Local Large-Margin Multi-Metric Learning for Face and Kinship Verification

Junlin Hu, Jiwen Lu, *Senior Member, IEEE*, Yap-Peng Tan, *Senior Member, IEEE*, Junsong Yuan, *Senior Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*—Metric learning has attracted wide attention in face and kinship verification and a number of such algorithms have been presented over the past few years. However, most existing metric learning methods learn only one Mahalanobis distance metric from a single feature representation for each face image and cannot make use of multiple feature representations directly. In many face-related tasks, we can easily extract multiple features for a face image to extract more complementary information, and it is desirable to learn distance metrics from these multiple features so that more discriminative information can be exploited than those learned from individual features. To achieve this, we present a large-margin multi-metric learning (LM^{3}L) method for face and kinship verification, which jointly learns multiple global distance metrics under which the correlations of different feature representations of each sample are maximized, and the distance of each positive pair is less than a low threshold and that of each negative pair is greater than a high threshold. To better exploit the local structures of face images, we also propose a local metric learning (LML) and a local large-margin multi-metric learning (L^{2}M^{3}L) methods to learn a set of local metrics. Experimental results on three face datasets show that the proposed methods achieve very competitive results compared with the state-of-the-art methods.**

***Index Terms*—Local metric learning, multi-metric learning, face verification, kinship verification.**

I. Introduction

LEARNING a promising distance metric from data itself plays an important role in computer vision and pattern recognition. Metric learning techniques have been widely used in many visual analysis applications such as face recognition [1], [2], image classification [3], human activity recognition [4], and kinship verification [5]. Over the past decade, a large number of metric learning algorithms have been proposed and some of them have been successfully applied to face and kinship verification [6], [1], [5], [7]. In facial image analysis, we are usually able to extract multiple feature representations for each face image and it is desirable to learn distance metrics from these multiple feature representations such that more discriminative information can be exploited than those learned from individual features. A widely used solution is to concatenate different features together into a new feature vector and then employ existing distance metric learning algorithms on this concatenated vector directly. However, this concatenation is not physically meaningful because each feature has its own statistical characteristic, and such a simple concatenation ignores the diversity of multiple features and cannot effectively explore the complementary information among the multiple features.

In this paper, we first present a large-margin multi-metric learning (LM^{3}L) method for face verification and kinship verification. Unlike the methods of learning a distance metric on concatenated feature vectors, we collaboratively learn multiple distance metrics from multiple feature representations of data, where one distance metric is learned for each feature and the correlations of different feature representations of each sample are maximized, and under the learned metric spaces the distance of each positive pair is less than a smaller threshold and that of each negative face pair is more than a larger threshold, respectively. In addition, we also propose two local distance metric learning approaches, i.e., local metric learning (LML) and a local large-margin multi-metric learning (L^{2}M^{3}L), to better exploit the local manifold structures of face images. Experimental results on three widely used face datasets show that our methods can obtain competitive results compared with state-of-the-art methods. See Fig. 1 of [8] for the basic pipeline of our multi-metric learning methods.

This paper is an extension to our conference paper [8], where we only learn a global distance metric for each single-view feature of samples. The new contributions of this paper are summarized as: 1) we have presented a local metric learning (LML) for each single-view feature by jointly learning a global and several local metrics to better exploit the local manifold structure that face images usually lie on; 2) we have proposed a local large-margin multi-metric learning (L^{2}M^{3}L) method by integrating the LML into the LM^{3}L method, and the LM^{3}L is a special case of the L^{2}M^{3}L method where only multiple global metrics are jointly learned; and 3) we have conducted more experiments on three datasets for face and kinship verification tasks to show the effectiveness of the proposed methods.

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074 and Grant 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program. (*Corresponding author: Jiwen Lu.*)

Junlin Hu, Yap-Peng Tan, and Junsong Yuan are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798. E-mail: jhu007@e.ntu.edu.sg; eyptan@ntu.edu.sg; jsyuan@ntu.edu.sg.

Jiwen Lu and Jie Zhou are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, 100084, China. E-mail: lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn.

II. Related Work

A. Face and Kinship Verification

These days face verification under uncontrolled conditions is a mainstream task of face recognition [9], [1], [10], [2], [11], [12], [13], which aims to determine if two face images/videos are from the same subject or not. Kinship verification from facial images is another challenging face analysis problem [14], [15], [16], [17], [18], and its goal is to decide whether there is a kinship relation between two individuals via their face images. In recent years, many methods have been proposed for face and kinship verification under uncontrolled conditions [10], [6], [19], [20], [21], [22]. Most of these methods mainly focus on feature representation and similarity/metric learning, which are two important steps in the pipeline of the face/kinship verification. Typical feature descriptors include local binary pattern (LBP) [23], probabilistic elastic matching (PEM) [24], etc. The similarity/metric learning step aims to learn one or more metrics from the training data to help improve the verification accuracy [1], [19], [17], [20]. In this paper, we propose a multi-metric learning method to learn multiple distance metrics for face and kinship verification under uncontrolled conditions.

B. Metric Learning

A number of metric learning methods have been introduced in the literature recently, and most of them seek an appropriate global distance metric to exploit discriminative information from the training samples. Representative metric learning methods include large margin nearest neighbor (LMNN) [25], information theoretic metric learning (ITML) [26], logistic discriminant metric learning (LDML) [1], pairwise constrained component analysis (PCCA) [27], neighborhood repulsed metric learning (NRML) [5], similarity metric learning (SM-L) [19], and deep metric learning [28], [29]. Recently several local metric learning methods [30], [31] have been proposed to model the local specificities of the data points by learning a set of local distance metrics for a single feature. While these methods have achieved encouraging performance in face verification, most of them learn one global metric or multiple local metrics from the single-view feature representation and cannot exploit multi-view feature representations directly. Unlike these single feature based methods, we present a multi-metric learning approach by collaboratively learning multiple global and local distance metrics to better exploit complementary information from multiple feature representations for face and kinship verification in the wild.

III. Large-Margin Multi-Metric Learning

Before detailing our method, we first list the notations used in this paper. Bold capital letters, e.g., $\mathbf{X}_1$, $\mathbf{X}_2$, represent matrices, and bold lower case letters, e.g., $\mathbf{x}_1$, $\mathbf{x}_2$, represent column vectors. Given a training set containing N data points, i.e., $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, each data point of this set can be easily represented by the multiple types of features, e.g., color, texture, shape, etc. Let $\mathcal{X}_k = \{\mathbf{x}_i^k \in \mathbb{R}^{d_k}\}_{i=1}^N$ be the k-th feature set of $\mathcal{X}$ from the k-th type of feature representation, and let $\mathbf{X}_k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \cdots, \mathbf{x}_N^k]$ be the feature matrix of set $\mathcal{X}_k$, where $\mathbf{x}_i^k$ is the feature vector of the data point $\mathbf{x}_i$ in the k-th feature space, $k = 1, 2, \cdots, K$; K is the total number of types of features; and d_k is feature dimension of $\mathbf{x}_i^k$.

A. Problem Formulation

For a feature set $\mathcal{X}_k = \{\mathbf{x}_i^k \in \mathbb{R}^{d_k}\}_{i=1}^N$ from the k-th feature representation, the squared Mahalanobis distance between a pair of samples $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ can be computed as:

$$d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k) = (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \mathbf{M}_k (\mathbf{x}_i^k - \mathbf{x}_j^k), \quad (1)$$

where $\mathbf{M}_k \in \mathbb{R}^{d_k \times d_k}$ is a positive definite matrix.

We seek a distance metric $\mathbf{M}_k$ such that the squared distance $d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k)$ for a face pair $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ in the kth feature space should be smaller than a given threshold $\mu_k - \tau_k$ ($\mu_k > \tau_k > 0$) if two samples are from the same subject, and larger than a threshold $\mu_k + \tau_k$ if these two samples are from different subjects, which can be formulated as the following constraints:

$$y_{ij}\big(\mu_k - d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k)\big) > \tau_k, \quad (2)$$

where pairwise label $y_{ij} = 1$ if $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ are from the same category (or similar pair), and $y_{ij} = -1$ if they are from different categories (or dissimilar pair). The parameter μ_k is an absolute threshold to decide whether two samples are similar pair or not. The parameter τ_k is a positive slack variable to guarantee a margin (i.e., $2\tau_k$) between a similar pair and a dissimilar pair.

To learn $\mathbf{M}_k$, we define the constraints (2) by a hinge loss, and formulate the following objective function to learn the k-th distance metric, named single metric learning (SML):

$$\min_{\mathbf{M}_k} J_k = \sum_{i,j} h\Big(\tau_k - y_{ij}\big(\mu_k - d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k)\big)\Big), \quad (3)$$

where $h(x) = \max(x, 0)$ represents the hinge loss function. The objective function (3) penalizes the violation of the constraint $d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k) > \mu_k - \tau_k$ for a similar pair and that of the constraint $d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k) < \mu_k + \tau_k$ for a dissimilar pair by using the hinge loss.

Given a face image, it is easy to extract multiple features for each image for multiple feature fusion. These features extracted from the same face image are usually highly correlated to each other even if they characterize face images from different aspects [32]. For multiple feature fusion, these highly correlated information should be preserved because they usually reflect the intrinsic information of samples. An important principle to perform multi-feature metric learning is to jointly learn multiple distance metrics by preserving the correlation between different feature pairs. Motivated by the success of canonical correlation analysis (CCA), we propose a large-margin multi-metric learning method to seek the commonality of multiple feature representations, which is consistent to the CCA-based multiple feature fusion methods [32], [33].

