Robust Distance Metric Learning with Auxiliary Knowledge*

Zheng-Jun Zha[*†‡] **Tao Mei**[†] **Meng Wang**[†] **Zengfu Wang**[*] **Xian-Sheng Hua**[†]
[*] Department of Automation, University of Science and Technology of China, Hefei, 230027, P. R. China
[†]Internet Media Group, Microsoft Research Asia, Beijing, 100190, P. R. China
[‡] MOE-Microsoft Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China, Hefei, 230027, P. R. China

Abstract

Most of the existing metric learning methods are accomplished by exploiting pairwise constraints over the labeled data and frequently suffer from the insufficiency of training examples. To learn a robust distance metric from few labeled examples, prior knowledge from unlabeled examples as well as the metrics previously derived from auxiliary data sets can be useful. In this paper, we propose to leverage such auxiliary knowledge to assist distance metric learning, which is formulated following the regularized loss minimization principle. Two algorithms are derived on the basis of manifold regularization and log-determinant divergence regularization technique, respectively, which can simultaneously exploit label information (i.e., the pairwise constraints over labeled data), unlabeled examples, and the metrics derived from auxiliary data sets. The proposed methods directly manipulate the auxiliary metrics and require no raw examples from the auxiliary data sets, which make them efficient and flexible. We conduct extensive evaluations to compare our approaches with a number of competing approaches on face recognition task. The experimental results show that our approaches can derive reliable distance metrics from limited training examples and thus are superior in terms of accuracy and labeling efforts.

1 Introduction

A lot of machine learning and pattern recognition methods, such as clustering, classification, and regression approaches, involve the use of a distance metric over the input feature space. The performance of these methods often depends highly on the choose of metric. Instead of determining a metric manually, a promising approach is to learn an appropriate metric from data automatically. Distance metric learning has become an emerging topic in which the goal is to induce a powerful distance metric from labeled examples. The distance metric is found by keeping "similar" objects close together while separating "dissimilar" ones. In recent years, a number of metric learning methods have been developed, which have shown to perform well when sufficient training data are available. However, in many real-world applications, training examples are very few, and in which the performance of these methods may significantly degrade due to the "overfitting" problem [Hoi *et al.*, 2008][Yang and Jin, 2006].

To derive reliable distance metric from the few labeled examples, one solution is to leverage auxiliary knowledge sources to assist distance metric learning, such as the structure information provided by the abundant unlabeled examples [Hoi *et al.*, 2008] and the knowledge derived from auxiliary data (e.g., the distance metrics learned from auxiliary data). Different from labeled data, unlabeled data are typically widely available, and existing studies have shown that this knowledge can be helpful in distance metric learning [Hoi *et al.*, 2008]. Here "auxiliary data" means the data that are collected from different sources and different from the target data in distribution. Such auxiliary data is easy to be obtained in practice and has proved useful in various applications [Pan *et al.*, 2008] [Gupta and Ratinov, 2008] [Talvitie and Singh, 2007] [Mansour *et al.*, 2007] [Satpal and Sarawagi, 2007].

However, directly applying the metric learned from auxiliary examples may not perform well since it might be biased to the distribution of the auxiliary data. Two straightforward approaches for leveraging the auxiliary data are : (1) to construct an "ensemble" combining the distance metric derived independently from the limited training data and the labeled auxiliary examples; and (2) to learn an "aggregated" distance metric from the combination of the limited labeled examples and auxiliary data. However, these methods may not work well mainly due to the distribution difference between the target and the auxiliary data. Moreover, learning an "aggregated" metric is typically expensive because the size of auxiliary data is usually large. Thus there is a need for a more efficient and effective approach for exploiting auxiliary knowledge to assist distance metric learning.

In this paper, we develop two novel algorithms for learning distance metric from only a small amount of labeled examples by simultaneously exploiting abundant unlabeled examples, as well as incorporating available prior knowledge from auxiliary data. Specifically, the distance metric learning is formulated as a regularized loss minimization problem. Two metric learning approaches are developed on the basis of

*This work was performed when Zheng-Jun Zha was visiting Microsoft Research Asia as a research intern.

manifold regularization and log-determinant divergence regularization technique, respectively, which can handle multiple knowledge, namely leveraging multiple auxiliary knowledge simultaneously to assist metric learning from few labeled examples. The weights of the auxiliary knowledge are optimized automatically to reflect the utility of these knowledge. We apply the proposed methods to face recognition task. The experimental results show that our methods can construct reliable distance metric and do significantly improve the performance when training data are limited.

