# Multiview Metric Learning with Global Consistency and Local Smoothness

DEMING ZHAI, Harbin Institute of Technology
HONG CHANG, SHIGUANG SHAN, and XILIN CHEN, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
WEN GAO, Harbin Institute of Technology

In many real-world applications, the same object may have different observations (or descriptions) from multiview observation spaces, which are highly related but sometimes look different from each other. Conventional metric-learning methods achieve satisfactory performance on distance metric computation of data in a single-view observation space, but fail to handle well data sampled from multiview observation spaces, especially those with highly nonlinear structure. To tackle this problem, we propose a new method called *Multiview Metric Learning with Global consistency and Local smoothness* (MVML-GL) under a semisupervised learning setting, which jointly considers global consistency and local smoothness. The basic idea is to reveal the shared latent feature space of the multiview observations by embodying global consistency constraints and preserving local geometric structures. Specifically, this framework is composed of two main steps. In the first step, we seek a global consistent shared latent feature space, which not only preserves the local geometric structure in each space but also makes those labeled corresponding instances as close as possible. In the second step, the explicit mapping functions between the input spaces and the shared latent space are learned via regularized locally linear regression. Furthermore, these two steps both can be solved by convex optimizations in closed form. Experimental results with application to manifold alignment on real-world datasets of pose and facial expression demonstrate the effectiveness of the proposed method.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Metric learning, multiview learning, global consistency, local smoothness

**ACM Reference Format:**
Zhai, D., Chang, H., Shan, S., Chen, X., and Gao, W. 2012. Multiview metric learning with global consistency and local smoothness. ACM Trans. Intell. Syst. Technol. 3, 3, Article 53 (May 2012), 22 pages.
DOI = 10.1145/2168752.2168767 http://doi.acm.org/10.1145/2168752.2168767

## 1. INTRODUCTION

Metric learning plays a crucial role in the computer vision and pattern recognition community. Many tasks, such as image classification, clustering, content-based image

This paper is partially supported by the Natural Science Foundation of China under contracts No. 60803084 and No. 60872077, and the National Basic Research Program of China (973 Program) under contract 2009CB320902.

Author's addresses: D. Zhai (corresponding author) and W. Gao, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; emails: {dmzhai, wgao}@jdl.ac.cn; H. Chang, S. Shan, and X. Chen, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China; emails: {hchang, sgshan, xlchen}@jdl.ac.cn.

DOI 10.1145/2168752.2168767 http://doi.acm.org/10.1145/2168752.2168767

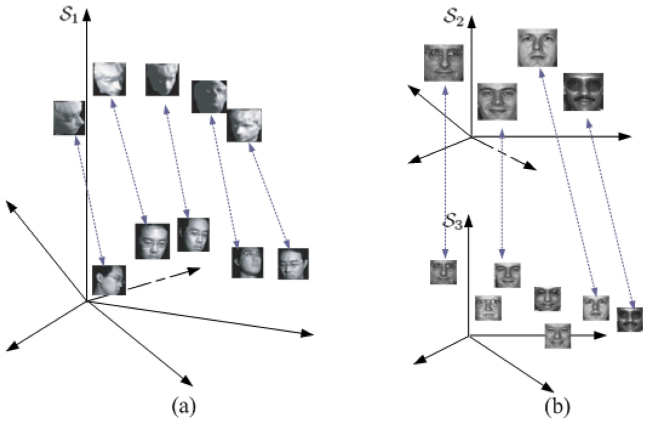


Fig. 1. Illustration of multiview metric learning problem. (a) Pose alignment of two observation sets in the same dimensional space; (b) low-high resolution face matching in two different dimensional spaces.

annotation and retrieval [Wu et al. 2011] depend critically on the choice of an appropriate distance metric. In the literature, various metric-learning methods have been proposed in the past few years [Bar-Hillel et al. 2003; Chang and Yeung 2007; Davis et al. 2007; Jin et al. 2009; Liu et al. 2010; Weinberger et al. 2006; Xing et al. 2003; Yeung et al. 2008; Zhan et al. 2009]. Nevertheless, these algorithms all deal with data lying in a single-view observation space.

In many real-world applications, the same object (e.g., face, pose) may have different observations (or descriptions) from multiple views which are highly related but sometimes look different from each other, for example, facial expression recognition for different people, pose estimation with image sequences from different objects[1], object recognition with pictures from different camera angles, and identity recognition with video and audio streams. Recently, some new applications have emerged, including face matching with near infrared (NIR) images and visual (VIS) ones [Lei and Li 2009], and the alignment of face images with different resolutions [Li et al. 2009]. All these applications naturally bring about a new problem: how to compute the distance metric or measure the relationship between multiview observations, which is referred to as *multiview metric learning*. To the best of our knowledge, Zheng et al. [2010] is the only previous work on multiview metric learning.

Multiview metric learning is a challenging task, since different observations may locate in the same or different dimensional space(s). As illustrated in Figure 1(a), two image sequences with similar pose variations are located in the same space (more generally, they can be in different dimensional spaces). Considering that the appearance of each view is disparate, it is difficult to produce optimal comparison by regarding multiple observations as one view and measure their relationship using single-view metric-learning algorithms. In the application of low-high resolution face image matching, as shown in Figure1(b), low-resolution images and high-resolution ones form two-view observation spaces with different dimensions. In such situations, even single-view metric-learning algorithms cannot work, since the dimensions of multiview observations are different.

In the literature, some algorithms have been proposed to measure the relationship between two related datasets. Canonical Correlation Analysis (CCA) [Hotelling 1936] is one of the most popular statistical methods correlating linear relationships between

[1] For this example, pose is the estimation target, while different objects correspond to different views.

two variables. It has been applied to a wide range of real-world applications with great success [Hardoon et al. 2004]. However, CCA is a global and linear methodology and fails to deal with data involving complex nonlinear correlation. Kernel CCA (KCCA) [Akaho 2001] extends the nonlinear processing ability by using the so-called "kernel trick". Non-Consolidating Correlation Analysis (NCCA) [Ek et al. 2008] also extends CCA by learning additional non-shared transformations for each view. NCCA comprises two steps: applying CCA to find shared embedded data and applying NCCA to find non-shared embedded data. In spirit, the shared embedding space learning in the first step is in the same as that in CCA. CCA, KCCA, and NCCA all deal with correlation in a global way. The advantage of these methods is that metric learning could be formulated as convex optimization problems with no local optima and could be solved using efficient algorithms. More recently, the work in Zheng et al. [2010] learns global Mahalanobis distance metrics, which extends Neighborhood Components Analysis (NCA) [Goldberger et al. 2004] to a multiview setting. Since the cost function in Zheng et al. [2010] is not convex, there is no guarantee of obtaining the global optimal solutions with gradient computation. In a nutshell, these algorithms all ignore the details of the local structures.

Actually, as pointed out by Bottou and Vapnik [1992], it is usually not easy to find a unique function which holds good predictability in the entire data space. Vapnik [1995] further demonstrates that the local learning algorithms usually achieve lower empirical errors than global ones. This is because nearby instances are more likely generated by the same data-generation model, while far away instances tend to differ in it. Accordingly, neighboring instances may have the same or similar distance metric, while for instances lying away in different neighboring spaces, distance metrics change heavily. In addition, it is proposed in Wu et al. [2007] that learning in a local manner can sufficiently boost capacity powers.

Some studies have been devoted to distance metric learning considering local information. Local Linear Embedding (LLE) [Saul et al. 2003] and Locality Preserving Projections (LPP) [He and Niyogi 2003] are two dimensionality reduction methods, which can be seen as learning distance metrics by preserving local geometric structures. Yang et al. [2006] propose another efficient algorithm for local distance metric learning in a probabilistic framework. However, LLE does not give explicit metrics for unseen data, while LPP and Yang's method are still global in the sense that the same transformation is applied to all instances during metric learning. Frome et al. [2006] propose a method to learn distinctive distance functions for different instances as a combination of elementary distances between patch-based visual features; they then extend it to enable the comparison between them [Frome et al. 2007]. Since label information is necessary in the learning process, the method could only compute local distance for labeled data. Chang and Yeung [2007] also propose a method called locally smooth metric learning, in which the learned local metrics vary smoothly and could preserve the intraclass topological structure of the data. However, a heuristic initialization stage is needed for setting the target locations of all labeled instances. Zhan et al. [2009] address the local metric learning problem using metric propagation under a transductive setting, which cannot generalize to new test data.

These local methods are all proposed for metric learning in a single-view observation space. Although the idea of local learning has been applied on multiview scenarios, current work mainly focuses on transductive classification [Wu and Schölkopf 2007], clustering [Wang et al. 2007; Wu and Schölkopf 2006], and dimensionality reduction [Wu et al. 2007]. As far as we know, there is little work dedicating to the topic of metric learning with local methodology on multiview observation spaces in the literature.

Inspired by the idea of "thinking globally and fitting locally" [Saul et al. 2003], in this article, we propose a novel method called Multiview Metric Learning with Global consistency and Local smoothness (MVML-GL) under a semi-supervised setting. Similar ideas have been adopted elsewhere, such as large margin classifiers [Huang et al. 2004], in which the model learns the decision boundary by considering the data in both a local and a global fashion. The basic idea of our method is to reveal the shared latent space of the multiview observations by embodying global consistency constraint and preserving local geometric structure. Our method decomposes the multiview metric learning as a two-step approach: graph-based embedding for multiview observations and regression-based mapping functions learning for each instance. Specially, in the first step, our method seeks a shared latent feature space to establish the relationship between data from multiview observation spaces according to the labeled instances pairs. In the second step, the explicit mapping functions between the input spaces and the shared latent space are learned via regularized locally linear regression, which allows different local metrics to be learned at different locations of each input space. Furthermore, the graph-Laplacian regularization term is incorporated to keep the learned metric varying smoothly. It should be noted that the first step is globally consistent, which simultaneously considers geometric structures contained in each view and connections between data from different views, and the second step is locally smooth, which enables each instance to have its own specific distance metric instead of applying a uniform one for all instances. In addition, the two steps can both be formulated as convex optimization problems with closed-form solutions.

