length of each video sequence. The number of k nearest affine hulls is set to neighboring 7. Then the experimental results are summarized and shown in Table 3.

From Table 3, it is apparent that the proposed MMML method achieves the best recognition rate 93.1% on average. With a discriminative learning procedure, DCC, MDA and the proposed MMML outperform MMD and

Table 3: Experimental results on the Honda/UCSD database.

Methods	50 Frames	100 Frames	Full Length	Average Results
DCC	79.5%	92.3%	94.9%	88.9%
MMD	82.1%	84.6%	94.9%	87.2%
MDA	76.9%	89.7%	100%	88.9%
AHISD	87.2%	84.6%	89.7%	87.2%
SANP	84.6%	92.3%	100%	92.3%
MMML	89.7%	92.3%	97.22%	93.1%

AHISD, which directly compare the models of image sets without any learning procedure. Although SANP does not involve any discriminative learning steps, it reconstructs a sample point using sparse affine coefficients that introduce the discriminative information into the sparse representations, thus SANP is also able to achieve a comparable performance of 92.3%. With a decrease in the length of the frames, i.e., 50 frames in each image set, the performance of MDA and DCC degrades heavily because there is not sufficient training samples for discriminative learning. On the other hand, MMML describes an image set using a collection of affine hulls that can “fill-in” the missing data, thus a reasonably good performance can be achieved, even with less training samples.

Experiments on the CMU MoBo Database: We adopt the following configurations. One image set per person is randomly selected for training, with the remaining three image sets for testing. We repeat the random selection 30 times for each method and the averaged results are illustrated in Table 4.

Table 4: Experimental results on the CMU MoBo database.

Methods	Average Results
DCC	94.9% $\pm$ 2.3%
MMD	94.4% $\pm$ 3.0%
MDA	95.8% $\pm$ 1.9%
AHISD	94.4% $\pm$ 1.9%
SANP	97.8% $\pm$ 1.1%
MMML	98.6% $\pm$ 2.5%

From Table 4, we observe that the proposed MMML achieves 98.6%,

which is the best performance of all methods. Compared to the Honda/UCSD database, the CMUMoBo database contains more noises and the proposed MMML method performs better than others in that it learns a class-specific metric that is able to better describe a noisy image set. Although a comparably good performance can be achieved by SANP that is only slightly inferior to MMML, SANP requires a much higher computational complexity. Similarly, without any learning procedure, MMD and AHISD perform not as well as the discriminative learning based MDA, DCC, SANP and MMML, where SANP employs sparse coefficients that introduce discriminative information.

Experiments on the YouTube Celebrities Database: We adopt the same configurations as in [48] where five-fold cross validation experiments are conducted on this database. We divide video sequences of each person into 5 groups and make sure that each group contains 9 videos (with minimum overlapping). In this way, the YouTube Celebrities Database are divided into 5 folds, each consists of 423 video sequences from 47 subjects. In each fold, we select 3 image sets per person randomly for training, with the remaining 6 image sets for testing. We repeat the random selection 10 times for each method and show the averaged recognition rates in Table 5.

In Table 5, the performance achieved by all set-based image classification methods is degraded on the YouTube Celebrities database, where the videos are obtained in real life scenarios with low quality, misalignments and large variations in appearances, poses and expressions. In such a challenging condition, the recognition rate achieved by MMML reaches 69.4%, which outperforms all other methods. Since the single affine hull model can not exactly describe an noisy image set of low quality and large appearance variations, the performance achieved by SANP on the Youtube Celebrities database is not as good as that on the Honda/CUCSD and CMU MoBo databases.

Table 5: Experimental results on the YouTube Celebrities database.

Methods	Average Results
DCC	62.2% $\pm$ 2.1%
MMD	63.6% $\pm$ 2.3%
MDA	61.9% $\pm$ 3.4%
AHISD	64.2% $\pm$ 4.1%
SANP	63.9% $\pm$ 2.3%
MMML	69.4% $\pm$ 2.8%

Table 6: Time consumption of different methods on the Honda/UCSD database (classification of one image set).