The rest of this paper is organized as follows. We review related work in Section 2. Section 3 describes the proposed distance metric learning methods. Experimental results are reported in Section 4, followed by concluding remarks in Section 5.

2 Related Work

Our work is closely related to the previous studies on supervised distance metric learning [Yang and Jin, 2006] [Yang, 2007]. Most existing methods learn a distance metric from side information that is presented in a set of pairwise constraints: equivalence constrains that include pairs of "similar" examples and inequivalence constraints for "dissimilar" examples. The optimal distance metric is derived by keeping "similar" examples close and enforcing the "dissimilar" examples well separated. We briefly review some representative work here.

In recent years, a number of algorithms have been proposed for supervised distance metric learning. [Bar-Hillel *et al.*, 2005] proposed Relevant Components Analysis (RCA) method to learn a linear transformation from the equivalence constrains, which can be used directly to compute the distance between two examples. Disciminative Component Analysis (DCA) and Kernel DCA [Hoi *et al.*, 2006] improved RCA by exploiting negative constraints and aim to capture nonlinear relationships using contextual information. [Schultz and Joachims, 2003] extended the support vector machine to distance metric learning by encoding the pairwise constraints into a set of linear inequalities. [Xing *et al.*, 2003] formulated distance metric learning as a constrained convex programming problem by minimizing the distance between the data points in the same classes under the constraint that the data points from different classes are well separated. Neighborhood Component Analysis (NCA) [Goldberger *et al.*, 2004] learned a distance metric by extending the nearest neighbor classifier. [Weinberger *et al.*, 2006] proposed the maximum-margin nearest neighbor (LMNN) method that extends NCA through a maximum margin framework. [Globerson and Roweis, 2006] learned a Mahalanobis distance by collapsing examples in the same class to a single point and keeping examples from different classes far away. [Davis *et al.*, 2007] formulated distance metric learning as a Bregman optimization problem. [Hillel and Weinshall, 2007] defined the similarity as the gain in coding length by shifting from pairwise independent encoding to joint encoding. [Yang and Jin, 2007] presented a Bayesian framework for distance metric learning that estimates a posterior distribution for the distance metric from labeled pairwise constraints. [Yeung *et al.*, 2007] proposed a nonlinear metric learning method based on the kernel approach. [Alipanahi *et al.*, 2008] show a strong relationship between distance metric learning methods and Fisher Discriminant Analysis (FDA). [Hoi *et al.*, 2008] proposed a semi-supervised distance metric learning method that integrates both labeled and unlabeled examples.

3 Distance Metric Learning with Auxiliary Knowledge

3.1 Problem Set-UP

Let $\mathcal{C} = \{x_1, x_2, \cdots, x_N\}$ denote a collection of N data points, where $x_i \in \mathbb{R}^d$ is the d dimensional feature vector. For the labeled examples in $\mathcal{C}$, two sets of pairwise constraints are available,which are denoted by $\mathcal{S}$ and $\mathcal{D}$, respectively.

$$\mathcal{S} = \{(x_i, x_j | x_i \text{ and } x_j \text{ are labeled to be similar}\}$$
$$\mathcal{D} = \{(x_i, x_j | x_i \text{ and } x_j \text{ are labeled to be dissimilar}\},$$

where $\mathcal{S}$ is the set of similar pairwise constraints, and $\mathcal{D}$ is the set of dissimilar pairwise constraints.

For any two data points x_i and x_j, a Mahalabobis distance between them can be expressed as:

$$d_{\mathbf{M}}(x_i, x_j) = \|x_i - x_j\|_{\mathbf{M}} = \sqrt{(x_i - x_j)^T \mathbf{M}(x_i - x_j)}, \quad (1)$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is the Mahalabobis metric, a symmetric matrix of size $d \times d$. In general, $\mathbf{M}$ must be positive smeidefinite ($\mathbf{M} \succeq \mathbf{0}$) to satisfy the properties of metric, i.e., nonnegativity and triangle inequality. When $\mathbf{M}$ is equal to the identity matrix $\mathbf{I}$, the distance in Eq.(1) reduces to the Euclidean distance.