The contribution of the article is highlighted in the following.

- This article proposes a robust and flexible multiview metric-learning method by jointly considering global consistency and local smoothness.
- The proposed method formulates global and local metric learning as two convex optimization problems which could be efficiently solved with closed-form solutions.
- The essential connection between multiview metric learning and manifold alignment is discussed.

The rest of this article is organized as follows. In Section 2, we present the proposed multiview metric-learning method in details. In Section 3, we discuss the relationship between multiview metric learning and manifold alignment. Section 4 shows the experimental results with application to manifold alignment for pose estimation and facial expression recognition. Finally, Section 5 gives some concluding remarks and discussions about future work.

## 2. MULTIVIEW METRIC LEARNING

The target of the multiview metric-learning method is to compute distance metrics between samples from multiple observation datasets. Without loss of generality, we take two-view observation spaces for example, and our algorithm could be naturally extended to metric learning for multiview observations, as depicted in Section 2.4.

Since two-view datasets are from different observations of the same object, they are highly related to each other. It is reasonable to assume that some common features across both spaces can be represented in a shared latent feature space where the intrinsic relationship of the two datasets is revealed. As a consequence, we resort to a shared latent feature space to establish the relationship among different observations. Although sometimes the two views may seem to be quite different in nature, the intrinsic connection between them still exists, since they are from the same object. Some well known works, such as CCA [Hotelling 1936] and KCCA [Akaho 2001], also exploit the assumption on the intrinsic relationship of data from different views, though it may be not suitable for all real-world cases. These methods have been successfully applied in many challenging applications. When the two views are quiet different, the performance of alignment from the low-dimension embeddings may also degrade heavily. In this scenario, we need more labeled corresponding pairs and better feature representation for each view in order to improve the accuracy of alignment.

Similar to spectral regression [Cai et al. 2007], we would accomplish multiview metric learning by two steps: (1) get the common low-dimensional embeddings for all labeled correspondence pairs, and (2) learn the relationships between the input space of each observation and the shared latent space for unlabeled and test data.

In the following sections, we first claim some notations and present our method in two steps: the globally consistent shared latent feature space learning and locally smooth multiview metric learning. Then, the overall algorithm is summarized, followed by a short discussion. In the end, we perform a time complexity analysis and provide several alternative ways to improve for real applications.

### 2.1. Notations

Let us represent two datasets $\mathcal{G}$, $\mathcal{P}$ with their matrix forms as $\mathbf{X}^g = [\mathbf{x}_1^g, \mathbf{x}_2^g, \cdots \mathbf{x}_{N_g}^g]$ and $\mathbf{X}^p = [\mathbf{x}_1^p, \mathbf{x}_2^p, \cdots \mathbf{x}_{N_p}^p]$, where the column vector $\mathbf{x}_i^g \in \mathbb{R}^{D_g}$ ($\mathbf{x}_j^p \in \mathbb{R}^{D_p}$) denotes a data point in the $D_g(D_p)$ dimensional input space. Suppose $K$ correspondence pairs are given in set $\mathcal{C}$, that is, $(i, j) \in \mathcal{C}$ if $\mathbf{x}_i^g$ corresponds to $\mathbf{x}_j^p$. With the prior knowledge, the training sets can be separated into labeled and unlabeled subsets, denoted as $\mathcal{G} = \{\mathcal{G}^l, \mathcal{G}^u\}$ and $\mathcal{P} = \{\mathcal{P}^l, \mathcal{P}^u\}$. Let $\mathbf{x}_i$ ($\mathbf{x}_i^p$ or $\mathbf{x}_i^g$) be an unlabeled or a new test data point either from set $\mathcal{P}$ or set $\mathcal{G}$, $\mathcal{S}^l(\mathbf{x}_i)$, and let $\mathcal{S}^u(\mathbf{x}_i)$ indicate the labeled and unlabeled datasets which are derived from the same input space as $\mathbf{x}_i$, that is, if $\mathbf{x}_i \in \mathcal{P}$, then $\mathcal{S}^l(\mathbf{x}_i) = \mathcal{P}^l$, $\mathcal{S}^u(\mathbf{x}_i) = \mathcal{P}^u$. In addition, we denote the unknown common embeddings of $\mathbf{X}^g$ and $\mathbf{X}^p$ by $\mathbf{Y}^g = [\mathbf{Y}_g^l, \mathbf{Y}_g^u]$ and $\mathbf{Y}^p = [\mathbf{Y}_p^l, \mathbf{Y}_p^u]$, respectively. $\mathbf{Y}^l = [\mathbf{Y}_g^l, \mathbf{Y}_p^l]$ indicates the low-dimensional embeddings for all labeled data.

### 2.2. Globally Consistent Shared Latent Feature Space Learning

In order to derive the common low-dimensional embeddings for all labeled correspondence pairs, two important issues should be taken into consideration: (1) the embedded correspondence pairs should be as close as possible and (2) the common embeddings should preserve the local geometric structures in all original input spaces. In this way, we also consider local smoothness for labeled data.

We achieve the overall objective by minimizing the following energy function:

$$\mathcal{J}(\mathbf{Y}^l) = 2 \sum_{(i,j)\in\mathcal{C}} ||\mathbf{y}_i^g - \mathbf{y}_j^p||^2 + \sum_{i,j\in\mathcal{G}^l} s'_{ij}||\mathbf{y}_i^g - \mathbf{y}_j^g||^2 + \sum_{i,j\in\mathcal{P}^l} s''_{ij}||\mathbf{y}_i^p - \mathbf{y}_j^p||^2, \tag{1}$$

where $s'_{ij}$ ($s''_{ij}$) defines the edge weight of the corresponding graph for each dataset. With the prior knowledge hidden in available labeled correspondence pairs, the two graphs can be connected by adding some edges between related vertices. Let $\mathbf{S}$ be the weight matrix of the connected graph, and each entry in $\mathbf{S}$ is defined as

$$s_{ij} = \begin{cases} 1 & \text{, if } (i, j) \in \mathcal{C} \\ e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma_i \sigma_j}} & \text{, if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i), \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

where $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of $k$-nearest-neighbors of $\mathbf{x}_i$, $\sigma_i$ ($\sigma_j$) is the distance from $\mathbf{x}_i$ ($\mathbf{x}_j$) to its $m$th nearest neighbor, and $m$ is a local scale factor. This way of defining the adjacency matrix is called local scaling [Zelnik-Manor and Perona 2005], and it controls the decreasing speed of $s_{ij}$ with the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Note that the local parameters $k$ and $m$ control the values of $s'_{ij}$ and $s''_{ij}$ and further influence the relative contribution of local structure preserving in each observation space. Accordingly, no additional turning parameters are introduced to balance the three terms in Eq. (1). With these notations, the objective function can be simplified and rewritten as

$$\begin{aligned} \mathcal{J}(\mathbf{Y}^l) &= \sum_{i,j\in\mathcal{G}^l\cup\mathcal{P}^l} s_{ij}||\mathbf{y}_i - \mathbf{y}_j||^2 \\ &= 2\left(\sum_{i,j\in\mathcal{G}^l\cup\mathcal{P}^l} \mathbf{y}_i{}^T s_{ij}\mathbf{y}_i - \sum_{i,j\in\mathcal{G}^l\cup\mathcal{P}^l} \mathbf{y}_i{}^T s_{ij}\mathbf{y}_j\right) \\ &= 2\mathrm{Tr}\left(\mathbf{Y}^l\mathbf{L}(\mathbf{Y}^l)^T\right), \end{aligned} \tag{3}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is called the Laplacian matrix, and $\mathbf{D}$ is a diagonal matrix with the entries $\mathbf{D}_{ii} = \sum_j s_{ij}$. In the literature a similar idea of using graph Laplacian for multiple-view data can be found in Sindhwani and Niyogi [2005]. However, our method aims to find a shared latent subspace for local multiview metric learning, which is different from that work.

The objective function gives a high penalty when the labeled correspondence pairs from different spaces or neighboring data points in the same space are mapped far apart. Therefore, it tries to find a shared latent feature space which reflects the intrinsic relationship between multiview observations and preserves the local geometric structures of data in each input space. For the preceding optimization problem, we set a constraint, $(\mathbf{Y}^l)^T\mathbf{D}\mathbf{Y}^l = \mathbf{I}$, to avoid trivial solutions. Then, the minimization problem is then formulated as

$$\mathbf{Y}^l = \arg\min \mathrm{Tr}\left(\frac{\mathbf{Y}^l\mathbf{L}(\mathbf{Y}^l)^T}{\mathbf{Y}^l\mathbf{D}(\mathbf{Y}^l)^T}\right). \tag{4}$$

Finally, the common low-dimensional embeddings for all labeled data can be obtained by solving a generalized eigenvector problem, which is expressed as

$$\mathbf{L}\mathbf{Y}^l = \lambda\mathbf{D}\mathbf{Y}^l. \tag{5}$$

### 2.3. Locally Smooth Multiview Metric Learning

*2.3.1. Multiview Metric Definition.* The embedding results obtained in the preceding section are only the solutions for labeled training data. To map the unlabeled and new test data into the shared latent feature space, we turn to learn the mapping function $f_i$ between $\mathbf{y}_i$ and $\mathbf{x}_i$, that is, $\mathbf{y}_i = f_i(\mathbf{x}_i)$. Specially, we consider a linear transformation function $f_i(\cdot; \mathbf{W}_i, \mathbf{b}_i)$ defined as $f_i(\mathbf{x}_i) = \mathbf{W}_i^T\mathbf{x}_i + \mathbf{b}_i$, where $\mathbf{W}_i$ is a transformation matrix, and $\mathbf{b}_i$ is a translation vector. One often deals with the bias term $\mathbf{b}_i$ by appending each instance with an additional dimension $\mathbf{x}_i \Leftarrow [\mathbf{x}_i; 1]$, $\mathbf{W}_i \Leftarrow [\mathbf{W}_i; \mathbf{b}_i^T]$. Then the linear transformation function becomes

$$f_i(\mathbf{x}_i) = \mathbf{W}_i^T\mathbf{x}_i. \tag{6}$$

Note that the transformation function $f$ is defined for each individual data point but not shared by all data points in each input space globally.