The proposed large-margin multi-metric learning (LM^{3}L) method aims to learn K distance metrics $\{\mathbf{M}_k \in \mathbb{R}^{d_k \times d_k}\}_{k=1}^K$ for a multi-feature dataset, such that 1) the discriminative information from each single feature can be exploited as much as possible; and 2) the differences of different feature representations of each sample in the learned distance metrics are minimized, because different features of each sample share the same semantic label.

Since the difference computation of the sample $\mathbf{x}_i$ from the k-th and ℓ-th ($1 \leq k, \ell \leq K,\ k \neq \ell$) feature representations relies on the distance metrics $\mathbf{M}_k$ and $\mathbf{M}_\ell$, which could be different in dimensions, it is infeasible to compute them directly. To address this, we use an alternative constraint to reflect the relationships of different feature representations. Since the distance between $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$, and that of $\mathbf{x}_i^\ell$ and $\mathbf{x}_j^\ell$ are expected to be as small as possible, hence, we formulate the following objective function to constrain the interactions of different distance metrics in our LM^{3}L method:

$$\begin{aligned}\min_{\{\mathbf{M}_k,\ w_k\}_{k=1}^K} \ & J = \sum_{k=1}^K w_k\ J_k \\ & + \lambda \sum_{\substack{k,\ell=1 \\ k<\ell}}^K \sum_{i,j} \Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell) \Big)^2, \\ \text{s.t.} \ & \sum_{k=1}^K w_k = 1,\ w_k \geq 0,\ \lambda > 0, \end{aligned} \tag{4}$$

where w_k is a nonnegative weighting parameter to reflect the importance of the k-th feature in the whole objective function, and λ weights the pairwise difference of the distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the learned distance metrics $\mathbf{M}_k$ and $\mathbf{M}_\ell$. The physical meaning of (4) is that we aim to learn K distance metrics $\{\mathbf{M}_k\}_{k=1}^K$ under which the difference of feature representations of each pair of face samples is enforced to be as small as possible. The reason that a sample x_i should be close in different feature spaces (k and ℓ) is to seek a commonality of multiple features and make all the features more robust, which is consistent to the CCA-based multiple feature fusion methods.

Having obtained the multiple distance metrics $\{\mathbf{M}_k\}_{k=1}^K$ and their weights $\{w_k\}_{k=1}^K$, the distance between two multi-feature data points $\mathbf{x}_i$ and $\mathbf{x}_j$ under the global metrics learned by the LM^{3}L is computed as:

$$\begin{aligned} d^2_{\mathrm{LM^3L}}(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{k=1}^K w_k\ d^2_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ &= \sum_{k=1}^K w_k\ (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \mathbf{M}_k (\mathbf{x}_i^k - \mathbf{x}_j^k). \end{aligned} \tag{5}$$

The trivial solution of (4) is $w_k = 1$, which corresponds to the minimum J_k over different feature representations, and $w_k = 0$ otherwise. This solution means that only one single feature that yields the best verification accuracy is selected, which does not satisfy our objective on exploring the complementary property of multi-feature data.

To address this shortcoming, we modify w_k to be w_k^p ($p > 1$), then the new objective function is rewritten as:

$$\begin{aligned}\min_{\{\mathbf{M}_k,\ w_k\}_{k=1}^K} \ & J = \sum_{k=1}^K w_k^p\ J_k \\ & + \lambda \sum_{\substack{k,\ell=1 \\ k<\ell}}^K \sum_{i,j} \Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell) \Big)^2, \\ \text{s.t.} \ & \sum_{k=1}^K w_k = 1,\ w_k \geq 0,\ \lambda > 0. \end{aligned} \tag{6}$$

When $p = 1$, it is not easy to obtain the optimal w_k. From (17), we obtain $J_1 = J_2 = \cdots = J_K = \eta$, then w_k can be an arbitrary value in the interval [0, 1]. Thus, the trivial solution is $w_k = 1$ for the minimum J_k over different feature representations, and $w_k = 1$ otherwise. When $p > 1$, we obtain a closed-form solution of w_k by (19), and each feature representation has a particular contribution to the final metric learning.

B. Alternating Optimization

To the best of our knowledge, it is non-trivial to seek a global optimal solution to (6) because there are K metrics to be learned simultaneously. In this work, we employ an iterative method by using the alternating optimization method to obtain a local optimal solution. The alternating optimization learns $\mathbf{M}_k$ and w_k in an iterative manner. In our experiments, we randomly select the order of different features to start the optimization procedure and our tests show that the influence of this order is not critical to the final verification performance.

1) Step 1: Fix $\mathbf{w} = [w_1, w_2, \cdots, w_K]$, *Update* $\mathbf{M}_k$*:* With the fixed $\mathbf{w}$, we can cyclically optimize (6) over different features. We sequentially optimize $\mathbf{M}_k$ with the fixed $\mathbf{M}_1$, $\cdots$, $\mathbf{M}_{k-1}$, $\mathbf{M}_{k+1}$, $\cdots$, $\mathbf{M}_K$. Hence, (6) can be rewritten as:

$$\begin{aligned}\min_{\mathbf{M}_k} \ & J = A_k\ +\ w_k^p\ J_k \\ & + \lambda \sum_{\substack{\ell=1 \\ \ell\neq k}}^K \sum_{i,j} \Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell) \Big)^2, \end{aligned} \tag{7}$$

where A_k is a constant term.

To learn the distance metric $\mathbf{M}_k$, we employ a gradient-based scheme. After some algebraic simplification, we can obtain the gradient as:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{M}_k} &= w_k^p \sum_{i,j} y_{ij} h'(z_{ij}) \mathbf{C}_{ij}^k \\ & + \lambda \sum_{\substack{\ell=1 \\ \ell\neq k}}^K \sum_{i,j} \Bigg(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)} \Bigg) \mathbf{C}_{ij}^k, \end{aligned} \tag{8}$$

where z_{ij} and $\mathbf{C}_{ij}^k$ can be calculated respectively by:

$$\mathbf{C}_{ij}^k = (\mathbf{x}_i^k - \mathbf{x}_j^k)(\mathbf{x}_i^k - \mathbf{x}_j^k)^T, \tag{9}$$

$$z_{ij} = \tau_k - y_{ij}\big(\mu_k - d^2_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)\big). \tag{10}$$

$\mathbf{C}_{ij}^k$ denotes the outer product of pairwise differences. $h'(x)$ is the derivative of $h(x)$, $h'(0) = 0$ at the non-differentiable point. In addition, we use derivations given as:

$$\frac{\partial}{\partial \mathbf{M}_k} d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) = \frac{1}{2\ d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)}\ \mathbf{C}_{ij}^k, \tag{11}$$

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{M}_k}\Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)\Big)^2 \\ &= 2\ \Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)\Big)\ \frac{\partial}{\partial \mathbf{M}_k} d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ &= \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)}\right)\ \mathbf{C}_{ij}^k. \end{aligned} \tag{12}$$

Then, matrix $\mathbf{M}_k$ can be obtained by using a gradient descent algorithm:

$$\mathbf{M}_k = \mathbf{M}_k - \beta\ \frac{\partial J}{\partial \mathbf{M}_k}, \tag{13}$$

where β is the learning rate.

In practice, directly optimizing the Mahalanobis distance metric $\mathbf{M}_k$ may suffer slow convergence and overfitting problems if data is very high-dimensional and the number of training samples is insufficient. Therefore, we propose an alternative method to jointly perform dimensionality reduction and metric learning, which means a low-rank linear projection matrix $\mathbf{L}_k \in \mathbb{R}^{s_k \times d_k}$ $(s_k < d_k)$ is learned to project each sample $\mathbf{x}_i^k$ from the high-dimensional input space to a low-dimensional embedding space, where the distance metric $\mathbf{M}_k = {\mathbf{L}_k}^T \mathbf{L}_k$. Then, we differentiate the objective function J with respect to $\mathbf{L}_k$, and obtain the gradient as follows:

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{L}_k} = 2\mathbf{L}_k\ \Bigg(& w_k^p \sum_{i,j} y_{ij} h'(z_{ij}) \mathbf{C}_{ij}^k \\ & + \lambda \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \sum_{i,j} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)}\right) \mathbf{C}_{ij}^k \Bigg). \end{aligned} \tag{14}$$

Lastly, $\mathbf{L}_k$ can be obtained by a gradient descent rule:

$$\mathbf{L}_k = \mathbf{L}_k - \beta\ \frac{\partial J}{\partial \mathbf{L}_k}. \tag{15}$$

To make sure the learned metric $\mathbf{M}_k$ is a positive semidefinite matrix after each iteration, we clip the spectrum of $\mathbf{M}_k = {\mathbf{L}_k}^T \mathbf{L}_k$ by singular value decomposition.

2) Step 2: Fix $\{\mathbf{M}_k\}_{k=1}^K$, Update $\mathbf{w} = [w_1, w_2, \cdots, w_K]$: Then, we update $\mathbf{w}$ with the fixed $\{\mathbf{M}_k\}_{k=1}^K$ by the method of Lagrange multipliers. We construct a Lagrange function as:

$$L(\mathbf{w}, \eta) = \sum_{k=1}^{K} w_k^p\ J_k\ +\ A\ -\ \eta\ \Big(\sum_{k=1}^{K} w_k - 1\Big), \tag{16}$$

in which A is a constant term.