Many recent studies on distance metric learning focus on learning the Mahalabois matric $\mathbf{M}$ by leveraging the similar and dissmilar pairwise relations in $\mathcal{S}$ and $\mathcal{D}$. However, in many real-world scenario, the labeled examples are usually insufficient. As a result, the typical metric learning methods may suffer from the "overfitting" problem and can not provide reliable distance metric. On the other hand, the abundant unlabeled data add some auxiliary knowledge, i.e., the previously learned metrics from auxiliary data, are typically available and beneficial to the metric learning task. Therefore, we propose to learn a robust metric under the assistance of these auxiliary metrics and the abundant unlabeled data in $\mathcal{C}$. Let $\mathcal{M} = \{\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_K\}$ denote the set of K available auxiliary metrics $\mathbf{M}_i \in \mathbb{R}^{d \times d}$. Our goal is to find an optimal distance metric from the data collection $\mathcal{C}$, the sets of pairwise constraints $\mathcal{S}$ and $\mathcal{D}$, and the set of auxiliary metrics $\mathcal{M}$. To achieve this goal, two novel distance metric learning algorithms are developed in the following subsections.

3.2 Our Approach

Given the above information $\mathcal{C}$, $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{M}$, we can formulate the distance metric learning problem into the following optimization framework:

$$\mathbf{M}^* = \min_{\mathbf{M}} f(\mathbf{M}, \mathcal{C}, \mathcal{S}, \mathcal{D}, \mathcal{M}), \text{s.t.} \mathbf{M} \succeq 0, \quad (2)$$

where f is some objective function defined over the given data and $\mathbf{M}^*$ is the desired distance metric. The key to find the optimal distance metric is to formulate a proper objective function f. Following the regularized loss minimization principle, we can define f as a regularized loss function as:

$$f(\mathbf{M}, \mathcal{S}, \mathcal{D}, \mathcal{C}, \mathcal{M}) = L(\mathbf{M}, \mathcal{S}, \mathcal{D}) + R(\mathbf{M}, \mathcal{C}, \mathcal{M}), \quad (3)$$

where $L(\cdot)$ is a loss function defined on the pairwise constraints while $R(\cdot)$ is the regularization term that takes advantage of unlabeled data and auxiliary knowledge to prevent "overfitting" and enhance the generalization and robustness of the distance metric.

Eq.(3) provides a generic solution for learning robust metric from limited training data by exploiting auxiliary knowledge and unlabeled data. This solution offers great flexibility and efficiency, because it directly manipulates the previously learned metric as an abstraction of the auxiliary data and requires no raw auxiliary data. This also makes it applicable even when the raw data are not accessible. From this generic solution, one can derive concrete algorithm by choosing certain loss functions $L(\cdot)$ and regularization function $R(\cdot)$. While the choices are virtually numerous, we discuss two specific algorithms here.

Loss Function

One common principle for metric learning is to keep the "similar" data points close and separate the "dissimilar" ones. Following this principle, the loss function $L(\cdot)$ should be defined in the way such that the minimization of the loss function will result in minimizing the distances between the data points with similar constraints and maximizing the distances between the data points with dissimilar constraints. We adopt the sum of squared distances expression for defining the loss function in terms of its effectiveness and efficiency:

$$\begin{aligned}
&L(\mathbf{M}, \mathcal{S}, \mathcal{D}) \\
&= \eta_s \sum_{(x_i,x_j)\in\mathcal{S}} \|x_i - x_j\|^2_{\mathbf{M}} - \eta_d \sum_{(x_i,x_j)\in\mathcal{D}} \|x_i - x_j\|^2_{\mathbf{M}} \\
&= \eta_s \sum_{(x_i,x_j)\in\mathcal{S}} (x_i - x_j)^T\mathbf{M}(x_i - x_j) \\
&-\eta_d \sum_{(x_i,x_j)\in\mathcal{D}} (x_i - x_j)^T\mathbf{M}(x_i - x_j) \\
&= \eta_s \sum_{(x_i,x_j)\in\mathcal{S}} \mathbf{tr}\left((x_i - x_j)(x_i - x_j)^T\mathbf{M}\right) \\
&-\eta_d \sum_{(x_i,x_j)\in\mathcal{D}} \mathbf{tr}\left((x_i - x_j)(x_i - x_j)^T\mathbf{M}\right) \\
&= \eta_s \mathbf{tr}\left(\mathbf{S}\cdot\mathbf{M}\right) - \eta_d \mathbf{tr}\left(\mathbf{D}\cdot\mathbf{M}\right)
\end{aligned} \quad (4)$$