For any two data points $\mathbf{x}_i^g \in \mathbb{R}^{D_g}$ and $\mathbf{x}_j^p \in \mathbb{R}^{D_p}$ from different observation sets, the distance metric between them is defined as their Euclidean distance in the shared latent feature space, which can be formulated as

$$\begin{aligned} d\left(\mathbf{x}_i^g, \mathbf{x}_j^p\right) &\doteq \left(\mathbf{y}_i^g - \mathbf{y}_j^p\right)^T\left(\mathbf{y}_i^g - \mathbf{y}_j^p\right) \\ &= \left(\left(\mathbf{W}_i^g\right)^T\mathbf{x}_i^g - \left(\mathbf{W}_j^p\right)^T\mathbf{x}_j^p\right)^T\left(\left(\mathbf{W}_i^g\right)^T\mathbf{x}_i^g - \left(\mathbf{W}_j^p\right)^T\mathbf{x}_j^p\right). \end{aligned} \tag{7}$$

It's worth noting that, in most cases, $\mathbf{W}_i^g$ is not equal to $\mathbf{W}_j^p$ for the multiview metric-learning problem, since using the same projection for the two types of data would not produce optimal comparison. Furthermore, the dimensions of the two projections would not be equal in the case that the data $\mathbf{x}_i^g$ and $\mathbf{x}_j^p$ are in different dimension spaces. When $\mathbf{W}_i^g = \mathbf{W}_j^p = \mathbf{W}$, the multiview distance metric defined previously will degrade to a traditional Mahalanobis distance metric defined as $d(\mathbf{x}_i^g, \mathbf{x}_j^p) = (\mathbf{x}_i^g - \mathbf{x}_j^p)^T \mathbf{M} (\mathbf{x}_i^g - \mathbf{x}_j^p)$, where $\mathbf{M} = \mathbf{W}\mathbf{W}^T$.

*2.3.2. Locally Smooth Metric Learning.* It is desirable that two adjacent data points in each observation space should be mapped to close locations in the shared latent feature space. In other words, the mapping functions should vary smoothly in the original input spaces so that the datasets after metric learning can preserve the original local geometry structure. Motivated by this expectation, we learn the mapping function via regularized locally linear regression.

Without loss of generalization, let $\mathbf{x}_i$ ($\mathbf{x}_i^p$ or $\mathbf{x}_i^g$) be an unlabeled or a new test data point either from set $\mathcal{P}$ or set $\mathcal{G}$, then the optimal local affine transformation $f_i$ for $\mathbf{x}_i$ is computed by minimizing the following objective function:

$$\mathcal{J}(f_i) = \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} ||\mathbf{y}_j - f_i(\mathbf{x}_j)||^2 + \lambda ||f_i||^2. \tag{8}$$

The first term in the objective function is the well known moving least square [Levin 1998]; the second term is the shrinkage constraint, also known as the Tikhonov regularizer [Hastie et al. 2001], which helps to improve the generalization of the solutions. The parameter $\lambda$ is used to balance the two terms. For moving the least square term, $\theta_{ij}$ is called the moving weight, which reflects the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ and is defined as $\theta_{ij} = 1/(\| \mathbf{x}_i - \mathbf{x}_j \|^2 + 1_{\mathbf{x}_i = \mathbf{x}_j} \cdot \varepsilon)$, where $1_{a=b}$ is an indicator function that takes the value 1 if $a = b$ and 0 otherwise, and $\varepsilon$ is a small adjustment value to prevent the denominator from degenerating to 0. $\mathbf{y}_j$ is the low-dimensional embedding for labeled data $\mathbf{x}_j$, which has already been obtained in Section 2.2.

The optimal transformation function $f_i$ is found by projecting itself onto the labeled data samples and minimizing the weighted reconstructed errors, as formulated in Eq. (8). During this procedure, moving weights are incorporated in order to express the relative importance of labeled data. It works well when there are enough labeled correspondence pairs in the training set. However, the performance degrades when the labeled pairs are scarce or not evenly distributed in original input spaces. Inspired by manifold regularization [Belkin et al. 2004; Shao et al. 2011], in the proposed method, an additional regularization term is imposed onto the objective function, which aims to utilize the large amount of unlabeled data to achieve better generalization ability.

Considering the smoothness of the transformation functions, an additional regularization term is imposed onto the objective function. Ultimately, we formulate the objective function which uses both labeled and unlabeled data as follows:

$$\mathcal{J}(f_i) = \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} ||\mathbf{y}_j - f_i(\mathbf{x}_j)||^2 + \lambda ||f_i||^2 + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip} ||f_i - f_p||^2, \tag{9}$$

where the last term imposes the penalty that preserves the smoothness of local mapping functions between point $\mathbf{x}_i$ and the unlabeled data in the shared latent space. Besides, $s_{ip}$ is the weight defined in Eq. (2), and $\Upsilon$ is a parameter that controls the balance between the fitting accuracy and regularization performance.

*2.3.3. Optimization Solution.* The optimal transformation matrix $\mathbf{W}_i$ can be obtained with a closed-form solution by solving a convex optimization problem. Substituting Eq. (6) into the loss function $J$, defined in Eq. (9), we can obtain

$$\mathcal{J}(\mathbf{W}_i) = \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} ||\mathbf{y}_j - \mathbf{W}_i^T \mathbf{x}_j||^2 + \lambda ||\mathbf{W}_i||^2 + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip} ||\mathbf{W}_i - \hat{\mathbf{W}}_p||^2. \tag{10}$$

Then, take the derivative of $J$ with respect to $\mathbf{W}_i^T$ and set it to zero:

$$\frac{\partial \mathcal{J}(\mathbf{W}_i)}{\partial \mathbf{W}_i^T} = 2 \left( \sum_{\mathbf{x}_j \in s^l(\mathbf{x}_i)} \theta_{ij} (\mathbf{W}_i^T \mathbf{x}_j - \mathbf{y}_j) \mathbf{x}_j^T + \lambda \mathbf{W}_i^T + \gamma \sum_{\mathbf{x}_p \in s^u(\mathbf{x}_i)} s_{ip} (\mathbf{W}_i - \hat{\mathbf{W}}_p)^T \right) = 0. \tag{11}$$

After expanding the equation, we have

$$\begin{aligned} &\sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} \mathbf{W}_i^T \mathbf{x}_j \mathbf{x}_j^T + (\lambda + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip}) \mathbf{W}_i^T \\ &= \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} \mathbf{y}_j \mathbf{x}_j^T + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip} \hat{\mathbf{W}}_p^T, \end{aligned} \tag{12}$$

and the optimal $\mathbf{W}_i^T$ can be finally represented as

$$\begin{aligned} \mathbf{W}_i^T = &\left( \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} \mathbf{y}_j \mathbf{x}_j^T + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip} \hat{\mathbf{W}}_p^T \right) \\ &\left( \sum_{\mathbf{x}_j \in \mathcal{S}^l(\mathbf{x}_i)} \theta_{ij} \mathbf{x}_j \mathbf{x}_j^T + \lambda \mathbf{I} + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip} \mathbf{I} \right)^{-1}, \end{aligned} \tag{13}$$

where $\mathbf{I}$ is the identity matrix, and $\hat{\mathbf{W}}_p$ is the transformation function for computed unlabeled data point $\mathbf{x}_p$.

Note the objective function $\mathcal{J}(f_i)$ in Eq. (9) is quadratic in $f_i$; therefore, it is theoretically possible to obtain a closed-form solution for the parameters of all $u$ unlabeled transformations by solving a set of $u$ equations. Nevertheless, this approach is undesirable as it requires inverting a possibly large $u \times u$ matrix. We propose here a more efficient alternative approach for obtaining an approximate solution. For the local transformations of unlabeled data, we first order them based on the distances to labeled data, and then the transformations for the data points closer to the labeled points are estimated before those that are farther away. This alternative approach may be regarded as a process of propagating the changes from the labeled points to the unlabeled points.

Consequently, if $\mathbf{x}_i$ is an unlabeled data sample, $\mathcal{S}^u(\mathbf{x}_i)$ denotes the $i-1$ local transformation functions estimated before; while if $\mathbf{x}_i$ is a new test data sample, $\mathcal{S}^u(\mathbf{x}_i)$ represents all unlabeled data since their transformation functions have already been computed during the training stage. Since the final solution $\mathbf{W}_i$ in Eq. (13) is linear with respect to all known values, we can express it in closed-form and compute efficiently without any iterative process.