Let $\frac{\partial L(\mathbf{w},\eta)}{\partial w_k} = 0$ and $\frac{\partial L(\mathbf{w},\eta)}{\partial \eta} = 0$, we have

$$\frac{\partial L(\mathbf{w}, \eta)}{\partial w_k} = p\ w_k^{p-1}\ J_k - \eta = 0, \tag{17}$$

$$\frac{\partial L(\mathbf{w}, \eta)}{\partial \eta} = \sum_{k=1}^{K} w_k - 1 = 0. \tag{18}$$

Algorithm 1: LM^{3}L

```
Input: Training set $\{\mathcal{X}_k\}_{k=1}^K$ from $K$ views; Learning
       rate $\beta$; Parameter $p$, $\lambda$, $\mu_k$ and $\tau_k$; Total iterative
       number $T$; Convergence error $\varepsilon$.
Output: Multiple metrics: $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_K$; and
        weights: $w_1, w_2, \cdots, w_K$.
// Initialization:
       Initialize $\mathbf{L}_k = \mathbf{I}^{s_k \times d_k}$,
       $w_k = 1/K$, $k = 1, \cdots, K$.
// Alternating optimization:
       for $t = 1, 2, \cdots, T$, do
           for $k = 1, 2, \cdots, K$, do
               Compute $\mathbf{L}_k$ by (14) and (15).
           end for
           Compute $\mathbf{w}$ according to (19).
           Computer $J^{(t)}$ via (6).
           if $t > 1$ and $|J^{(t)} - J^{(t-1)}| < \varepsilon$
               Go to Output.
           end if
       end for
// Output distance metrics and weights:
       $\mathbf{M}_k = {\mathbf{L}_k}^T \mathbf{L}_k$, $k = 1, 2, \cdots, K$.
       Output $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_K$ and $\mathbf{w}$.
```

According to (17) and (18), w_k can be updated as:

$$w_k = \frac{\big(1/J_k\big)^{1/(p-1)}}{\sum\limits_{k=1}^{K} \big(1/J_k\big)^{1/(p-1)}}. \tag{19}$$

We repeat the above two steps until the algorithm meets a certain convergence condition. The proposed LM^{3}L algorithm is summarized in Algorithm 1, where $\mathbf{I} \in \mathbb{R}^{s_k \times d_k}$ is a matrix with 1's on the diagonal and zeros elsewhere.

IV. Local Large-Margin Multi-Metric Learning

A. Local Metric Learning

Considering a single feature set $\mathcal{X}_k = \{\mathbf{x}_i^k \in \mathbb{R}^{d_k}\}_{i=1}^N$ of $\mathcal{X}$ which is represented by the k-th type of feature, the squared Mahalanobis distance between a pair of samples $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ under a specific local metric $\mathbf{M}_k^{(q)} \in \mathbb{R}^{d_k \times d_k}$, $q = 1, 2, \cdots, Q_k$, can be calculated by:

$$d^2_{\mathbf{M}_k^{(q)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) = (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \mathbf{M}_k^{(q)} (\mathbf{x}_i^k - \mathbf{x}_j^k), \tag{20}$$

where $\mathbf{M}_k^{(q)} \in \mathbb{R}^{d_k \times d_k}$ is a positive semi-definite (PSD) matrix ($\mathbf{M}_k^{(q)} \succeq \mathbf{0}$), and Q_k is the total number of local distance metrics corresponding to the k-th type of feature representation.

Based on a set of local distance metrics $\{\mathbf{M}_k^{(q)}\}_{q=1}^{Q_k}$, the distance of each sample pair $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ can be defined as a convex combination by soft dividing the whole input space into different regions:

$$\begin{aligned} & d^2_{\{\mathbf{M}_k^{(q)}\}_{q=1}^{Q_k}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ & = \sum_{q=1}^{Q_k} \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k)\ d^2_{\mathbf{M}_k^{(q)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ & = (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \left(\sum_{q=1}^{Q_k} \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{M}_k^{(q)} \right) (\mathbf{x}_i^k - \mathbf{x}_j^k), \end{aligned} \tag{21}$$

where $\alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k)$ is nonnegative weight to measure the importance of the q-th local metric to both $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$, which ensures the learned dissimilarity function of $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ to be local. In addition, a global metric $\mathbf{M}_k^{(0)}$ with a positive constant weight $\alpha_k^{(0)}(\mathbf{x}_i^k, \mathbf{x}_j^k) = c_k$ is complemented into (21) to handle the part of the dissimilarity function shared by the whole input space. Thus, the final distance (or dissimilarity function) of a pair $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ is given as:

$$\begin{aligned} & d^2_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ & = (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \left(\sum_{q=0}^{Q_k} \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{M}_k^{(q)} \right) (\mathbf{x}_i^k - \mathbf{x}_j^k) \\ & = (\mathbf{x}_i^k - \mathbf{x}_j^k)^T\ \mathbf{M}_k(\mathbf{x}_i^k, \mathbf{x}_j^k)\ (\mathbf{x}_i^k - \mathbf{x}_j^k), \end{aligned} \tag{22}$$

where the PSD matrix $\mathbf{M}_k(\cdot,\cdot) \in \mathbb{R}^{d_k \times d_k}$ is a matrix-valued function, and it is weighted by $Q_k + 1$ matrices for a sample pair $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$ as follows:

$$\mathbf{M}_k(\mathbf{x}_i^k, \mathbf{x}_j^k) = \left(\sum_{q=0}^{Q_k} \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{M}_k^{(q)} \right). \tag{23}$$

Importantly, the weight $\alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k)$ in the local distance function (23) is defined as:

$$\alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) = \left\{ \begin{array}{ll} c_k & \text{if } q = 0 \\ u_k^{(q)}(\mathbf{x}_i^k)\ u_k^{(q)}(\mathbf{x}_j^k) & \text{otherwise} \end{array} \right., \tag{24}$$

in which $u_k^{(q)}(\mathbf{x}_i^k)$ is a gating model to assign weight to the q-th local distance metric in a data-dependent way. In our experiments, we adopt a softmax gating function [34], which is given as follows:

$$u_k^{(q)}(\mathbf{x}_i^k) = \frac{\exp\left({\mathbf{v}_k^{(q)}}^T \mathbf{x}_i^k + b_k^{(q)} \right)}{\sum_{m=1}^{Q_k} \exp\left({\mathbf{v}_k^{(m)}}^T \mathbf{x}_i^k + b_k^{(m)} \right)}, \tag{25}$$

for the q-th local metric, where $\mathbf{v}_k^{(q)}$ and $b_k^{(q)}$ are the weighting and bias parameters of this gating function respectively, and we have $u_k^{(q)}(\mathbf{x}_i^k) \geq 0$ for $1 \leq q \leq Q_k$ and $1 \leq k \leq K$. This gating function parametrized by the parameters $\{\mathbf{v}_k^{(q)}, b_k^{(q)}\}_{q=1}^{Q_k}$ is used to compute the weight (or probability) that an input vector $\mathbf{x}_i^k$ belongs to each local distance metric space. For example, a face image can be assigned to multiple metric spaces (e.g., age, expression, gender, race, etc.) with different weights. After having learned the parameters $\{\mathbf{v}_k^{(q)}, b_k^{(q)}\}_{q=1}^{Q_k}$ on the training data, this softmax gating function can be easily applied to other input vectors in a data-dependent way.

Then, we formulate our local metric learning (LML) method with respect to the single feature type k under the same large margin framework as used in (3):

$$\begin{aligned} \min_{\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k},\ \{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k}} & J_k = \\ & \sum_{i,j} h\Big(\tau_k - y_{ij}\big(\mu_k - d^2_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)\big) \Big). \end{aligned} \tag{26}$$

The objective function (26) is not jointly convex to $\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k}$ and $\{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k}$, and it is non-trivial to find a global solution. To obtain these parameters, we employ the alternating optimization strategy and the gradient descent based method to achieve the local optimal solution.

1) Step 1: Fix $\{\mathbf{M}_k^{(r)}\}_{r=0}^{Q_k} \setminus \mathbf{M}_k^{(q)}$ *and* $\{\mathbf{v}_k^{(r)},\ b_k^{(r)}\}_{r=1}^{Q_k}$, *Update* $\mathbf{M}_k^{(q)}$*:* The partial derivative of J_k with regard to $\mathbf{M}_k^{(q)}$, $0 \leq q \leq Q_k$, can be calculated by:

$$\frac{\partial J_k}{\partial \mathbf{M}_k^{(q)}} = \sum_{i,j} y_{ij} h'(z_{ij}) \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{C}_{ij}^k, \tag{27}$$

where $z_{ij} = \tau_k - y_{ij}\big(\mu_k - d^2_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)\big)$, and $\mathbf{C}_{ij}^k$ is the outer product of pairwise differences given by (9).

2) Step 2: Fix $\{\mathbf{M}_k^{(r)}\}_{r=0}^{Q_k}$ *and* $\{\mathbf{v}_k^{(r)},\ b_k^{(r)}\}_{r=1}^{Q_k} \setminus \{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}$, *Update* $\mathbf{v}_k^{(q)}$ *and* $b_k^{(q)}$*:* The partial derivative of J_k with regard to $\mathbf{v}_k^{(q)}$ and $b_k^{(q)}$ can be computed by:

$$\begin{aligned} \frac{\partial J_k}{\partial \mathbf{v}_k^{(q)}} = & \sum_{i,j} y_{ij} h'(z_{ij}) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ & \times \Big([\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k)]\mathbf{x}_i^k + [\delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k)]\mathbf{x}_j^k \Big), \end{aligned} \tag{28}$$

$$\begin{aligned} \frac{\partial J_k}{\partial b_k^{(q)}} = & \sum_{i,j} y_{ij} h'(z_{ij}) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \\ & \times \Big(\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k) + \delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k) \Big), \end{aligned} \tag{29}$$

in which the delta function $\delta(q-m) = 1$ if $q = m$ and 0 otherwise for $q = 1, 2, \cdots, Q_k$.