where $\mathbf{tr}(\cdot)$ means the trace operation on matrix, η_s and η_d are two trading-off parameters balancing the similar and dissimilar constraints, and

$$\mathbf{S} = \sum_{(x_i,x_j)\in\mathcal{S}} (x_i - x_j)(x_i - x_j)^T$$
$$\mathbf{D} = \sum_{(x_i,x_j)\in\mathcal{D}} (x_i - x_j)(x_i - x_j)^T$$

Next, we move our effort to define proper regularization function $R(\cdot)$.

Log-Determinant Regularization Function

As aforementioned, the training data are usually limited in practice and the available auxiliary knowledge (i.e., the auxiliary metrics $\mathcal{M}$) can be utilized to assist the distance metric learning. To encode $\mathcal{M}$ into f, we resolve to the regularization function $R(\cdot)$, which is a key to prevent "overfitting" and enhance generalization and robustness of the learned metric. We define $R(\cdot)$ in the way such that the minimization of $R(\cdot)$ will result in minimizing the divergence between the target metric $\mathbf{M}$ and the auxiliary metric $\mathbf{M}_k \in \mathcal{M}$. In other words, we aim to regularize $\mathbf{M}$ as close as possible to the auxiliary distance metrics. Here we adopt Bregman divergence [Davis *et al.*, 2007] to measure the difference between $\mathbf{M}$ and $\mathbf{M_k}$ as:

$$D_g(\mathbf{M}||\mathbf{M}_k) = g(\mathbf{M}) - g(\mathbf{M}_k) - \langle \nabla_g(\mathbf{M}_k), \mathbf{M} - \mathbf{M}_k \rangle,$$

where $g(\cdot)$ is a strict convex and continuously differentiable function. We define $g(\cdot)$ as $-\mathbf{logdet}(\cdot)$ and get the log-determinant divergence between $\mathbf{M}$ and $\mathbf{M}_k$:

$$D_g(\mathbf{M}||\mathbf{M}_k) = \mathbf{tr}(\mathbf{M}_k^{-1}\mathbf{M}) - \log\det\mathbf{M}, \quad (5)$$

where we ignore the constant term regarding $\mathbf{M}_k$.

Based on Eq.(5), we formulate $R(\cdot)$ as a combination of the divergence between $\mathbf{M}$ and each auxiliary knowledge $\mathbf{M}_k$:

$$\begin{aligned}
R(\mathbf{M}, \mathcal{C}, \mathcal{M}) &= \sum_{k=1}^{K} \mu_k D_g(\mathbf{M}||\mathbf{M}_k) \\
&= \sum_{k=1}^{K} \mu_k \left(\mathbf{tr}(\mathbf{M}_k^{-1}\mathbf{M}) - \log\det\mathbf{M}\right),
\end{aligned} \quad (6)$$

Substituting Eq.(4) and Eq.(6) into Eq.(3), we can get the concrete algorithm named **L-DML** (i.e., **L**og-determinant regularized **D**istance **M**etric **L**earning):

$$\begin{aligned}
\mathbf{M}^* = \min_{\mathbf{M}} &\sum_{k=1}^{K} \mu_k \left(\mathbf{tr}(\mathbf{M}_k^{-1}\mathbf{M})\right) - \log\det\mathbf{M} \\
&+ \eta_s \mathbf{tr}\left(\mathbf{S}\cdot\mathbf{M}\right) - \eta_d \mathbf{tr}\left(\mathbf{D}\cdot\mathbf{M}\right) + \gamma\|\mu\|^2 \\
\text{s.t.}\mathbf{M} &\succeq 0,\ \sum_{k=1}^{K} \mu_k = 1, \mu_k \geq 0, k = 1, 2, \cdots, K,
\end{aligned} \quad (7)$$