### 2.4. The Algorithm

The algorithm flow of the proposed method is summarized in Table I. In our method, two optimization problems are defined (in Step 3, Step 4 and 5) and both take into consideration preserving of local structure. In Step 3, the low-dimensional embeddings

Table I. Algorithm of the Proposed Method

```
Input: Training Set $\mathcal{G}$ and $\mathcal{P}$ with $k$ labeled correspondence pairs in $\mathcal{C}$,
       test data points $\mathbf{x}_i^g \in \mathbb{R}^{D_g}$ and $\mathbf{x}_j^p \in \mathbb{R}^{D_p}$,
Output: The distance metric induced by $\mathbf{W}_i^g$, $\mathbf{W}_j^p$ and $d(\mathbf{x}_i^g, \mathbf{x}_j^p)$.
Training Stage:
  Phase I: Globally consistent shared latent feature space learning
    Step 1: Construct the adjacency graph $G^g$ and $G^p$ for each training set;
    Step 2: Construct graph $G$ by adding edges between the labeled
            correspondence vertices;
    Step 3: Compute the optimal embedding results $\{\{\mathbf{y}_i^g\}_{i=1}^k, \{\mathbf{y}_j^p\}_{j=1}^k\}$ for all
            labeled data of the two sets according to Eq. (5);
  Phase II: Locally smooth multiview metric learning for unlabeled data
    Step 4: Compute the locally linear transformations for all unlabeled data
            of both views according to Eq. (13);
Test Stage: Locally smooth multiview metric learning for test data
    Step 5: Compute the locally linear transformations $\mathbf{W}_i^g$ and $\mathbf{W}_j^p$ for
            data $\mathbf{x}_i^g$ and $\mathbf{x}_j^p$ according to Eq. (13);
    Step 6: Get the embeddings $\mathbf{y}_i^g = (\mathbf{W}_i^g)^T\mathbf{x}_i^g$ and $\mathbf{y}_j^p = (\mathbf{W}_j^p)^T\mathbf{x}_j^p$ in the shared
            latent feature space;
    Step 7: Compute distance metric $d(\mathbf{x}_i^g, \mathbf{x}_j^p)$ according to Eq. (7).
```

are learned for all labeled data through local-scaling-based graph Laplacian matrix. In Steps 4, and 5, different local metrics are learned at different locations of the input spaces via regularized locally linear regression. Consequently, the overall transformations of all points in input spaces are locally linear but globally nonlinear. With such a property, our method not only can keep the advantage of easy computation due to the local processing manner, but also has a strong nonlinear processing ability for complex status.

The proposed MVML-GL method can be easily generalized to metric learning in more than two observation spaces. The generalized algorithm has two differences compared with that of Table I: multiple adjacency graphs are constructed for different observation datasets in Step 1; they are then connected into one graph according to the labeled correspondence pairs among multiple datasets in Step 2. In particular, when there is only one observation space, our method will degrade to the conventional single-space-based metric-learning method.

In addition, our method is not restricted to semi-supervised learning (SSL) settings and can also be used in supervised learning (which can also be considered as the extreme case of SSL). For our method, SSL can make use of unlabeled data to achieve better performance, especially when the labeled data is sparse and unevenly distributed.

### 2.5. Complexity Analysis

In the offline phase, the optimal solutions for all labeled samples of both views are obtained by solving a generalized eigenvector problem in Eq. 5, of which the computational cost is $O((2l)^3)$, and $l$ is the number of labeled data in each view.

In the online phase, for each local smooth transformation, the main computation burden comes from the computation of the matrix inverse in Eq. (13), where $d$ is the dimension of the input space. Let $T(d)$ be the time complexity of computing the inverse of a matrix in $\mathbb{R}^{d\times d}$, and $T(d) = O(d^3)$ using the standard method, or $T(d) = O(d^{2.376})$

Table II. Average Running Time in MVML-GL

| Online Phase | | |
|---|---|---|
| Method | Standard Inverse | Woodbury identity |
| $d = 256, l = 16$ | 0.0223 | 0.0098 |
| $d = 1024, l = 16$ | 0.6591 | 0.1580 |

*Note*: Unit: second/sample.

with the method of Coppersmith and Winogard. Thus the algorithm is efficient as long as $d$ is not exceptionally large.

When $d$ is large, we can reduce the computational complexity by several alterative ways. One straightforward way is to perform principal component analysis (PCA) [Jolliffe 2002] first to reduce the feature dimensions. Another way is to use the principle of the Woodbury matrix identity [Petersen and Pedersen 2008]. The inverse part denoted as $\mathbf{M}$ in Eq. (13) can be rewritten as

$$\mathbf{M} = (\mathbf{X}_l \mathbf{D}_\theta \mathbf{X}_l^T + a\mathbf{I})^{-1}, \tag{14}$$

where $\mathbf{X}_l$ is the labeled data matrix in one view with size of $d \times l$; $\mathbf{D}_\theta$ is a $l \times l$ diagonal matrix with each diagonal entry as $\theta_{ij}$, $\mathbf{I}$ is the identity matrix, and $a = \lambda + \Upsilon \sum_{\mathbf{x}_p \in \mathcal{S}^u(\mathbf{x}_i)} s_{ip}$. According to the Woodbury matrix identity, the inverse computation in Eq. (14) has an equivalent form expressed as

$$(\mathbf{X}_l \mathbf{D}_\theta \mathbf{X}_l^T + a\mathbf{I})^{-1} = \frac{1}{a}\mathbf{I} - \frac{1}{a^2}\mathbf{X}_l(\mathbf{D}_\theta^{-1} + \frac{1}{a}\mathbf{X}_l^T \mathbf{X}_l)^{-1}\mathbf{X}_l^T. \tag{15}$$

Since $\mathbf{D}_\theta$ is a diagonal matrix, the inverse compuation is very efficient. In this case, the inverse operation is conducted on an $l \times l$ matrix with the complexity of $O(l^3)$. As a consequence, the overall computation complexity in the online phase is $O(l^3 + ld^2)$, where $O(ld^2)$ is the complexity for matrix multiplication. When the labeled data number $l$ is smaller than input dimension $d$, we could utilize the Woodbury matrix identity instead to facilitate the computation.

The average running time of the proposed method on a typical computer (2.53GHz CPU, 4G memory) is shown in Table II. We can see that the running time decreases dramatically by using the Woodbury matrix identity, especially when the dimension of features is high.

In addition, to get the trade-off between accuracy and efficiency, we can also generalize the proposed method by assuming that the samples in a local region share the same projection rather than that each sample has its own one.

## 3. MULTIVIEW METRIC LEARNING AND MANIFOLD ALIGNMENT

From the perspective of manifold learning, the entire set of data in each view could be regarded as a manifold. Thus, multiview metric learning could be interpreted as a distance measurement among the data points of different manifolds. In spirit, multiview metric learning has a deep connection with a typical manifold learning application called manifold alignment, which aims to learn the correspondences between samples from different manifolds.

Broadly speaking, manifold alignment algorithms mainly fall into two categories: implicit-mapping-based solution and explicit-mapping-based. On one hand, the implicit-mapping-based methods, with no explicit projective mapping to be learnt, generate nonlinear low-dimensional embedding for each manifold. The nonlinear low-dimensional embedding can be interpreted as the data representation $\mathbf{y}_i^g$ and $\mathbf{y}_j^p$, as defined in Eq. (7). Since no mapping functions could be directly derived, this kind of methods could only do distance computation for data in the training set. Some representative work [Gong et al. 2005; Ham et al. 2005; Shon et al. 2006; Xiong et al. 2007], tend to directly align multiple data manifolds into a shared latent space or predefined target coordinates, and then match instances in correspondence. The main drawback of implicit-mapping-based methods is that they cannot process new test data without retraining. On the other hand, explicit-mapping-based methods, which learn linear transformations for each manifold, can be interpreted as learning two projective matrices $\mathbf{W}_i^g$ and $\mathbf{W}_j^p$, as defined in Eq. (7), and essentially solving the similar problem as multiview distance metric learning. Manifold alignment using Procrustes analysis [Wang and Mahadevan 2008] is a representative explicit-mapping-based method, which can be generalized to new data points. However, Procrustes analysis, which only learns a single affine transformation between one manifold and the other, fails to work well when the relationship between the two manifolds is beyond the affine transformation.

Conclusively, explicit-mapping-based manifold alignment methods are solving a similar problem as multiview metric learning; and implicit-mapping-based ones also have its essential connections to multiview metric learning. Taking the relationship into consideration, we could evaluate the proposed multiview metric-learning algorithm by comparing with some state-of-the-art manifold alignment algorithms.

## 4. EXPERIMENTAL RESULTS

To demonstrate the superiority of the proposed MVML-GL algorithm, we conduct various experiments on two typical computer vision applications: pose estimation and facial expression recognition. For comprehensive comparison, the MVML-GL algorithm is compared with some state-of-the-art methods on real-world datasets. It should be noted that in the following experiments, we do not compare with the method of Zheng et al. [2010]. This is because that we aim to align the instances based on the consistency of multiple views, while the method of Zheng et al. [2010], additional labeled data in each view is needed in order to achieve a better discriminative ability. Our method is under the same setting as CCA where only labeled correspondence pairs are needed during the training stage, which is more general compared with that in Zheng et al. [2010].

In experiments, the distance between one point $x_j^p$ in the probe and each point $x_i^g$ in the gallery set is calculated according to Table I; then the nearest one is aligned as the counterpart of $x_j^p$.

There are a few parameters involved in the MVML-GL algorithm. The dimension of the common embeddings is fixed to five, and the regularization coefficients $\lambda$ and $\Upsilon$ are set to some values in $(0, 1]$ which can be finally determined by cross-validation.

In practical implementation, we use the source code provided by Magnus Borga and M.B. Blaschko et al. for CCA[2] and KCCA[3], respectively. The LPA algorithm mainly takes two steps: manifold dimensionality reduction and alignment using Procrustes Analysis. Regarding the second step in LPA, we utilize the source code provided by Miguel A. Carreira-Perpinn to do Procrustes analysis.[4] The NMA algorithm is implemented by us. In the comparative studies, the parameters are tuned to achieve the best performance.

[2] http://www.imt.liu.se/~magnus/cca

[3] http://www.robots.ox.ac.uk/~blaschko/

[4] http://faculty.ucmerced.edu/mcarreira-perpinan/software.html

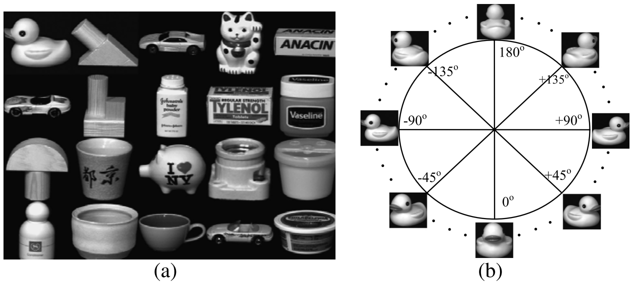


Fig. 2. (a) Sample images for 20 subjects in the COIL-20 dataset; (b) coordinate settings of poses in the COIL-20 dataset.