We repeat the above two steps until the algorithm reaches certain convergence conditions. Moreover, we also decompose $\mathbf{M}_k^{(q)}$ into $\mathbf{M}_k^{(q)} = {\mathbf{L}_k^{(q)}}^T \mathbf{L}_k^{(q)}$ in the objective function (26) to reduce the number of parameters in optimization.

B. Local Large-Margin Multi-Metric Learning

Local metric learning (LML) only learns a set of local distance metrics for each type of feature such that it cannot exploit discriminative information of other types of features. To utilize multiple features, we also propose a local large-margin multi-metric learning (L^{2}M^{3}L) method by integrating the local metric learning (LML) (26) and the large-margin multi-metric learning (LM^{3}L) (6) into a unified framework.

```
Algorithm 2: L$^2$M$^3$L
Input: Training set $\{\mathcal{X}_k\}_{k=1}^K$ from $K$ views; Local
    metric number $\{Q_k\}_{k=1}^K$; Learning rate $\beta$;
    Parameter $p$, $\lambda$, $\mu_k$, $\tau_k$; Total iterative number $T$;
    Convergence error $\varepsilon$.
Output: Metrics: $\Big\{\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k}\Big\}_{k=1}^K$; Weights: $\{w_k\}_{k=1}^K$;
    Gating model: $\Big\{\{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k}\Big\}_{k=1}^K$.
// Initialization:
    Initialize $\{\mathbf{M}_k^{(q)} = \mathbf{I}^{d_k \times d_k}\}_{q=0}^{Q_k}$ and $w_k = 1/K$,
    $\{\mathbf{v}_k^{(q)} \sim U(0,1),\ b_k^{(q)} = 0\}_{q=1}^{Q_k}$, $k = 1, 2, \cdots, K$.
// Alternating optimization:
    for $t = 1, 2, \cdots, T$, do
        for $k = 1, 2, \cdots, K$, do
            for $q = 0, 1, \cdots, Q_k$, do
                // Step 1: Update $\mathbf{M}_k^{(q)}$
                Calculate $\partial J / \partial \mathbf{M}_k^{(q)}$ by (31).
                $\mathbf{M}_k^{(q)} \longleftarrow \mathbf{M}_k^{(q)} - \beta\ \partial J / \partial \mathbf{M}_k^{(q)}$.
            end for
        end for
        for $k = 1, 2, \cdots, K$, do
            for $q = 1, 2, \cdots, Q_k$, do
                // Step 2: Update $\mathbf{v}_k^{(q)}$ and $b_k^{(q)}$
                Compute $\partial J / \partial \mathbf{v}_k^{(q)}$ by (32).
                Compute $\partial J / \partial b_k^{(q)}$ by (33).
                $\mathbf{v}_k^{(q)} \longleftarrow \mathbf{v}_k^{(q)} - \beta\ \partial J / \partial \mathbf{v}_k^{(q)}$.
                $b_k^{(q)} \longleftarrow b_k^{(q)} - \beta\ \partial J / \partial b_k^{(q)}$.
            end for
        end for
        // Step 3: Update $\{w_k\}_{k=1}^K$
        Calculate $\{w_k\}_{k=1}^K$ by (34).
        Calculate objective $J^{(t)}$ using (30).
        if $t > 1$ and $|J^{(t)} - J^{(t-1)}| < \varepsilon$
            Go to Output.
        end if
    end for
    Output $\Big\{\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k},\ \{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k},\ w_k\Big\}_{k=1}^K$.
```

The objective function of the L^{2}M^{3}L is formulated as:

$$\min_{\Big\{\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k},\ \{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k},\ w_k\Big\}_{k=1}^K} J = \sum_{k=1}^K w_k^p\ J_k\ +\ \lambda \sum_{\substack{k,\ell=1\\ k<\ell}}^K \sum_{i,j} \Big(d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k) - d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)\Big)^2,$$

$$\text{s.t.} \quad \sum_{k=1}^K w_k = 1,\ w_k \geq 0,\ \lambda > 0. \tag{30}$$

It is obviously that the LM^{3}L is a special case of the L^{2}M^{3}L method where only several global distance metrics are jointly solved. To minimize the optimization problem (30), we adopt similar methods as used in both LM^{3}L and LML.

1) Step 1: Fix $\Big\{\{\mathbf{M}_k^{(r)}\}_{r=0}^{Q_k},\ \{\mathbf{v}_k^{(r)},\ b_k^{(r)}\}_{r=1}^{Q_k},\ w_k\Big\}_{k=1}^K \setminus \mathbf{M}_k^{(q)}$*, Update* $\mathbf{M}_k^{(q)}$*:* We update $\mathbf{M}_k^{(q)}$, $1 \leq k \leq K$, $0 \leq q \leq Q_k$, by fixing other parameters. The partial derivative of J with regard to $\mathbf{M}_k^{(q)}$ can be calculated by:

$$\frac{\partial J}{\partial \mathbf{M}_k^{(q)}} = w_k^p \sum_{i,j} y_{ij} h'(z_{ij}) \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{C}_{ij}^k + \lambda \sum_{\substack{\ell=1\\ \ell\neq k}}^K \sum_{i,j} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)}\right) \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{C}_{ij}^k, \tag{31}$$

where $z_{ij} = \tau_k - y_{ij}\big(\mu_k - d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k)\big)$, and $\mathbf{C}_{ij}^k$ is the outer product of pairwise differences given by (9).

2) Step 2: Fix $\Big\{\{\mathbf{M}_k^{(r)}\}_{r=0}^{Q_k},\ \{\mathbf{v}_k^{(r)},\ b_k^{(r)}\}_{r=1}^{Q_k},\ w_k\Big\}_{k=1}^K \setminus \{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}$*, Update* $\mathbf{v}_k^{(q)}$ *and* $b_k^{(q)}$*:* The partial derivative of J with respect to $\mathbf{v}_k^{(q)}$ and $b_k^{(q)}$, $1 \leq k \leq K$, $1 \leq q \leq Q_k$, can be computed by (32) and (33), where $z_{ij} = \tau_k - y_{ij}\big(\mu_k - d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k)\big)$.

3) Step 3: Fix $\Big\{\{\mathbf{M}_k^{(r)}\}_{r=0}^{Q_k},\ \{\mathbf{v}_k^{(r)},\ b_k^{(r)}\}_{r=1}^{Q_k}\Big\}_{k=1}^K$*, Update* $\{w_k\}_{k=1}^K$*:* Following the same procedures as in LM^{3}L, we update w_k by utilizing the method of Lagrange multipliers. In this fashion, the closed-form solution can be obtained, and the w_k, $1 \leq k \leq K$, is given by:

$$w_k = \frac{\big(1/J_k\big)^{1/(p-1)}}{\sum\limits_{k=1}^K \big(1/J_k\big)^{1/(p-1)}}. \tag{34}$$

Then, we alternately update above three steps until the proposed L^{2}M^{3}L method reaches a certain convergence condition, and then we will find the optimal solution, $\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k}$, $\{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k}$, w_k, $k = 1, 2, \cdots, K$, of the L^{2}M^{3}L method. The Algorithm 2 lists the main steps of the L^{2}M^{3}L method.

After obtaining the multiple local and global distance metrics: $\{\mathbf{M}_k^{(q)}\}_{q=0}^{Q_k}$, gating model: $\{\mathbf{v}_k^{(q)},\ b_k^{(q)}\}_{q=1}^{Q_k}$, and weight w_k for all the $k = 1, 2, \cdots, K$, the distance of two multi-feature data points $\mathbf{x}_i$ and $\mathbf{x}_j$ under the learned local metrics by L^{2}M^{3}L can be calculated as follows:

$$\begin{aligned} & d_{\mathrm{L^2M^3L}}^2(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{k=1}^K w_k\ (\mathbf{x}_i^k - \mathbf{x}_j^k)^T \left(\sum_{q=0}^{Q_k} \alpha_k^{(q)}(\mathbf{x}_i^k, \mathbf{x}_j^k) \mathbf{M}_k^{(q)}\right) (\mathbf{x}_i^k - \mathbf{x}_j^k) \\ &= \sum_{k=1}^K w_k\ d_{\mathbf{M}_k}^2(\mathbf{x}_i^k, \mathbf{x}_j^k). \end{aligned} \tag{35}$$

V. Experiments

To evaluate the effectiveness of the proposed LM^{3}L, LML and L^{2}M^{3}L methods, we conducted face and kinship verification under unconstrained conditions on three real-world face datasets, i.e., the Labeled Faces in the Wild (LFW) [9], the YouTube Faces (YTF) [10], and the KinFaceW-II [5].