where $\mu = \{\mu_k\}_{k=1}^K$ are the weights that reflect the utility of the auxiliary metrics, $\|\mu\|^2$ is the L-2 norm of the weights and γ is a scalar. The new regularizer $\|\mu\|^2$ is adopted to penalize large weights on auxiliary metrics, and it thus prevents "overfitting" on the auxiliary knowledge. How to determine the weights $\{\mu_k\}_{k=1}^K$ is a problem, given the difficulty of knowing each auxiliary metric's utility to the target metric. To overcome this problem, **L-DML** is motivated to learn the weights $\{\mu_k\}_{k=1}^K$ automatically and simultaneously with the target metric $\mathbf{M}$. The solution will be discussed later.

Manifold Regularization Function

The aforementioned **L-DML** algorithm has leveraged the auxiliary metrics on the basis of log-determinant regularization technique. However, it does not take advantage of any

information of unlabeled data which can be useful for distance metric learning, especially when labeled data are limited. Therefore, we put effort to define a new regularizer such that it can simultaneously encode the auxiliary metrics $\mathcal{M} = \{\mathbf{M}_k\}_{k=1}^K$ and the unlabeled examples in the data collection $\mathcal{C}$. We define the new regularization term following manifold regularization principle.

Given auxiliary metric $\mathbf{M}_k$, the distance between two data points x_i and x_j under $\mathbf{M}_k$ can be calculated as $d_{\mathbf{M}_k}(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{M}_k (x_i - x_j)}$. Based on such distance measure, a data adjacency graph $\mathbf{W}_k \in \mathbb{R}^{N \times N}$ can be derived from the data collection $\mathcal{C}$, wherein each element $W_k(i,j)$ is the edge weight between two samples x_i and x_j. Considering the target metric $\mathbf{M}$, the distance between x_i and x_j is $d_{\mathbf{M}}(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{M} (x_i - x_j)}$ that can be further written as $d_{\mathbf{M}}(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{P}\mathbf{P}^\mathbf{T} (\mathbf{x_i} - \mathbf{x_j})}$ with $\mathbf{P} \in \mathbb{R}^{d \times m}$ is some corresponding linear mapping and $\mathbf{M} = \mathbf{P}\mathbf{P}^\mathbf{T}$. We can find that learning of $\mathbf{M}$ is equivalent to the learning of a linear projective mapping $\mathbf{P}$ in the feature space. Following the manifold regularization principle, we formulate the regularizer in the way such that the minimization of the regularizer will result in making the linear projective mapping being smooth over the data adjacency graph. Mathematically, we write the regularizer as:

$$\begin{aligned} & R_k(\mathbf{M}, \mathcal{C}, \mathbf{M}_k) \\ &= \tfrac{1}{2} \sum_{i,j=1}^{N} \left\| \mathbf{P}x_i - \mathbf{P}x_j \right\|^2 W_k(i,j) \\ &= \sum_{l=1}^{m} \mathbf{p}_l^T \mathbf{X}(\mathbf{D}_k - \mathbf{W}_k)\mathbf{X}^T \mathbf{p}_l = \sum_{l=1}^{m} \mathbf{p}_l^T \mathbf{X}\mathbf{L}_k\mathbf{X}^T \mathbf{p}_l \\ &= \mathbf{tr}(\mathbf{X}\mathbf{L}_k\mathbf{X}^T\mathbf{P}\mathbf{P}^T) = \mathbf{tr}(\mathbf{X}\mathbf{L}_k\mathbf{X}^T\mathbf{M}), \end{aligned} \tag{8}$$

where $\mathbf{L}_k$ is the graph Laplacian defined as $\mathbf{D}_k - \mathbf{W}_k$, $\mathbf{D}_k \in \mathbb{R}^{N \times N}$ is a diagonal matrix with $D_k(i,i) = \sum_{i=1}^{N} W_k(i,j)$, and $\mathbf{p}_l$ is the l-th column of matrix $\mathbf{P}$

Based on Eq.(8), we can define the regularization term $R(\cdot)$ as the combination of $R_k(\cdot)$ to incorporate all the auxiliary metrics and unlabeled data information.