Methods	Training	Testing
DCC	6.50s	0.23s
MMD	<i>N/A</i>	30.10s
MDA	8.90s	0.18s
AHISD	<i>N/A</i>	5.20s
SANP	<i>N/A</i>	2.80s
MMML	60.08s	0.13s

For both MAD and DCC, only a single discriminative subspace is learned for all the samples from different classes, this leads to that the large variations and noises can not be covered in this single subspace. Thus MDA and DCC demonstrate relatively low recognition rates on the YouTube Celebrities database. In MMML, the proposed affine hull-based distance metric is designed for each class such that the variations of the current class are well characterized by the class-specific model. This leads to a robust performance of MMML across the Honda/UCSD, CMU MoBo and YouTube Celebrities databases.

4.7. Discussions

Computational Complexity: In Table 6, we compare the time consumption by different methods on the Honda/UCSD database. We can see that MMD spends the longest time 30.1 seconds in the testing step of classifying one image set, because two operations, i.e., constructing a collection of linear subspaces through PCA and matching image sets by computing their principal angles, are very time-consuming. From the recognition rates shown in Table 3 and Table 4, we observe that a comparable performance can be achieved by SANP and MMML, but SANP requires more time (2.8 seconds) in classifying one image set than our MMML method (0.13 seconds), as shown in Table 6. Although the proposed MMML needs to learn a distance metric for each class, where more time is consumed for training, this step can be finished offline without affecting the time required for classifying each probe image set.

Sample Size: All methods assume that there are sufficient images in each image set such that a set of affine hulls can be used to approximate the manifold. Questions arise when there are not sufficient samples in some

image sets, then the number of affine hulls of each manifold will be reduced. Furthermore, for each subset of a manifold, there are two key factors to effectively describe its affine hull: mean vector μ and orthonormal bases $\mathbf{W}$. If there are not sufficient samples in a subset, the mean vector is able to better describe a subset than the orthonormal bases. In view of this, we assign a larger weight to the mean vector and a smaller weight to the orthonormal bases, such that the proposed method is able to adapt to different sizes of image sets.

5. Conclusion

In this paper, we propose a multi-manifold metric learning (MMML) method for face recognition based on image sets. Firstly, an image set is modeled as a nonlinear manifold that is divided into several subsets by using a spectral clustering algorithm. Then each subset is described by an affine hull such that a nonlinear manifold can be represented using a collection of affine hulls. On this basis, multiple person-specific distance metrics can be learned to maximize the scatterness of neighboring manifolds from different classes and minimize the compactness of manifolds from the same class simultaneously, such that more discriminative information can be exploited for classification. Experimental results on three popular databases are presented to show the effectiveness and robustness of our proposed method.

References

- [1] M. Turk, A. Pentland, Eigenfaces for recognition, *Journal of Cognitive Neuroscience* (1991) 71–86.
- [2] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on PAMI* (1997) 711–720.
- [3] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, *IEEE Transactions on PAMI* (2005) 328–340.
- [4] J. Tenenbaum, V. De Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* (2000) 2319–2323.
- [5] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* (2000) 2323–2326.

- [6] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, NIPS (2001) 585–591.
- [7] D. Xu, S. Yan, D. Tao, S. Lin, H.-J. Zhang, Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval, IEEE Transactions on IP 16 (11) (2007) 2811–2821.
- [8] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on PAMI 29 (1) (2007) 40–51.
- [9] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: CVPR, 2010, pp. 2567–2573.
- [10] T. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on PAMI (2007) 1005–1018.
- [11] S. Zhou, R. Chellappa, From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space, IEEE Transactions on PAMI (2006) 917–929.
- [12] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, in: CVPR, 2005, pp. 581–588.
- [13] G. Shakhnarovich, J. Fisher, T. Darrell, Face recognition from long-term observations, ECCV (2002) 851–865.
- [14] O. Arandjelović, R. Cipolla, An information-theoretic approach to face recognition from face motion manifolds, Image and Vision Computing (2006) 639–647.
- [15] S. Satoh, Comparative evaluation of face sequence matching for content-based video access, in: FG, 2000, pp. 163–168.
- [16] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: CVPR, 2011, pp. 529–534.
- [17] R. Duda, P. Hart, D. Stork, Pattern Classification, 2001.