Baseline: We evaluated the proposed methods with three baseline methods using different metric learning strategies:

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{v}_k^{(q)}} &= w_k^p \sum_{i,j} y_{ij} h'(z_{ij}) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \Big([\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k)]\mathbf{x}_i^k + [\delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k)]\mathbf{x}_j^k \Big) \; + \\
&\quad \lambda \sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \sum_{i,j} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)} \right) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \Big([\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k)]\mathbf{x}_i^k + [\delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k)]\mathbf{x}_j^k \Big) \\
&= \sum_{i,j} \left(w_k^p y_{ij} h'(z_{ij}) + \lambda \sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)} \right) \right) \\
&\quad \times \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \left([\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k)]\mathbf{x}_i^k + [\delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k)]\mathbf{x}_j^k \right), \qquad (32)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J}{\partial b_k^{(q)}} &= w_k^p \sum_{i,j} y_{ij} h'(z_{ij}) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \Big(\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k) + \delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k) \Big) \; + \\
&\quad \lambda \sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \sum_{i,j} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)} \right) \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \Big(\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k) + \delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k) \Big) \\
&= \sum_{i,j} \left(w_k^p y_{ij} h'(z_{ij}) + \lambda \sum_{\substack{\ell=1\\ \ell\neq k}}^{K} \left(1 - \frac{d_{\mathbf{M}_\ell}(\mathbf{x}_i^\ell, \mathbf{x}_j^\ell)}{d_{\mathbf{M}_k}(\mathbf{x}_i^k, \mathbf{x}_j^k)} \right) \right) \\
&\quad \times \sum_{m=1}^{Q_k} u_k^{(m)}(\mathbf{x}_i^k) u_k^{(m)}(\mathbf{x}_j^k) d^2_{\mathbf{M}_k^{(m)}}(\mathbf{x}_i^k, \mathbf{x}_j^k) \left(\delta(q-m) - u_k^{(q)}(\mathbf{x}_i^k) + \delta(q-m) - u_k^{(q)}(\mathbf{x}_j^k) \right), \qquad (33)
\end{aligned}$$

- Single Metric Learning (SML): we learned a single distance metric by using objective function (3) with single-view feature representation;
- Concatenated Metric Learning (CML): we first concatenated different feature representations into a long feature vector and then employed objective function (3) to learn a distance metric;
- Individual Metric Learning (IML): we learned the distance metric for each feature representation under the objective function (3) and then adopted the same weight (i.e., $w_k = 1/K$) to calculate the dissimilarity between a pair of face images by (5).

In addition, several free parameters p, β, λ, μ_k, τ_k, Q_k and c_k of our LM^{3}L, LML and L^{2}M^{3}L methods were empirically set as 2, 0.001, 0.1, 5, 1, 3 and 1 for all the $k = 1, 2, \cdots, K$ respectively, unless stated otherwise. The following subsections detail the experimental settings and results on three datasets.

A. Face Verification on LFW

1) Dataset and Settings: The LFW dataset [9] contains more than 13000 face images of 5749 subjects collected from the web with large variations in expression, pose, age, illumination, resolution, etc. There are two training paradigms for supervised learning on this dataset: *image restricted* and *image unrestricted*. In our experiments, we used the *image restricted* setting where only the pairwise label information is provided to train our methods. We followed the standard evaluation protocol on the "View 2" dataset [9] which consists of 3000 matched pairs (or positive pairs) and 3000 mismatched pairs (or negative pairs), and all these pairs were divided into 10 folds and each fold contains 300 positive pairs and 300 negative pairs of face iamges. We used the LFW-a dataset for our experimental evaluation, and thus our setting on this dataset falls into the category of *image-restricted, label-free outside data*. For each face image, we first cropped it into 80×150 pixels from its center to remove the background information, and then extracted three types of feature representations:

- Dense SIFT (DSIFT) [35]: Firstly, we densely sampled SIFT descriptors on each 16×16 patch without overlapping and obtained 45 SIFT descriptors. Then, we concatenated these SIFT descriptors to form a 5760-dimensional feature vector;
- LBP [23]: We divided each image into 8×15 non-overlapping blocks, where the size of each block is 10×10. Then, we extracted a 59-dimensional uniform pattern LBP feature for each block and concatenated them to form a 7080-dimensional feature vector;
- Sparse SIFT (SSIFT): We used SSIFT feature provided by [1], which first localized nine fixed landmarks in each image and extracted SIFT descriptors over three scales at these landmarks, and then concatenated these 27 SIFT descriptors to result a 3456-dimensional vector.

For these three kinds of features, we employed whitened PCA (WPCA) to project each feature vector into a 200-dimensional feature subspace, respectively. Note that we first employed the WPCA on the training set to compute the projection matrix, and then we used this projection matrix to reduce the dimension of each sample in the training set and testing set.

2) Comparison with Baseline Methods: Table I records the verification accuracy with standard error of our methods and baseline methods by different metric learning strate-

TABLE I
COMPARISONS OF THE MEAN VERIFICATION ACCURACY (%) WITH SEVERAL BASELINE METHODS ON THE LFW UNDER CATEGORY OF IMAGE-RESTRICTED, LABEL-FREE OUTSIDE DATA.

Method	Feature	Accuracy (%)
SML	DSIFT	84.30 ± 0.69
SML	LBP	83.83 ± 0.41
SML	SSIFT	84.58 ± 0.36
CML	All	82.40 ± 0.51
IML	All	87.78 ± 0.58
LML	DSIFT	86.33 ± 0.66
LML	LBP	85.98 ± 0.44
LML	SSIFT	86.75 ± 0.34
LM^{3}L	All	89.57 ± 0.48
L^{2}M^{3}L	All	$\mathbf{90.23 \pm 0.55}$

gies on the LFW dataset under category of image-restricted, label-free outside data. We see that the LM^{3}L and L^{2}M^{3}L methods consistently outperforms these baseline methods in terms of the mean verification accuracy. Compared with SML (or LML), the LM^{3}L (or L^{2}M^{3}L) learns multiple distance metrics with multi-feature representations, such that more discriminative information can be exploited for verification. Compared with CML and IML, our LM^{3}L and L^{2}M^{3}L jointly learn multiple distance metrics so that the distance metrics learned for different features can interact each other, therefore more complementary information can be extracted for face verification. We also observe that local metric leaning methods (i.e., LML and L^{2}M^{3}L) obtain the better performance than their global companions (i.e., SML and LM^{3}L). These results show that the LML and L^{2}M^{3}L can exploit local specificities of data to improve performance of face verification.

3) Comparison with State-of-the-Art Methods: We also compared our LM^{3}L and L^{2}M^{3}L methods with several state-of-the-art methods on the LFW dataset. These methods can be categorized into metric learning based methods containing PCCA [27], DML-eig combined [36], CSML+SVM [7], SFRD+PMML [6], Sub-SML [19], large margin local metric learning (LMLML) [31], and discriminative deep metric learning (DDML) [37]; and descriptor based methods including pose adaptive filter (PAF) [38], high dimensional vector multiplication (VMRS) [39], Hybrid on LFW3D [40], and Spartans [41]. Table II tabulates the mean verification accuracy with standard error of different methods and Fig. 1 shows ROC curves of several state-of-the-art methods on this dataset. We see that the proposed LM^{3}L and L^{2}M^{3}L methods achieve competitive results compared with these state-of-the-art methods except two methods: Sub-SML + Hybrid on LFW3D [40] and HPEN + HD-LBP + DDML [11]. The reason is that they both employed the powerful face alignment techniques, and Sub-SML + Hybrid on LFW3D [40] adopted more than 10 types of features and HPEN + HD-LBP + DDML [11] exploited the over-complete high-dimensional feature (i.e., 100K-dim HD-LBP) for face verification.

The reason that VMRS [39] outperforms L^{2}M^{3}L is that: 1) VMRS adopted 10 different features (i.e., LBP, TPLBP, OCLBP, SIFT, Scattering, and their "sqrt root" versions); 2) VMRS used high-dimensional features, e.g., the 40887-dimensional OCLBP, and the 96520-dimensional Scattering feature; and 3) VMRS combined the non-linear dimensionality reduction technique called Diffusion Maps (DM) [39] and WPCA to obtain an additional improvement in accuracy. In our methods, we used three low-dimensional features, where the WPCA is used to reduce the dimensionality. The reason that the performance of DDML [37] outperforms L^{2}M^{3}L is that 1) The DDML employs 6 different features; and 2) The DDML adopts a nonlinear distance metric learning method via the neural network to exploit the nonlinearity of data points. In our methods, we use three low-dimensional features to learn several linear distance metrics.

TABLE II
COMPARISONS OF THE MEAN VERIFICATION ACCURACY (%) WITH STATE-OF-THE-ART RESULTS ON THE LFW UNDER CATEGORY OF IMAGE-RESTRICTED, LABEL-FREE OUTSIDE DATA, WHERE NOF DENOTES THE NUMBER OF FEATURE USED IN EACH METHOD.