$$R(\mathbf{M}, \mathcal{C}, \mathcal{M}) = \sum_{k=1}^{K} \alpha_k \mathbf{tr}(\mathbf{X}\mathbf{L}_k\mathbf{X}^T\mathbf{M}) \tag{9}$$

Substituting the loss function in Eq.(4) and the regularizer in Eq.(9) into Eq.(3), we can get the following algorithm. We name it as **M-DML** (i.e., **M**anifold regularized **D**istance **M**etric **L**earning) algorithm.

$$\begin{aligned} \mathbf{M}^* = \min_{\mathbf{M}} & \sum_{k=1}^{K} \alpha_k \mathbf{tr}(\mathbf{X}\mathbf{L}_k\mathbf{X}^T\mathbf{M}) \\ & + \eta_s \mathbf{tr}\left(\mathbf{S} \cdot \mathbf{M}\right) - \eta_d \mathbf{tr}\left(\mathbf{D} \cdot \mathbf{M}\right) + \beta \left\|\alpha\right\|^2 \\ \text{s.t.} \mathbf{M} \succeq 0, & \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \geq 0, k = 1, 2, \cdots, K \end{aligned} \tag{10}$$

where $\alpha = \{\alpha_k\}_{k=1}^K$ are the weights that reflect the utility of the auxiliary metrics, β is a scalar, and regularizer $\left\|\alpha\right\|^2$ is the L-2 norm of the weights $\{\alpha_k\}_{k=1}^K$ that prevents "overfitting" on auxiliary metrics. We can find that the LRDML algorithm in [Hoi *et al.*, 2008] is a special case of our **M-DML** method. Specifically, **M-DML** will reduce to LRDML when only Euclidian metric is available as the auxiliary metric, i.e., $\mathcal{M} = \{\mathbf{I}\}$.

3.3 Learning by Alternating Optimization

Up to now we have shown that multiple auxiliary metrics, unlabeled data, and the pairwise constraints on labeled examples can be integrated into a regularized loss minimization framework. Two distance metric learning algorithms **L-DML** and **M-DML** have been developed by exploiting log-determinant and manifold regularization technique, respectively. Now we discuss the solution of them.

We first consider **M-DML**. Equation (10) is a standard formulation of Semidefinite Programs (SDP) [Stephen and Lieven, 2003] under any fixed α and can be solved efficiently using existing convex optimization packages, such as SeDuMi [Sturm, 1999]. However, α is crucial to the performance of **M-DML**. Since the utility may vary intensively among different auxiliary metrics, α should vary as well as according to their utility. Thus **M-DML** is motivated to learn both the distance metric $\mathbf{M}$ and the linear combination coefficients α simultaneously. We can realize the joint learning of $\mathbf{M}$ and α by adopting the alternating optimization technique to solve Eq.(10) in an iterative manner. Specifically, we solve **M-DML** as follows: solving Eq.(10) with respect to $\mathbf{M}$ with fixed α; then optimizing Eq.(10) with respect to α with $\mathbf{M}$ taking the value obtained before; and alternatively iterating the above two steps until the decrement of the objective function is zero. This process will converge to the optimal solution since the objective function is convex to both $\mathbf{M}$ and α. Analogously, the **L-DML** method can also be solved via alternating optimization.

4 Experiments

In this section, we investigate the performance of the proposed **L-DML** and **M-DML** methods for face recognition.

4.1 Data Sets

We employ four face image data sets in our experiments: ORL [Samaria and Harter, 1994], Yale [Belhumeur *et al.*, 1997], Extended Yale-B [Georghiades *et al.*, 2001], and CMU PIE [Sim *et al.*, 2003] collections. The ORL data set contains 40 distinct human subjects and each subject has ten gray images. The Yale corpus consists of 165 gray images of 15 subjects and there are 11 images per person. The Extended Yale-B data set contains 161,289 gray images of 38 human subjects under nine poses and 64 illumination conditions. We choose the frontal pose and use all the images under different illumination, thus there are 64 images for each person. The CMU PIE corpus contains 41,368 images of 69 subjects under 13 different poses, 43 different illumination conditions, and with four different expressions. We choose ten subjects under two near frontal poses and all the images under different illuminations and expressions. Thus there are 68 images per subject and 680 images in total. For computational efficiency, all the

face images are manually aligned and cropped. The size of each cropped image is 20×20 pixels with 256 gray levels per pixel. The feature (pixel values) are then normalized to [0,1].