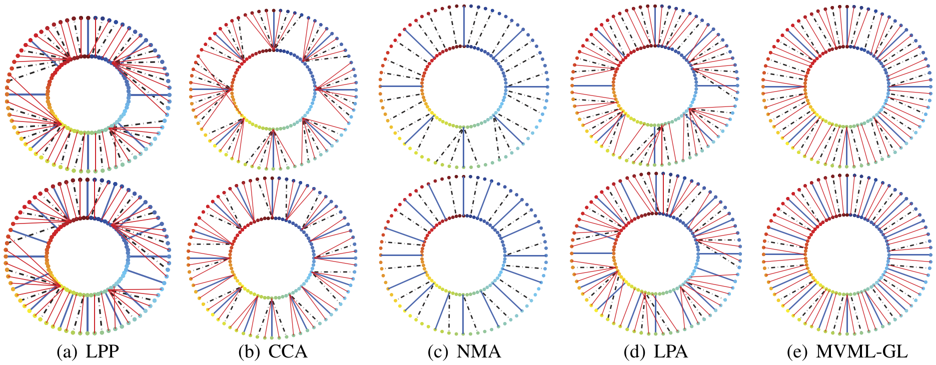

(a) LPP (b) CCA (c) NMA (d) LPA (e) MVML-GL

Fig. 3. The results of aligning "duck" and "block" inlayed on two concentric circles. The rows from top to bottom correspond to the number of the labeled correspondence pairs 8 and 16. Each column corresponds to the results of the algorithms (a) LPP, (b) CCA, (c) NMA, (d) LPA, and (e) our method.

### 4.1. Experiments on the COIL-20 Dataset

COIL-20 [Nene et al. 1996] is a dataset which contains 1,440 images of 20 objects. As shown in Figure 2, the pose coordinates for each object specify the camera movements around it, which contain 72 different sites at intervals of 5°. In experiments, we attempt to align the images with various poses from different objects for pose estimation. For each observation, 32 images are selected evenly for the training set, and the rest of the 40 images for testing. Each image is resized to 16 × 16 pixels, and image intensity is used as the feature.

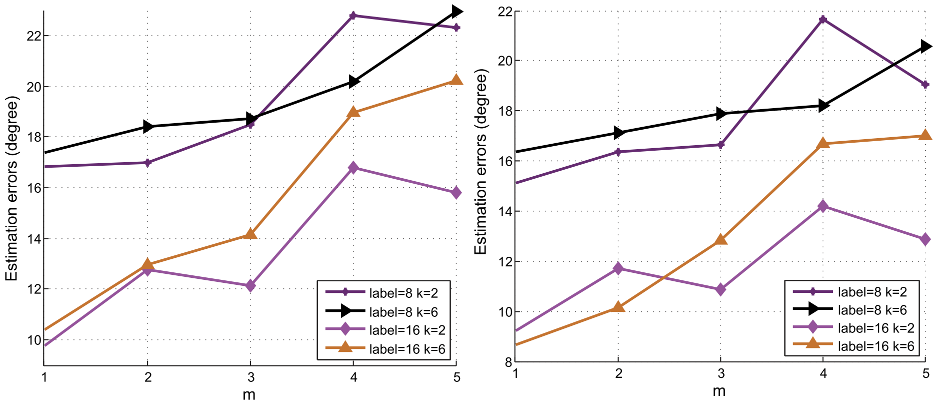
To illustrate the alignment results in an intuitve way, we inlay two aligned sequences on two concentric circles. As indicated in Figure 3, the blue-bold lines connect the labeled correspondence pairs, while the black-dashed and red lines denote the connections for aligned unlabeled and test points, respectively. In perfect alignment, the connections should be along the direction of radius (of the rays emitting from the center of the concentric circles). Due to the space limitation, the Euclidean results are omitted in Figure 3. Under this criterion, it can be seen that our method has no line cross and the lines connecting unlabeled and test data are more regular compared with other methods. It demonstrates that our method is locally smooth and has a strong out-of-sample ability to process new test data. Moreover, the quantitative comparison of the pose estimation is given in Table III in the case that the labeled correspondence number is 8 and 16.

From the comparative results shown in Figure 3 and Table III, we can see that the baseline method, which simply adopts Euclidean distance metric in input spaces,

Table III. Pose Estimation Errors on COIL-20 Dataset

| Method | ♯ labels = 8 | | ♯ labels = 16 | |
|---|---|---|---|---|
| | Unlabeled Error | Test Error | Unlabeled Error | Test Error |
| Euclidean | 36.90±3.5156 | 42.24±3.7698 | 36.69±4.5457 | 42.24±3.7698 |
| LPP | 34.54±5.739 | 39.86±5.34 | 34.54±5.739 | 39.86±5.34 |
| CCA | 24.05±5.8553 | 21.14±4.4208 | 19.84±8.0048 | 13.41 ±6.6283 |
| NMA | 18.43±11.2015 | - | 12.65±12.7656 | - |
| LPA | 18.71±9.7871 | 17.63±8.3203 | 17.75±8.998 | 17.43±8.7068 |
| MVML-GL | **17.38±8.0356** | **16.36±7.3151** | **9.75 ±5.9663** | **9.22±4.7556** |

*Note*: Unit: degree; mean ± std-dev.



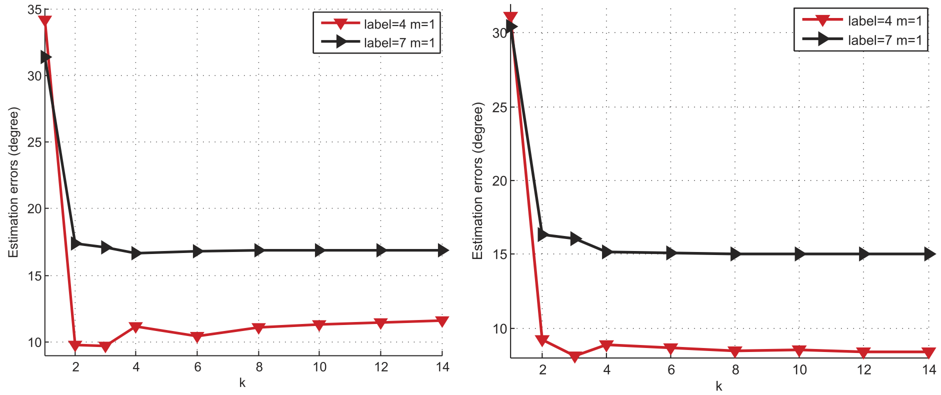
(a) Unlabeled data.

(b) Test data.

Fig. 4. Parameter selection for $m$ in the COIL-20 dataset.

yields poor performance. The Locality Preserving Projections (LPP) method [He and Niyogi 2003], which can uncover the essential manifold structure in a single observation space, fails to handle the scenario of multiview observation spaces well. The CCA method, which is linear and global, cannot provide satisfactory results either, since it only considers the relationship between the correspondence pairs without preserving local structure for each input space. The Nonlinear Manifold Alignment (NMA) method [Ham et al. 2005], which performs better than CCA, takes into account local structure preserving, but it cannot process the new test data since no explicit mapping is learned. As for Linear Procrustes Analysis (LPA) [Wang and Mahadevan 2008], it cannot give satisfactory results, since the relationship between the two manifolds is beyond the affine transformation. Among all the presented methods, the proposed MVML-GL achieves the best results. The metrics learned in the shared latent feature space via regularized locally linear regression vary smoothly and leads to significant performance benefits.

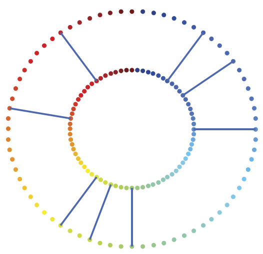
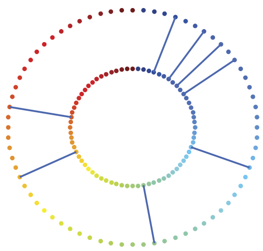
In the experiments, there are two important parameters, the local scalar $m$ and the number of nearest neighbors $k$. We first select $m$ with different labeled data numbers and different $k$. Then, the parameter $k$ is selected after determining the optimal $m$. As depicted in Figure 4, the pose estimation errors of the overall trend increase with the number of local scalar $m$, for different labeled data number and $k$-nearest-neighbors' settings. This may be due to the fact that the differences in appearance are large at an interval of five degrees of movement in the COIL-20 dataset. Accordingly, the width in the Gaussian kernel should be small enough to make the peak sharp. As depicted in Figure 5, the pose estimation errors decrease and finally level off with the increasing number of the $k$-nearest-neighbors. In addition, it is better to set even numbers for $k$



(a) Unlabeled data. (b) Test data.

Fig. 5. Parameter selection for $k$ in the COIL-20 dataset.

Table IV. The Pose Estimation Errors on COIL-20 Dataset When the Labeled Data is Unevenly Distributed

| | Method | ♯ labels = 8 | ♯ labels = 8 |
|---|---|---|---|
| | | Unlabeled Error | Test Error |
|  | Euclidean | 41.4254±2.9912 | 42.2368±3.7698 |
| | LPP | 37.1272± 4.8171 | 39.8553±5.340 |
| | CCA | 25.3728±3.2287 | 23.4605±4.3431 |
| | KCCA | 22.7961±8.0248 | 18.8947±5.6741 |
| | NMA | 22.6316±3.7981 | - |
| | LPA | 21.7215±6.0312 | 20.0066±6.0818 |
| | MVML-GL | **19.9013±4.961** | **17.7105±5.1084** |
| | Method | ♯ labels = 8 | ♯ labels = 8 |
| | | Unlabeled Error | Test Error |
|  | Euclidean | 40.4825±5.1763 | 42.2368±3.7698 |
| | LPP | 39.1996±5.6324 | 39.8553±5.340 |
| | CCA | 25.2741±5.1996 | 23.8487±3.7341 |
| | KCCA | 24.3531±8.6351 | 20.2895±7.5154 |
| | NMA | 24.8904±3.858 | - |
| | LPA | 26.7873±5.0505 | 25.1711±4.5644 |
| | MVML-GL | **23.2018±4.9948** | **19.9408±4.9607** |

*Note*: Unit: degree; mean ± std-dev.

than odd ones, since the pose variations are symmetric. As a result, we set $m = 1$ and $k = 2$ for the COIL-20 dataset.