- [18] L. Wang, X. Wang, J. Feng, Subspace distance analysis with application to adaptive bayesian algorithm for face recognition, *Pattern Recognition* (2006) 456–464.
- [19] F. Li, Q. Dai, W. Xu, G. Er, Weighted subspace distance and its applications to object recognition and retrieval with image sets, *IEEE Signal Processing Letters* (2009) 227–230.
- [20] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: *FG*, 1998, pp. 318–323.
- [21] X. Ning, G. Karypis, The set classification problem and solution methods, in: *Data Mining Workshops, 2008. ICDMW’08. IEEE International Conference on*, IEEE, 2008, pp. 720–729.
- [22] J. Louradour, H. Larochelle, Classification of sets using restricted boltzmann machines, *arXiv preprint arXiv:1103.4896*.
- [23] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, *Robotics Research* (2005) 192–201.
- [24] Y.-C. Chen, V. M. Patel, P. J. Phillips, R. Chellappa, Dictionary-based face recognition from video, in: *Computer Vision—ECCV 2012*, Springer, 2012, pp. 766–779.
- [25] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, *Journal of Machine Learning Research* (2003) 913–931.
- [26] T. Kozakaya, O. Yamaguchi, K. Fukui, Development and evaluation of face recognition system using constrained mutual subspace method, *IPSJ* (2004) 951–959.
- [27] J. Hamm, D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: *ICML*, 2008, pp. 376–383.
- [28] J. Lu, Y.-P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35 (1) (2013) 39–51.
- [29] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, P. J. Phillips, Video-based face recognition via joint sparse representation, in: *Automatic*

- Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
- [30] P. Zhu, W. Zuo, L. Zhang, S. Shiu, D. Zhang, Image set based collaborative representation for face recognition, *Information Forensics and Security*, IEEE Transactions on.
 - [31] E. G. Ortiz, A. Wright, M. Shah, Face recognition in movie trailers via mean sequence sparse representation-based classification, in: *Computer Vision and Pattern Recognition (CVPR)*, 2013 IEEE Conference on, IEEE, 2013, pp. 3531–3538.
 - [32] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern classification*, Wiley-interscience, 2012.
 - [33] R. Wang, X. Chen, Manifold discriminant analysis, in: *CVPR*, 2009, pp. 429–436.
 - [34] E. P. Xing, A. Y. Ng, M. I. Jordan, S. J. Russell, Distance metric learning with application to clustering with side-information, in: *NIPS*, 2002, pp. 505–512.
 - [35] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: *NIPS*, 2005.
 - [36] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, *JMLR* 10 (2009) 207–244.
 - [37] J. V. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon, Information-theoretic metric learning, in: *ICML*, 2007.
 - [38] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: *CVPR*, 2012.
 - [39] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: *CVPR*, 2012.
 - [40] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: *CVPR*, 2008, pp. 1–8.

- [41] U. Von Luxburg, A tutorial on spectral clustering, *Statistics and computing* 17 (4) (2007) 395–416.
- [42] A. Y. Ng, M. I. Jordan, Y. Weiss, et al., On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems* 2 (2002) 849–856.
- [43] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, California, USA, 1967, p. 14.
- [44] K. Lee, J. Ho, M. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: *CVPR*, 2003, pp. 313–220.
- [45] R. Gross, J. Shi, The cmu motion of body (mobo) database.
- [46] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: *CVPR*, 2008, pp. 1–8.
- [47] P. Viola, M. Jones, Robust real-time face detection, *IJCV* (2004) 137–154.
- [48] Y. Hu, A. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: *CVPR*, 2011, pp. 121–128.
- [49] D. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, *IJCV* (2008) 125–141.