4.2 Experimental Setting and Results

Among the four data collections, we choose ORL, Yale and CMU PIE as the auxiliary data sets and Extended Yale-B as the target data. We learn the auxiliary metrics using Relevance Component Analysis (RCA) [Bar-Hillel *et al.*, 2005] from ORL, Yale, and CMU PIE data. Then we randomly select t images per subject from Extended Yale-B data set with the labels to form the training data and the remaining images are considered to be the test examples. We gradually increase t ($t = 2, 4, 6, 8, 10$), and for each t we perform ten trials and then compute the average performance. The tradeoff parameters are empirically set as: η_d and γ (β) are about one-forth and one-eighth of η_s, respectively.

For comparison purpose, we also evaluate seven existing methods for recognizing the face images in target data: Euclidean distance (**EU**), Xing's method (**Xing**) [Xing *et al.*, 2003], **RCA** [Bar-Hillel *et al.*, 2005], **DCA** [Hoi *et al.*, 2006], **NCA** [Goldberger *et al.*, 2004], **LMNN** [Weinberger *et al.*, 2006], and **LRDML** [Hoi *et al.*, 2008]. Each of them is applied in the following four ways.

- "Target" (**Tar**): Learn the distance metric from target data.
- "Auxiliary" (**Aux**): Learn the distance metrics from labeled auxiliary data.
- "Aggregate" (**Agg**): Learn the distance metrics from the combination of target and auxiliary data.
- "Ensemble" (**Ens**): Construct an "ensemble" combining the distance metrics derived independently from target and auxiliary data.

The semi-supervised metric learning method LRDML has not been applied in "Auxiliary" and "Ensemble" ways since the auxiliary data are fully-labeled and thus no unlabeled example is available.

We treat face recognition as a multi-class classification problem and use the nearest neighborhood classifier to perform the classification. Table 1 provides the performance of all the approaches. We only show the average performance over all t due to the lack of space. From the results, we can find that, by effectively leveraging auxiliary knowledge, the proposed **L-DML** method outperforms all the existing approaches. By exploiting auxiliary knowledge and unlabeled data at the same time, **M-DML** achieves the best overall performance. Fig. (1) shows the detailed performance of **L-DML** and **M-DML** vary with the size of training data. For comparison purpose, the performance of LRDML, which is a semi-supervised method, and LMNN, which achieves the best performance among existing supervised methods, are also illustrated in Fig.(1). From the results, it can be found that **L-DML** and **M-DML** can achieve better performance than LMNN and LRDML with the same number of training examples. From another perspective, the proposed methods need much less training data to reach the same performance than the existing approaches.

Table 1: Performance of two proposed methods, i.e., **L-DML** and **M-DML**, and seven existing approaches in terms of the average accuracy over all the t.

Approach	Experimental Setting			
	Tar	Aux	Agg	Ens
EU	41.9	41.9	41.9	41.9
Xing	44.3	49.7	52.2	46.1
RCA	46.7	51.9	58.2	49.6
DCA	48.3	50.5	57.9	47.4
NCA	53.6	55.1	61.4	51.8
LMNN	56.5	60.8	63.4	59.2
LRDML	61.1	–	67.9	–
L-DML	**69.3**			
M-DML	**75.4**			

Table 2: Performance comparison between using multiple auxiliary knowledge and using single auxiliary knowledge.

Approach	Auxiliary Data				
	ORL	Yale	CMU PIE	Uniform Weighting	Automatic Weighting
L-DML	66.2	61.6	65.3	63.8	69.3
M-DML	72.4	68.1	70.5	70.9	75.4

We further investigate the performance comparison between utilizing multiple auxiliary metrics and single auxiliary metric. Table 2 provides the results of **L-DML** and **M-DML** when they exploit multiple auxiliary metrics and each individual auxiliary metric. We also illustrate the results of **L-DML** and **M-DML** that adopt uniform weights, i.e., assign equal weights to all the three auxiliary metrics (see Eq.(7) and Eq.(10). From the results we can clearly see that exploiting the three auxiliary metrics leads to better results than only using each individual auxiliary metric. We can also see that our automatic weight learning approach outperforms the method of adopting uniform weights. This indicates that the proposed **L-DML** and **M-DML** are able to weight auxiliary knowledge properly.