Method	NoF	Accuracy (%)
PCCA [27]	1	83.80 ± 0.40
Hybrid on LFW3D [40]	12	85.63 ± 0.53
DML-eig combined [36]	8	85.65 ± 0.56
LMLML [31]	1	86.13 ± 0.53
PAF [38]	1	87.77 ± 0.51
CSML+SVM [7]	6	88.00 ± 0.37
SFRD+PMML [6]	8	89.35 ± 0.50
Spartans [41]	1	89.69 ± 0.36
Sub-SML [19]	6	89.73 ± 0.38
TSML with feature fusion [42]	12	89.80 ± 0.47
DDML [37]	6	90.68 ± 1.41
VMRS [39]	10	91.10 ± 0.59
Sub-SML + Hybrid on LFW3D [40]	12	91.65 ± 1.04
HPEN + HD-LBP + DDML [11]	1	92.57 ± 0.36
LM^{3}L	3	89.57 ± 0.48
L^{2}M^{3}L	3	$\mathbf{90.23 \pm 0.55}$

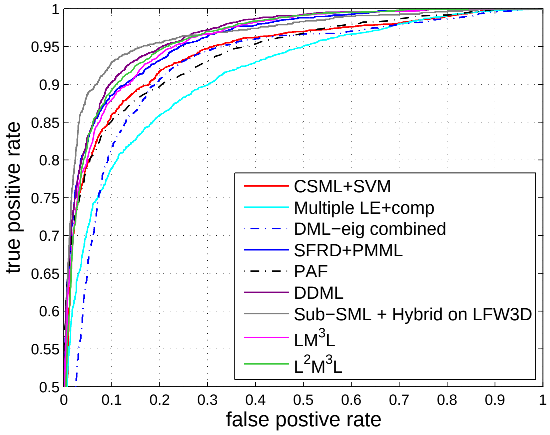


Fig. 1. ROC curves of our methods and several state-of-the-art methods on the LFW under category of image-restricted, label-free outside data.

4) Comparison of LM^{3}L and L^{2}M^{3}L with $\lambda = 0$: When $\lambda = 0$, we evaluated the LM^{3}L ($\lambda = 0$) and L^{2}M^{3}L ($\lambda = 0$) on LFW dataset under category of image-restricted, label-free outside data (see Table III). Table III shows that the regularization term of the LM^{3}L and L^{2}M^{3}L can help improve the verification accuracy. The reason is that the LM^{3}L and L^{2}M^{3}L seek a commonality of multiple features and make all features more robust for face verification.

TABLE III
COMPARISON OF LM^3L AND $\text{L}^2\text{M}^3\text{L}$ WHEN $\lambda = 0$ ON LFW DATASET UNDER CATEGORY OF IMAGE-RESTRICTED, LABEL-FREE OUTSIDE DATA.

Method	Feature	Accuracy (%)
LM^3L ($\lambda = 0$)	LBP, DSIFT, SSIFT	88.02 ± 0.56
LM^3L		89.57 ± 0.48
$\text{L}^2\text{M}^3\text{L}$ ($\lambda = 0$)	LBP, DSIFT, SSIFT	88.95 ± 0.60
$\text{L}^2\text{M}^3\text{L}$		90.23 ± 0.55

TABLE IV
COMPARISON OF THE MEAN VERIFICATION ACCURACY (%) WITH BASELINE METHODS USING DIFFERENT METRIC LEARNING STRATEGIES ON THE YTF UNDER THE IMAGE RESTRICTED SETTING.

Method	Feature	Accuracy (%)
SML	CSLBP	73.66 ± 1.52
SML	FPLBP	75.02 ± 1.67
SML	LBP	78.46 ± 0.94
CML	All	75.36 ± 2.37
IML	All	80.12 ± 1.33
LML	CSLBP	75.76 ± 1.59
LML	FPLBP	75.78 ± 2.19
LML	LBP	80.08 ± 2.06
LM^3L	All	81.28 ± 1.17
$\text{L}^2\text{M}^3\text{L}$	All	$\mathbf{81.72 \pm 1.53}$

B. *Video-based Face Verification on YTF*

1) Dataset and Settings: The YTF dataset [10] consists of 3425 videos of 1595 different people collected from YouTube site. There are also large variations in pose, illumination, and expression in each video, and the average length of each video clip is 181.3 frames. In our experiments, we followed the standard evaluation protocol and evaluated our methods for unconstrained video-based face verification on the 5000 video pairs. These 5000 pairs are equally divided into 10 folds and each fold contains 250 intra-personal pairs (positive pairs) and 250 inter-personal pairs (negative pairs). We adopted the *image restricted* and *image unrestricted* protocols to evaluate the proposed methods. For the image restricted setting, we directly used three feature descriptors including LBP, Center-Symmetric LBP (CSLBP) [10] and Four-Patch LBP (FPLBP) [43] which are provided by [10]. Since all face images have been aligned by the detected facial key points, we simply averaged all the feature vectors within one video clip to result a mean feature vector for each type of feature. Then, we employed WPCA to reduce each feature into a 200-dimensional feature vector.

2) Comparison with Baseline Methods: As in LFW dataset, we also compared our methods with several baseline methods using different metric learning strategies, i.e., SML, CML and IML on the YTF dataset under the image restricted setting. Table IV records the mean verification accuracy with standard error of these metric learning methods on the YTF under the image restricted setting. We see that the proposed LM^3L and $\text{L}^2\text{M}^3\text{L}$ methods consistently perform better than these baseline methods in terms of the mean verification accuracy; and local based methods, LML and $\text{L}^2\text{M}^3\text{L}$, can make use of local structures of samples to enhance the verification accuracy. These two observations are in agreement with results obtained on the LFW dataset.

TABLE V
COMPARISONS OF THE MEAN VERIFICATION ACCURACY WITH STANDARD ERROR (%) WITH SEVERAL STATE-OF-THE-ART RESULTS ON THE YTF UNDER THE IMAGE RESTRICTED SETTING.

Method	Accuracy (%)
MBGS (LBP) [10]	76.40 ± 1.80
APEM (LBP) [24]	77.44 ± 1.46
APEM (fusion) [24]	79.06 ± 1.51
STFRD+PMML [6]	79.48 ± 2.52
MBGS+SVM⊖ [44]	78.90 ± 1.90
VSOF+OSS (Adaboost) [45]	79.70 ± 1.80
DDML (combined) [37]	82.34 ± 1.47
LMKMML [48]	82.70 ± 1.50
DMM+CFN [47]	82.80 ± 0.90
Eigen-PEP [46]	84.80 ± 1.40
LM^3L	81.28 ± 1.17
$\text{L}^2\text{M}^3\text{L}$	$\mathbf{81.72 \pm 1.53}$

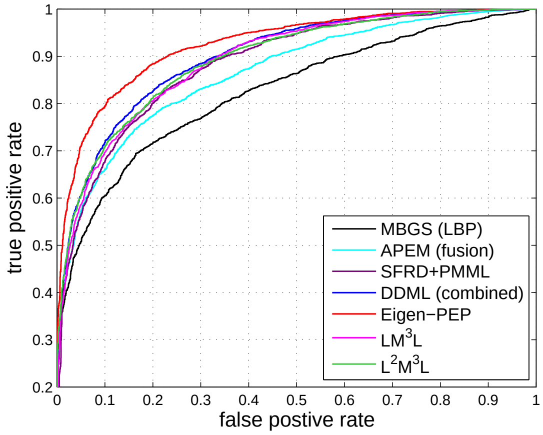


Fig. 2. ROC curves of our methods and several state-of-the-art methods on the YTF under the image restricted setting.

3) Comparison with State-of-the-Art Methods: We then compared LM^3L and $\text{L}^2\text{M}^3\text{L}$ with state-of-the-art methods on the YTF dataset under the image restricted setting. The compared methods include matched background similarity (MBGS) [10], APEM [24], STFRD+PMML [6], MBGS+SVM⊖ [44], VSOF+OSS (Adaboost) [45], DDML [37], Eigen-PEP [46], deep mixture model and convolutional fusion network (DMM+CFN) [47], and LMKMML [48]. Table V lists the mean verification accuracy with the standard error, and Fig. 2 shows ROC curves of our methods and several state-of-the-art methods on the YTF dataset, respectively. We can observe that our $\text{L}^2\text{M}^3\text{L}$ method achieves competitive results compared with most of these state-of-the-art methods on this dataset under the image restricted setting. Additionally, the Eigen-PEP obtains the best accuracy, the reason is that it exploits intra-class variations between frames of each video clip while our implementation simply takes the mean of all frames for a video in feature representation.

4) Comparison with Deep Learning based Methods: We also evaluated our LML, LM^3L and $\text{L}^2\text{M}^3\text{L}$ methods using convolutional neural network (CNN) feature which recently has achieved various promising results on face verification [2], [12], [13]. In our implementation, we employed the VGG-Face CNN model provided by [13] to compute CNN descriptor.

TABLE VI
COMPARISON WITH DEEP LEARNING BASED METHODS ON THE YTF DATASET UNDER IMAGE UNRESTRICTED SETTING.

Method	Feature	Accuracy (%)
SML	CNN	93.92 ± 1.18
LML	CNN	94.56 ± 1.24
LM^{3}L	CNN, CSLBP, FPLBP, LBP	94.75 ± 1.21
L^{2}M^{3}L	CNN, CSLBP, FPLBP, LBP	$\mathbf{94.90 \pm 1.09}$
DFD-SID+JB [49]	CNN	89.10 ± 0.40
DeepFace-single [2]	CNN	91.4 ± 1.1
FaceNet [12]	CNN	95.12 ± 0.39
Softmax (L2) [13]	CNN	91.6
Embedding loss [13]	CNN	97.3

Specifically, we only extracted CNN feature on the first 100 frames of each video at single scale. For each face image in YTF dataset, we first resized it to size of 200×200 pixels and cropped 100×100 region from its center, then we resized it to 224×224 pixel image to compute 4096-dimensional CNN feature vector. Finally, we averaged these CNN feature vectors of the first 100 frames for each video, and each video was represented by a 4096-dimensional vector. Moreover, each feature vector was reduced to the size of 200 by PCA. Table VI shows the mean verification accuracy of our proposed methods and several deep learning based methods using CNN feature (e.g., DeepFace [2], FaceNet [12], and VGG-Face CNN [13]) on the YTF under the image unrestricted setting. We see that our LM^{3}L and L^{2}M^{3}L can be comparable to these state-of-the-art results on this dataset under this unrestricted setting.