In the experiments just presented, the labeled data is evenly sampled. However, in real-world computer vision problems, the labeled data is scarce and not homogeneously sampled in general. In this situation, most of the existing methods suffer from serious performance drops. As shown in Table IV, we give two examples when eight labeled points are unevenly distributed, while the unlabeled data is evenly distributed. From the results, we can see that the proposed MVML-GL method achieves relatively low estimation errors by utilizing a large amount of unlabeled data for better generalization.

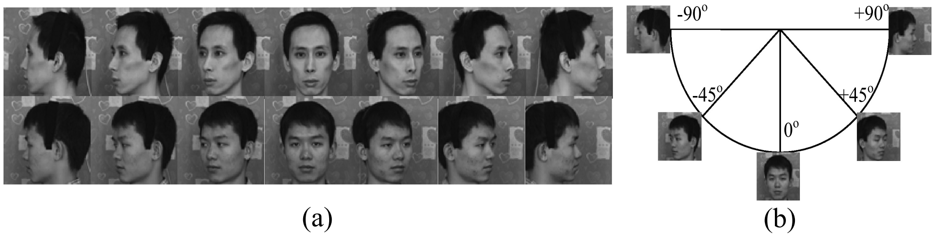
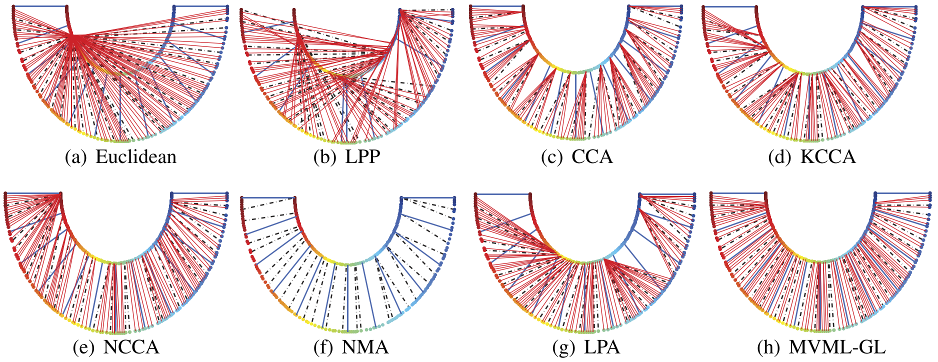


Fig. 6. (a) Two example sequences in the multi-pose face dataset; (b) coordinate settings of poses in our multi-pose face dataset.


(a) Euclidean (b) LPP (c) CCA (d) KCCA

(e) NCCA (f) NMA (g) LPA (h) MVML-GL

Fig. 7. The results of aligning 'person 1' and 'person 2' inlayed on two half concentric circles when labeled data number is 13 and evenly distributed. (a) Euclidean, (b) LPP, (c) CCA, (d) KCCA, (e) NCCA, (f)NMA, (g) LPA, and (h) our method.

### 4.2. Experiments on Face Pose Dataset

In this section, we will illustrate our method on a real-world face pose matching problem. The database we used is a private multi-pose database that we created. It consists of 1,011 images of ten people taken under normal indoor lighting conditions and fixed background with a Sony EVI-D31 camera. The poses are almost continuous (from $-90^\circ$ to $+90^\circ$, as shown in Figure 6(b)), and two example sequences are shown in Figure 6(a). In this experiment, the settings are almost the same as that for the COIL-20 dataset except that two additional comparative experiments are conducted and more quantitative results are given in order to achieve a more intensive evaluation.

Some qualitative results are illustrated in Figure 7, where the labeled data is evenly distributed, and the number of correspondce pairs is 13. Under the same visual criterion for the COIL-20 dataset, the proposed MVML-GL achieves the best alignment results among all comparative methods.

Furthermore, the quantity results are summarized in Table V. As shown in this table, Euclidean and LPP, both of which are single-view observation-space-based methods, exhibit high errors. The KCCA method, which has nonlinear processing ability, gets better results than CCA in most cases. NCCA, which introduces additional private latent spaces, also decreases the errors compared to CCA. NMA gets relatively good results for unlabeled data points but fails to handle new test data. LPA still cannot give satisfactory results in this situation. Among all these methods, the proposed MVML-GL method achieves the lowest errors.

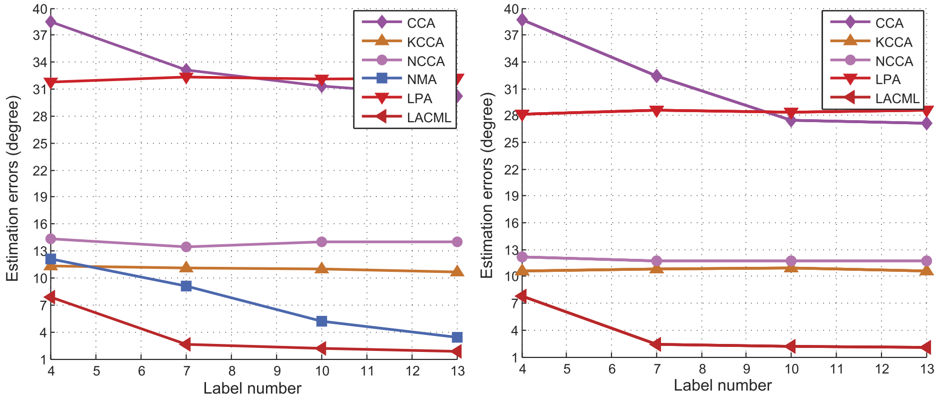
Moreover, the changes of estimation errors with the number of labeled data are shown in Figure 8. For a more clear illustration, the unsupervised methods Euclidean

Table V. Pose Estimation Errors on Multi-Pose Face Dataset

| Method | ♯ labels = 4 | | ♯ labels = 7 | |
|---|---|---|---|---|
| | Unlabeled Error | Test Error | Unlabeled Error | Test Error |
| Euclidean | 62.13±7.7499 | 68.94±15.4088 | 62.33±7.5137 | 68.94±15.4088 |
| LPP | 37.62±19.4957 | 39.34±16.9248 | 37.62±19.4957 | 39.34±16.9248 |
| CCA | 38.51±18.345 | 38.67±19.8798 | 33.13±19.2914 | 32.36±16.7348 |
| KCCA | 11.3229±8.2995 | 10.611±7.8843 | 11.1519±7.9169 | 10.868±7.572 |
| NCCA | 14.37±1.8158 | 12.20±1.8828 | 13.44±2.92 | 11.67±2.2336 |
| NMA | 12.11±0.78 | - | 9.09±0.4113 | - |
| LPA | 31.80±13.8377 | 28.12±12.101 | 32.35±15.2765 | 28.56±13.8047 |
| MVML-GL | **7.89±0.801** | **7.78±0.9557** | **2.57±0.8953** | **2.37±0.3519** |

| Method | ♯ labels = 10 | | ♯ labels = 13 | |
|---|---|---|---|---|
| | Unlabeled Error | Test Error | Unlabeled Error | Test Error |
| Euclidean | 62.49±7.7654 | 68.94±15.4088 | 62.08±7.9118 | 68.94±15.4088 |
| LPP | 37.62±19.4957 | 39.34±16.9248 | 37.62±19.4957 | 39.34±16.9248 |
| CCA | 31.38±15.5542 | 27.44±14.7403 | 30.18±16.6794 | 27.16±15.0844 |
| KCCA | 10.9794±8.1553 | 10.901±8.0018 | 10.713±7.4331 | 10.6±6.9615 |
| NCCA | 14.02±2.8728 | 11.70±2.3966 | 14.06±3.2071 | 11.77±1.7611 |
| NMA | 5.14±0.7618 | - | 3.37±0.2280 | - |
| LPA | 32.16±16.2053 | 28.41±14.7693 | 32.27±17.2452 | 28.55±15.4314 |
| MVML-GL | **2.22±0.6875** | **2.17±0.4924** | **1.87±0.5011** | **2.04±0.4243** |

*Note*: Unit: degree; mean ± std-dev.



(a) Unlabeled data (b) Test data

Fig. 8. Estimation errors as a function of the number of labeled data on the multi-pose face dataset.

and LPP results are omitted, since the results do not change with the number of labeled data. Overall, the estimation errors decrease with the number of labeled data. When the labeled data number is relatively small, the changes are obvious.

In addition, some visual results when the labeled data is unevenly distributed are illustrated in Figure 9. In this scenario, the Euclidean, LPP, CCA, and KCCA all have line-crosses of large angles. In contrast, our method keeps relatively good alignment results. It further demonstrates that the proposed method provides more accurate and stable results for unlabeled and test data and is more suitable for practical pose-alignment tasks compared with other methods.

Note that the first step, that is, the learning of the shared latent subspace (Phase I in Table I), can also be implemented by other methods, such as CCA and KCCA. We evaluate the contributions of the first step by replacing the proposed shared latent subspace learning method with CCA and KCCA, respectively, to get the common

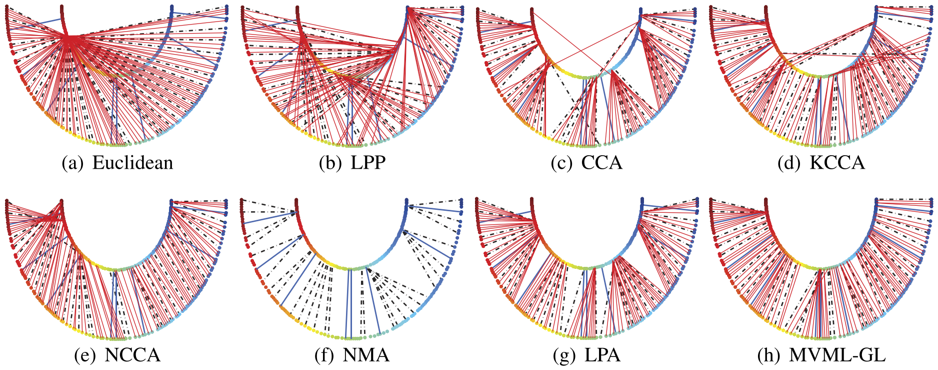


Fig. 9. The results of aligning ‘person 1’ and ‘person 2’ inlayed on two half concentric circles when the labeled data number is 8 and unevenly distributed. (a) Euclidean, (b) LPP, (c) CCA, (d) KCCA, (e) NCCA, (f)NMA, (g) LPA, and (h) our method.