5 Conclusions

We have developed two novel approaches for learning reliable distance metric from limited training data by exploiting auxiliary knowledge. These approaches are efficient and flexible as they directly utilize the metrics learned from the auxiliary data and require no raw auxiliary examples. Moreover, the proposed methods can effectively take advantage of multiple auxiliary metrics. We apply the proposed approaches to face recognition and compare them with various existing methods. The results show that our methods are more effective than the state-of-art methods for learning robust distance metric from few labeled examples.

References

[Alipanahi *et al.*, 2008] B. Alipanahi, M. Biggs, and A.Ghodsi. Distance metric learning versus fisher discriminant analysis. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[Bar-Hillel *et al.*, 2005] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence

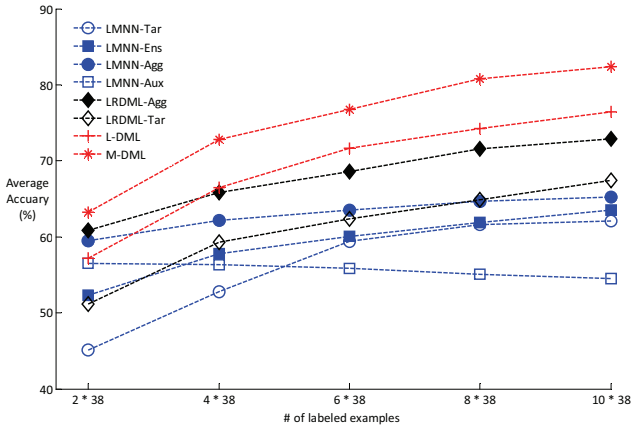

Figure 1: Performance of LRDML, LMNN, and the proposed **L-DML** and **M-DML** methods.

constraints. *Journal of Machine Learning Research (JMLR)*, 6, 2005.

[Belhumeur *et al.*, 1997] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 19(7), 1997.

[Davis *et al.*, 2007] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

[Georghiades *et al.*, 2001] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 23(6), 2001.

[Globerson and Roweis, 2006] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[Goldberger *et al.*, 2004] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[Gupta and Ratinov, 2008] R. Gupta and L.-A. Ratinov. Text categorization with knowledge transfer from heterogeneous data sources. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[Hillel and Weinshall, 2007] A. B. Hillel and D. Weinshall. Learning distance function by coding similarity. In *Proceedings of the International conference on Machine learning (ICML)*. ACM, 2007.

[Hoi *et al.*, 2006] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[Hoi *et al.*, 2008] S. C.H. Hoi, W. Liu, and S.-F. Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Mansour *et al.*, 2007] Y. Mansour, M.r Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[Pan *et al.*, 2008] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[Samaria and Harter, 1994] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *In Proc. IEEE Workshop Applications of Computer Vision*, 1994.

[Satpal and Sarawagi, 2007] S. Satpal and S. Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2007.

[Schultz and Joachims, 2003] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[Sim *et al.*, 2003] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression database. *IEEE Transaction on Pattern Analysis and Machine Intelligence (T-PAMI)*, 25(12):1615–1618, 2003.

[Stephen and Lieven, 2003] B. Stephen and V. Lieven. *Convex Optimization*, volume 5. Cambridge University Press, 2003.

[Sturm, 1999] J. F. Sturm. Using sedumi 1.20, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(12), 1999.

[Talvitie and Singh, 2007] E. Talvitie and S. Singh. An experts algorithm for transfer learning. In *Proceedings of the 20th Int. Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[Weinberger *et al.*, 2006] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NIPS)*. 2006.

[Xing *et al.*, 2003] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[Yang and Jin, 2006] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Michigan State University, 2006.

[Yang and Jin, 2007] L. Yang and R. Jin. Bayesian active distance metric learning. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[Yang, 2007] L. Yang. An overview of distance metric learning. Technical report, Michigan State University, 2007.

[Yeung *et al.*, 2007] D.-Y. Yeung, H. Chang, and G. Dai. A scalable kernel-based algorithm for semi-supervised metric learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.