The Face Net [12] and Embedding loss [13] are two current state-of-the-art methods on YTF dataset. The reason that they outperformed L^{2}M^{3}L is that:

- The Face Net [12] used about 200 million face images in the model training, and the VGG-Face CNN model used in our method used about 2.6 million face images.
- To extract CNN features for each face video, the Embedding loss [13] selected the top 100 frames of this video by ordering the faces by their facial landmark confidence score. In our L^{2}M^{3}L, we only took the first 100 frames by following the setting in the Face Net [12].
- The Face Net [12] and Embedding loss [13] are two strongly-supervised learning methods, because they employ the triplet loss function which exploits the label information of each face video of YTF dataset. Unlike these methods, our L^{2}M^{3}L method is a weakly-supervised learning method which only exploits the pair-wise supervision from face video pairs.

In Table VI, the main aim of L^{2}M^{3}L and LML is to show that 1) our metric learning methods learn the favorable distance metrics to improve the performance of Softmax (L2) [12] (91.6%) when the CNN feature is used; and 2) L^{2}M^{3}L can further improve the verification accuracy by integrating low-level features and high-level features into a unified framework.

C. Kinship Verification on KinFaceW-II

TABLE VII
COMPARISONS OF THE MEAN VERIFICATION ACCURACY (%) WITH BASELINE METHODS USING DIFFERENT METRIC LEARNING STRATEGIES ON THE KINFACEW-II DATASET.

Method	Feature	F-S	F-D	M-S	M-D	Mean
SML	LE	76.2	70.1	72.4	71.8	72.6
SML	LBP	66.9	65.5	63.1	68.3	66.0
SML	TPLBP	71.8	63.3	63.0	67.6	66.4
SML	SIFT	68.1	63.8	67.0	63.9	65.7
CML	All	76.3	67.5	74.3	75.4	73.4
IML	All	79.4	71.5	76.3	77.3	76.1
LML	LE	76.8	74.2	76.6	73.8	75.4
LML	LBP	66.0	64.8	67.8	66.8	66.4
LML	TPLBP	68.6	66.2	65.4	70.8	67.8
LML	SIFT	72.2	66.0	68.2	66.2	68.2
LM^{3}L	All	**82.4**	74.2	**79.6**	78.7	78.7
L^{2}M^{3}L	All	**82.4**	**78.2**	78.8	**80.4**	**80.0**

1) Dataset and Settings: The KinFaceW-II [5] is a kinship face dataset collected from the public figures or celebrities and their parents or children. There are four kinship relations in the KinFaceW-II datasets: Father-Son (F-S), Father-Daughter (F-D), Mother-Son (M-S) and Mother-Daughter (M-D), and each relation includes 250 pairs of kinship images. Following the experimental settings in [5], we constructed 250 positive pairs (with kinship) and 250 negative pairs (without kinship) for each relation. For each face image, we extracted four types of feature representations as follows:

- LEarning-based descriptor (LE) [50]: Following the same parameter settings used in [50], [5], we first obtained 200 cluster centers by k-means clustering, and then performed vector quantization to obtain a 200-bin histogram feature for the whole face image;
- LBP: A 256-bin histogram feature was extracted;
- TPLBP [43]: We obtained a 256-bin histogram feature for each image by adopting the default setting in [43].
- SIFT: We densely sampled SIFT descriptors on 16×16 blocks with space of 8 pixels, and then computed a 200-bin histogram feature for each image by adopting the bag-of-visual-words model.

We adopted the 5-fold cross validation strategy for each of the four relations in this dataset and the final results were reported by the mean verification accuracy.

2) Comparison with Baseline Methods: We first compared our method with SML, CML, and IML on the KinFaceW-II dataset. Table VII records the mean verification accuracy of our methods and baseline methods using various metric learning strategies on the KinFaceW-II dataset for four kinship relations, respectively. We also see that the LML, LM^{3}L and L^{2}M^{3}L consistently outperforms baseline metric learning strategies on four relations in mean verification accuracy.

3) Comparison with Multiple Metric Learning Methods: We further compared the LM^{3}L and L^{2}M^{3}L methods with several multiple metric learning methods for kinship verification. These multiple metric learning methods include multiple canonical correspondence analysis (MCCA) [33], multiple NRML (MNRML) [5], discriminative multimetric learning (DMML) [20], and PMML [6]. Table VIII reports the mean verification accuracy of our methods and these multiple metric learning methods. We observe that L^{2}M^{3}L achieves about 1.7% improvement over DMML and 1.3% over LM^{3}L in terms of the mean verification accuracy.

TABLE VIII
COMPARISONS OF THE MEAN VERIFICATION ACCURACY (%) WITH SEVERAL MULTIPLE METRIC LEARNING METHODS ON THE KINFACEW-II DATASET.

Method	Feature	F-S	F-D	M-S	M-D	Mean
MCCA [33]	All	74.0	72.1	74.8	75.3	74.1
PMML [6]	All	77.7	72.4	76.3	74.8	75.3
MNRML [17]	All	76.9	74.3	77.4	77.6	76.5
DMML [20]	All	78.5	76.5	78.5	79.5	78.3
LM^{3}L	All	**82.4**	74.2	**79.6**	78.7	78.7
L^{2}M^{3}L	All	**82.4**	**78.2**	78.8	**80.4**	**80.0**

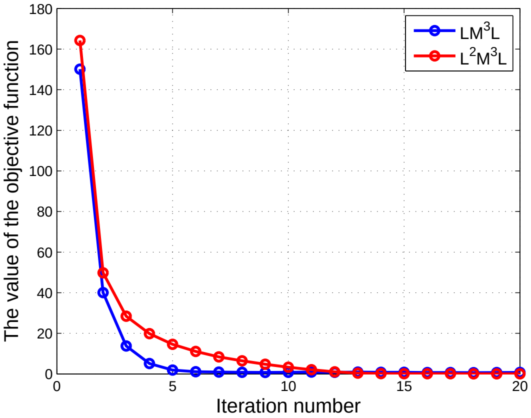


Fig. 3. The value of the objective function of LM^{3}L and L^{2}M^{3}L versus different number of iterations on the LFW dataset.

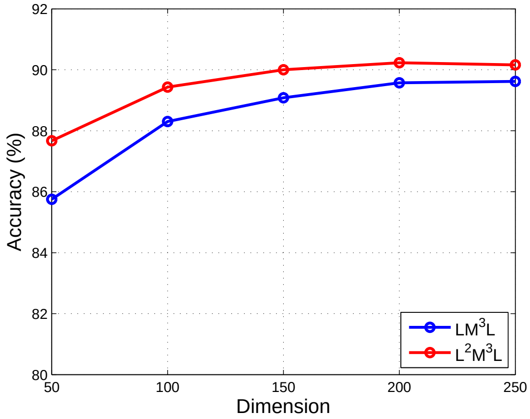


Fig. 4. The mean verification accuracy of LM^{3}L and L^{2}M^{3}L versus different feature dimensions on the LFW dataset.

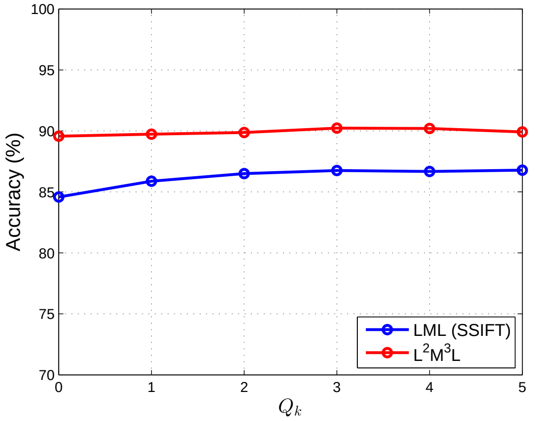


Fig. 5. The mean verification accuracy of LM^{3}L and L^{2}M^{3}L versus various Q_k (i.e., number of local metrics) on the LFW dataset.

D. Discussion and Parameter Analysis

We examined several parameters that may affect the performance of the LM^{3}L and L^{2}M^{3}L methods on the LFW dataset under category of image-restricted, label-free outside data.

1) Convergence Analysis: We first evaluated the convergence of the LM^{3}L and L^{2}M^{3}L methods with different number of iterations. Fig. 3 shows the value of the objective function of the LM^{3}L and L^{2}M^{3}L versus different number of iterations on the LFW dataset. We see that the convergence speed of our methods is acceptable. The LM^{3}L converges in $5 \sim 6$ iterations and the L^{2}M^{3}L method begins to keep stable after the 10 iterations on the training set.

2) Effect of Different Feature Dimensions: Then we investigated the performance of the LM^{3}L and large-margin versus different feature dimensions. Fig. 4 shows the mean verification accuracy of our multi-metric learning methods versus different feature dimensions on the LFW dataset. We notice that our methods can achieve stable performance when the feature dimension of various features reaches 200. This is the reason that we select 200 dimension for each feature via WPCA in our experiments on this dataset.

3) Effect of Different Number of Local Metrics: Lastly, we evaluated how the various number of local metrics (i.e., Q_k) affects the LML and L^{2}M^{3}L methods. For the LML method, we chose SSIFT feature for this evaluation due to its good performance. Fig. 5 lists the mean verification accuracy versus various Q_k on the LFW dataset. We see that increasing Q_k improves the accuracy of LML and L^{2}M^{3}L, but the performance of our local metric learning based methods remains stable or may even degrade if a too large number of local metrics is adopted. The reason may be that learning more local metrics requires a sufficient number of training samples in model training. In the experiments, we adopted three local metrics because it not only obtains acceptable results but also reduces the computational time.