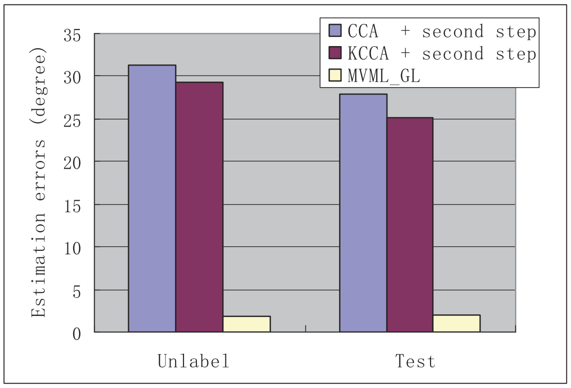
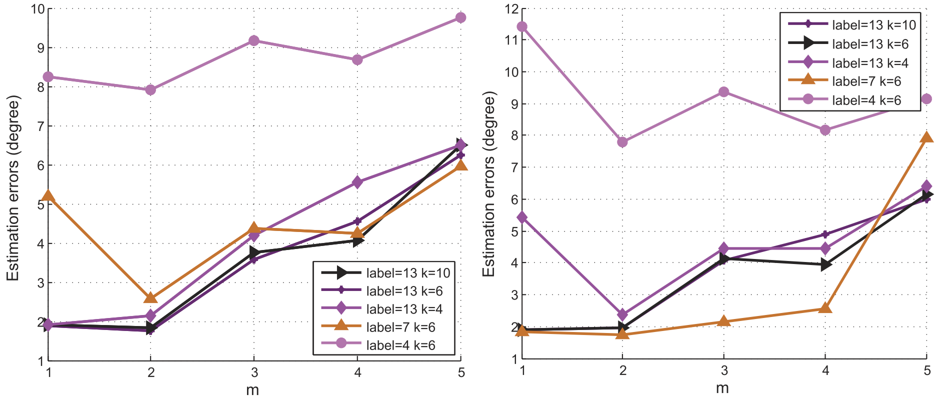


Fig. 10. The contribution of the first step on the multi-pose face dataset when there are 13 evenly distributed labeled data.

low-dimensions for all labeled data. Then the same metric-learning method is performed in the second step (Phase II and Test Stage in Table I) for unlabeled and test data. As depicted in Figure 10, the estimation errors of our proposed method decrease significantly compared with other comparative studies, which is mainly because the proposed shared latent subspace learning method introduces two local geometric structure preserving constraints for each view. In contrast, CCA and KCCA only preserve the sample correspondences but no topological structures when establishing the relationship between two views. Since in the second step, the mapping function for unlabeled or test data is a locally smooth learning procedure in the shared latent space, the performance will seriously decrease if the local structure cannot be preserved. Therefore, the shared latent subspace learning is a basis for the second step, and it also plays a very important role for the ultimate alignment performance.

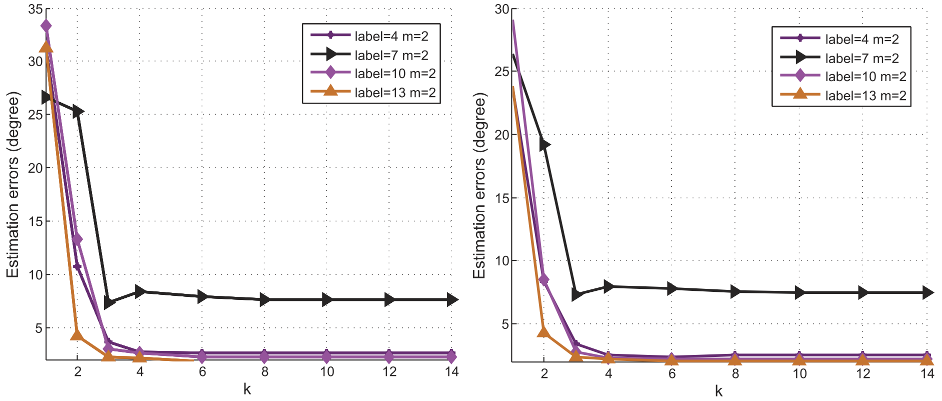
The impacts of parameters $k$ and $m$ for the multi-pose face dataset are illustrated in Figure 11 and 12. From the result, we can see that the pose estimation errors first decrease and then increase with the number of $m$. The changes of estimation errors with the number of $k$ are similar with that in COIL-20. Since the poses are almost continuous in the multi-pose face dataset, exploiting a relatively large number of nearest neighbor points will lead to a better approximation for the data manifold



(a) Unlabeled data.

(b) Test data.

Fig. 11. Parameter selection for $m$ in the multi-pose face dataset.



(a) Unlabeled data.

(b) Test data.

Fig. 12. Parameter selection for $k$ in the multi-pose face dataset.

structure. Thus, the optimal parameter is selected as $m = 2, k = 6$ to benefit from the local structure preserving for each view.

### 4.3. Experiments on the Facial Expression Dataset

Experiments are also conducted on a public facial expression dataset. The JAFFE dataset [Lyons et al. 1998] contains 213 images of seven facial expressions posed by ten Japanese female models. There are three or four images for each expression of each person. We randomly select one image of each expression as labeled data and the rest as unlabeled data. In our experiment, one person is labeled with known expressions, and we recognize the expressions of the other nine persons just by aligning their facial images to the standard image set.

Since sadness and disgust are omitted as hard-to-evoke expressions, we illustrate our method with five expressions including anger, fear, happiness, neutral, and surprise. Table VI shows the overall recognition results of the proposed MVML-GL method in the form of classification matrix. In our experiments, there is only one recognition failure for the expressions anger, neutral, and surprise, respectively. For happiness expressions, the success is 100%.

Table VI. The Overall Classification Matrix of MVML-GL on the JAFFE Dataset

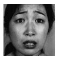
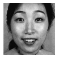
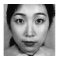
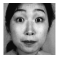
| Expression<br>Template | Anger<br> | Fear<br> | Happiness<br> | Neutral<br> | Surprise<br> |
|---|---|---|---|---|---|
| Anger | **17** | 0 | 0 | 0 | 0 |
| Fear | 0 | **15** | 0 | 1 | 0 |
| Happiness | 1 | 0 | **17** | 0 | 0 |
| Neutral | 0 | 3 | 0 | **17** | 1 |
| Surprise | 0 | 1 | 0 | 0 | **17** |
| Success | 94.44% | 78.95% | 100% | 94.44% | 94.44% |

Table VII. Recognition Rates (%) on the JAFFE Dataset

| Expression | Anger | Fear | Happiness | Neutral | Surprise | Average |
|---|---|---|---|---|---|---|
| Euclidean | 44.44±0.46 | 75.93± 0.38 | 0 | 0 | 0 | 24.07±0.17 |
| LPP | 33.33±0.43 | 55.56±0.46 | 55.56± 0.39 | 16.67± 0.35 | 38.89± 0.49 | 40.00± 0.43 |
| CCA | 94.44±0.17 | 77.78± 0.36 | 94.44±0.17 | 83.33± 0.35 | 72.22±0.44 | 84.44± 0.30 |
| KCCA | 83.33±0.35 | 66.67± 0.43 | **100** | 77.78± 0.36 | **100** | 85.56± 0.23 |
| NMA | **100** | **87.04**± 0.26 | 94.44± 0.17 | 88.89± 0.22 | 94.44± 0.17 | 92.96± 0.16 |
| LPA | 94.44±0.17 | 83.33± 0.35 | **100** | 83.33± 0.35 | **100** | 92.22± 0.17 |
| MVML-GL | 94.44±0.17 | 83.33± 0.35 | **100** | **94.44**± 0.17 | 94.44± 0.17 | **93.33**± 0.17 |

*Note*: mean ± std-dev.

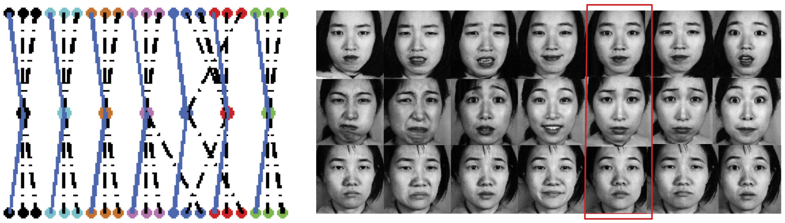
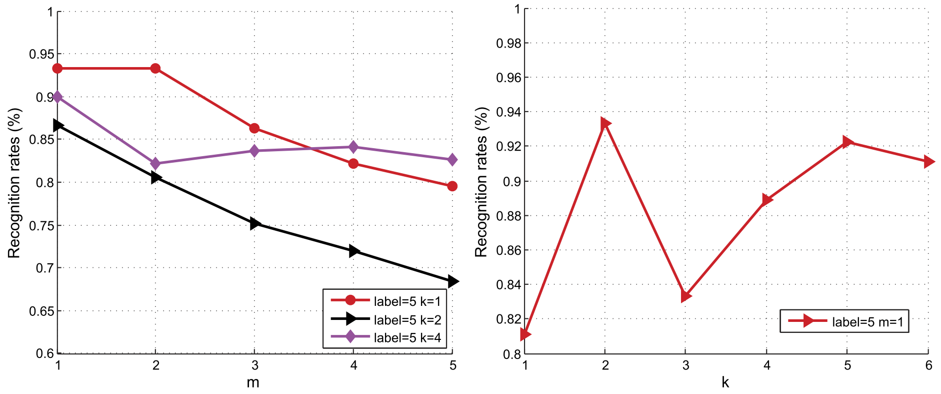

(a) (b)

Fig. 13. Alignment results among three-view observations. (a) Shows the alignment results of our method, where dots with the same color represent one kind of expression. (b) Shows some samples of matched images where the red box highlights the error matching.

We further compare our method with other related work. The recognition results of the comparative studies are summarized in Table VII. From the results, we can see that the Euclidean distance and LPP metric yield poor performances. CCA and KCCA are better than LPP as two general multiview metric-learning methods. NMA and LPA achieve even higher recognition rates, benefiting from the local preserving in each view. Our method gets comparable results as NMA and LPA and slightly better than KCCA, concerning the average recognition rates.

In addition, our method could be extended to distance metric learning in more than two observation spaces. In Figure 13(a), we show the the alignment results among three observations of all expressions. Each dot represents a face image, and the color specifies its expression. In perfect alignment, the dots with the same color should be connected. It can be seen that our method achieves good results except for some expressions, which are similar and hard to distinguish. Figure 13(b) shows some examples of the aligned expression pairs; the red box highlights the wrong alignment.

The impacts of parameters $k$ and $m$ for the JAFFE dataset are shown in Figure 14. Similar to COIL-20, the pose estimation errors increase with the number of local scalar $m$. And the best recognition rate is achieved when $k$ equals two. Note the changes of



(a) Unlabeled data. (b) Unlabeled data.

Fig. 14. Parameter selection for $m$ and $k$ in the multi-pose face dataset.

$k$ in the JAFFE dataset are different with that of $k$ in the pose dataset. This may be due to the fact that expression recognition is a multiclass problem, and the expression images do not lie on a continuous manifold in general, which makes the change of parameters $k$ more complicated.

## 5. CONCLUDING REMARKS

In this article, we have proposed a new multiview metric-learning algorithm to establish the relationship between multiview observations. Our method first reveals the globally consistent shared latent space of the multiview data by considering the geometric structure of each view and the connections between data from different views. Then regularized locally linear regression is performed to learn the explicit mapping functions between the input spaces and the shared latent space, in which the graph-Laplacian regularization term is incorporated to keep the learned metric functions various smoothly. These two procedures both can be formulated as convex optimization problems which could be solved efficiently with closed-form solutions. Experimental results on pose and facial expression matching provide empirical evidence for the effectiveness of our approach.

Despite its promising performance, there is still room for us to further improve our method. To trade accuracy for efficiency, the proposed multiview metric-learning method can be generalized to piece-wise linear, in which the samples in a local region share the same projection instead of each sample having its own one, which could bring some computation advantages for large-scale practical applications. Another interesting extension is to introduce discriminative information into the regularization framework for some classification problems.

## ACKNOWLEDGMENTS

The authors would like to thank three anonymous reviewers for their constructive suggestions which helped us improving our manuscript and also to thank Prof. Dit-Yan Yeung for useful advice in revising the paper.

## REFERENCES

AKAHO, S. 2001. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society (IMPS'01)*.

BAR-HILLEL, A., HERTZ, T., SHENTAL, N., AND WEINSHALL, D. 2003. Learning distance functions using equivalence relations. In *Proceedings of the 20th International Conference on Machine Learning*. 11–18.

BELKIN, M., NIYOGI, P., AND SINDHWANI, V. 2004. Manifold regularization : A geometric framework for learning from examples. *J. Machine Learn. Res.*, 2399–2434.

BOTTOU, L. AND VAPNIK, V. 1992. Local learning algorithms. *Neural Computat. 4*, 6, 888–900.

CAI, D., HE, X., AND HAN, J. 2007. Spectral regression for efficient regularized subspace learning. In *Proceedings of the 11th IEEE International Conference on Computer Vision*.

CHANG, H. AND YEUNG, D. 2007. Local smooth metric learning with application to image retrieval. In *Proceedings of the 11th IEEE International Conference on Computer Vision*.

DAVIS, J. V., KULIS, B., JAIN, P., SRA, S., AND DHILLON, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*. 209–216.

EK, C. H., RIHAN, J., TORR, P. H. S., ROGEZ, G., AND LAWRENCE., N. D. 2008. Ambiguity modeling in latent spaces. In *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction (MLMI'08)*. Springer-Verlag, Berlin, 62–73.

FROME, A., SINGER, Y., AND MALIK, J. 2006. Image retrieval and classification using local distance functions. In *Advances in Neural Information Processing Systems 19*, 417–424.

FROME, A., SHA, F., SINGER, Y., AND MALIK, J. 2007. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *Proceedings of the 11th IEEE International Conference on Computer Vision*.

GOLDBERGER, J., ROWEIS, S., HINTON, G., AND SALAKHUTDINOV, R. 2004. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, MIT Press, 513–520.

GONG, H., PAN, C., YANG, Q., LU, H., AND MA, S. 2005. A semi-supervised framework for mapping data to the intrinsic manifold. In *Proceedings of the 10th IEEE International Conference on Computer Vision*. Vol. 1, 98–105.

HAM, J., LEE, D. D., AND SAUL, L. K. 2005. Semisupervised alignment of manifolds. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. 120–127.

HARDOON, D., SZEDMAK, S., AND SHAWE-TAYLOR, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput. 16*, 2639–2664.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

HE, X. AND NIYOGI, P. 2003. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*, MIT Press.

HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika 28*, 312–377.

HUANG, K., YANG, H., KING, I., AND LYU, M. R. 2004. Learning large margin classifiers locally and globally. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*. ACM, New York, NY, 401–408.

JIN, R., WANG, S., AND ZHOU, Y. 2009. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta Eds., 862–870.

JOLLIFFE, I. 2002. *Principal Component Analysis* 2nd Ed. Springer, New York.

LEI, Z. AND LI., S. Z. 2009. Coupled spectral regression for matching heterogeneous faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

LEVIN, D. 1998. The approximation power of moving least squares. *Math. Computat. 67*, 224, 1517–1531.

LI, B., CHANG, H., SHAN, S., AND CHEN, X. 2009. Coupled metric learning for face recognition with degraded images. In *Proceedings of the 1st Asian Conference on Machine Learning*.

LIU, W., MA, S., TAO, D., LIU, J., AND LIU, P. 2010. Semi-supervised sparse metric learning using alternating linearization optimization. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. 1139–1148.

LYONS, M. J., KAMACHI, M., GYOBA, J., AND AKAMATSU, S. 1998. Coding facial expressions with gabor wavelets. In *Procedings of the 3rd IEEE Automatic Face and Gesture Recognition*.

NENE, S., NAYAR, S., AND MURASE, H. 1996. Columbia object image library: Coil-20. Tech. rep. CUCS-006-96, Columbia University.

PETERSEN, K. B. AND PEDERSEN, M. S. 2008. The matrix cookbook. `http://matrixcookbook.com`.

SAUL, L. K., ROWEIS, S. T., AND SINGER, Y. 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res. 4*, 119–155.

SHAO, Y., ZHOU, Y., AND CAI, D. 2011. Variational inference with graph regularization for image annotation. *ACM Trans. Intell. Syst. Technol. 2*, 11:1–11:21.

SHON, A. P., K. GROCHOW, A. H., AND RAO, R. 2006. Learning shared latent structure for image synthesis and robotic imitation. In *Advances in Neural Information Processing Systems 18*, 1233–1240.

SINDHWANI, V. AND NIYOGI, P. 2005. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.

VAPNIK, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

WANG, C. AND MAHADEVAN, S. 2008. Manifold alignment using procrustes analysis. In *Proceedings of the 25th International Conference on Machine Learning*. 1120–1127.

WANG, F., ZHANG, C., AND LI, T. 2007. Clustering with local and global regularization. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*. 657–662.

WEINBERGER, K., BLITZER, J., AND SAUL, L. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*.

WU, M. AND SCHÖLKOPF, B. 2006. A local learning approach for clustering. In *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA, 1529–1536.

WU, M. AND SCHÖLKOPF, B. 2007. Transductive classification via local learning regularization. In *Proceedings of the 11th International Workshop on Artificial Intelligence and Statistics*.

WU, M., YU, K., YU, S., AND SCHÖLKOPF, B. 2007. Local learning projections. In *Proceedings of the 24th International Conference on Machine Learning*. 1039–1046.

WU, L., HOI, S. C., JIN, R., ZHU, J., AND YU, N. 2011. Distance metric learning from uncertain side information for automated photo tagging. *ACM Trans. Intell. Syst. Technol. 2*, Article 13.

XING, E., NG, A., JORDAN, M., AND RUSSELL, S. 2003. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer Eds., MIT Press, Cambridge, MA, 505–512.

XIONG, L., WANG, F., AND ZHANG, C. 2007. Semi-definite manifold alignment. In *Proceedings of the 18th European Conference on Machine Learning (ECML)*. 773–781.

YANG, L., JIN, R., SUKTHANKAR, R., AND LIU, Y. 2006. An efficient algorithm for local distance metric learning. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*. AAAI Press, 543–548.

YEUNG, D.-Y., CHANG, H., AND DAI, G. 2008. A scalable kernel-based semi-supervised metric learning algorithm with out-of-sample generalization ability. *Neural Computat. 20*, 11.

ZELNIK-MANOR, L. AND PERONA, P. 2005. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*. MIT Press, 1601–1608.

ZHAN, D., LI, M., LI, Y.-F., AND ZHOU, Z. 2009. Learning instance specific distances using metric propagation. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*. ACM, New York, NY, 1225–1232.

ZHENG, H., WANG, M., AND Z.LI. 2010. Audio-visual speaker identification with multiview distance metric learning. In *Proceedings of the IEEE 17th International Conference on Image Processing*. 4561–4564.

Received November 2011; revised March 2011; accepted May 2011