VI. CONCLUSION

In this paper, we have introduced a large-margin multi-metric learning (LM^{3}L) method for face and kinship verification under unconstrained conditions. The LM^{3}L jointly learns multiple distance metrics under which more discriminative and complementary information can be exploited. Moreover, to better exploit the local structures of face images, we have proposed a local metric learning (LML) and a local large-margin multi-metric learning (L^{2}M^{3}L) methods to learn a set of local metrics. Experimental results on three datasets show that our method can achieve competitive results compared with the state-of-the-art methods. For future work, we are interested in applying our methods to other computer vision applications such as person re-identification, action recognition and object tracking to further show their effectiveness.

References

[1] M. Guillaumin, J. J. Verbeek, and C. Schmid, “Is that you? metric learning approaches for face identification,” in *IEEE International Conference on Computer Vision*, 2009, pp. 498–505.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] Z. Wang, Y. Hu, and L.-T. Chia, “Image-to-class distance metric learning for image classification,” in *ECCV*, 2010, pp. 706–719.

[4] D. Tran and A. Sorokin, “Human activity recognition with metric learning,” in *European Conference on Computer Vision*, 2008, pp. 548–561.

[5] J. Lu, J. Hu, X. Zhou, Y. Shang, Y. Tan, and G. Wang, “Neighborhood repulsed metric learning for kinship verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2594–2601.

[6] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, “Fusing robust face region descriptors via multiple metric learning for face recognition in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3554–3561.

[7] H. V. Nguyen and L. Bai, “Cosine similarity metric learning for face verification,” in *Asian Conference on Computer Vision*, 2010, pp. 709–720.

[8] J. Hu, J. Lu, J. Yuan, and Y. Tan, “Large margin multi-metric learning for face and kinship verification in the wild,” in *Asian Conference on Computer Vision*, 2014, pp. 252–267.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[10] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[11] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conference*, 2015, pp. 41.1–41.12.

[14] R. Fang, K. Tang, N. Snavely, and T. Chen, “Towards computational models of kinship verification,” in *International Conference on Image Processing*, 2010, pp. 1577–1580.

[15] X. Zhou, J. Hu, J. Lu, Y. Shang, and Y. Guan, “Kinship verification from facial images under uncontrolled conditions,” in *ACM Conference on Multimedia*, 2011, pp. 953–956.

[16] S. Xia, M. Shao, J. Luo, and Y. Fu, “Understanding kin relationships in a photo,” *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1046–1056, 2012.

[17] J. Lu, X. Zhou, Y. Tan, Y. Shang, and J. Zhou, “Neighborhood repulsed metric learning for kinship verification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 331–345, 2014.

[18] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah, “Who do I look like? determining parent-offspring resemblance via gated autoencoders,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1757–1764.

[19] Q. Cao, Y. Ying, and P. Li, “Similarity metric learning for face recognition,” in *IEEE International Conference on Computer Vision*, 2013, pp. 2408–2415.

[20] H. Yan, J. Lu, W. Deng, and X. Zhou, “Discriminative multimetric learning for kinship verification,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1169–1178, 2014.

[21] S. Chakraborty, S. K. Singh, and P. Chakraborty, “Local directional gradient pattern: a local descriptor for face recognition,” *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 1201–1216, 2017.

[22] S. Chakraborty, S. Singh, and P. Chakraborty, “Local gradient hexa pattern: A descriptor for face recognition and retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[23] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: Application to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[24] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, “Probabilistic elastic matching for pose variant face verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3499–3506.

[25] K. Q. Weinberger, J. Blitzer, and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *Advances in Neural Information Processing Systems*, 2005, pp. 1473–1480.

[26] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *International Conference on Machine Learning*, 2007, pp. 209–216.

[27] A. Mignon and F. Jurie, “Pcca: A new approach for distance learning from sparse pairwise constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

[28] J. Hu, J. Lu, and Y.-P. Tan, “Deep metric learning for visual tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2056–2068, 2016.

[29] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, “Deep transfer metric learning,” *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5576–5588, 2016.

[30] X. You, Q. Li, D. Tao, W. Ou, and M. Gong, “Local metric learning for exemplar-based object detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1265–1276, 2014.

[31] J. Bohné, Y. Ying, S. Gentric, and M. Pontil, “Large margin local metric learning,” in *European Conference on Computer Vision*, 2014, pp. 679–694.

[32] Y. Fu, L. Cao, G. Guo, and T. S. Huang, “Multiple feature fusion by subspace learning,” in *ACM International Conference on Image and Video Retrieval*, 2008, pp. 127–134.

[33] A. Sharma, A. Kumar, H. Daume III, and D. Jacobs, “Generalized multiview analysis: a discriminative latent space,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1867–1875.

[34] M. Gönen and E. Alpaydin, “Localized multiple kernel learning,” in *International Conference on Machine Learning*, 2008, pp. 352–359.

[35] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[36] Y. Ying and P. Li, “Distance metric learning with eigenvalue optimization,” *Journal of Machine Learning Research*, vol. 13, pp. 1–26, 2012.

[37] J. Hu, J. Lu, and Y. Tan, “Discriminative deep metric learning for face verification in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1875–1882.

[38] D. Yi, Z. Lei, and S. Z. Li, “Towards pose robust face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3539–3545.

[39] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, “Fast high dimensional vector multiplication face recognition,” in *IEEE International Conference on Computer Vision*, 2013, pp. 1960–1967.

[40] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[41] F. Juefei-Xu, K. Luu, and M. Savvides, “Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios,” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4780–4795, 2015.

[42] L. Zheng, K. Idrissi, C. Garcia, S. Duffner, and A. Baskurt, “Triangular similarity metric learning for face verification,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–7.

[43] L. Wolf, T. Hassner, and Y. Taigman, “Descriptor based methods in the wild,” in *Real-Life Images workshop at the European Conference on Computer Vision*, 2008.

[44] L. Wolf and N. Levy, “The svm-minus similarity score for video face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3523–3530.

[45] H. Mendez-Vazquez, Y. Martinez-Diaz, and Z. Chai, “Volume structured ordinal features with background similarity measure for video face recognition,” in *International Conference on Biometrics*, 2013, pp. 1–6.

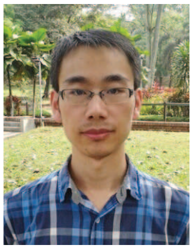
[46] H. Li, G. Hua, X. Shen, Z. L. Lin, and J. Brandt, “Eigen-pep for video face recognition,” in *Asian Conference on Computer Vision*, 2014, pp. 17–33.

[47] C. Xiong, L. Liu, X. Zhao, S. Yan, and T.-K. Kim, “Convolutional fusion network for face verification in the wild,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 517–528, 2016.

[48] J. Lu, G. Wang, and P. Moulin, “Localized multifeature metric learning for image-set-based face recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 529–540, 2016.

[49] Z. Lei, D. Yi, and S. Z. Li, “Learning stacked image descriptor for face recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1685–1696, 2016.

[50] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2707–2714.



Junlin Hu received the B.Eng. degree from the Xi'an University of Technology, Xi'an, China, in 2008, and M.Eng. degree from Beijing Normal University, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision, pattern recognition, and biometrics.

Jiwen Lu (S'10-M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from the Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From March 2011 to November 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 150 scientific papers in these areas, where 41 were the IEEE Transactions papers. He serves/has served as an Associate Editor of Pattern Recognition Letters, Neurocomputing, and the IEEE Access, a Guest Editor of five journals such as Pattern Recognition, Computer Vision and Image Understanding, and Image and Vision Computing, and an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is/was a Workshop Chair/Special Session Chair/Area Chair for more than 10 international conferences. He was a recipient of the National 1000 Young Talents Plan Program in 2015.

Yap-Peng Tan (S'95-M'97-SM'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1995 and 1997, respectively, all in electrical engineering. From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, and Sharp Laboratories of America, Camas, WA. In November 1999, he joined the Nanyang Technological University of Singapore, where he is currently Associate Professor and Associate Chair (Academic) of the School of Electrical and Electronic Engineering. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, and data analytics. He served as the Chair of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society from 2012 to 2014, and Chairman of the IEEE Signal Processing Singapore Chapter from 2009 to 2010. He has also served as Associate Editor of the IEEE Signal Processing Letters, IEEE Transactions on Multimedia, and the IEEE Access, an Editorial Board Member of the EURASIP Journal on Advances in Signal Processing and EURASIP Journal on Image and Video Processing, Guest Editor for special issues of several journals including the IEEE Transactions on Multimedia. He is the Tutorial Co-Chair of ICME'2016 and Technical Program Co-Chair of ICIP'2019, and was the Finance Chair of ICIP'2004, General Co-Chair of ICME'2010, Technical Program Co-Chair of ICME'2015, and General Co-Chair of VCIP'2015.

Junsong Yuan (M'08-SM'14) received his Ph.D. from Northwestern University in 2009 and M.Eng. from National University of Singapore in 2005, both in Electrical and Computer Engineering. Before that, he graduated from the Special Class for the Gifted Young of Huazhong University of Science and Technology (HUST), Wuhan, China, in 2002. He is currently an Associate Professor at School of Electrical and Electronics Engineering (EEE), Nanyang Technological University (NTU). His research interests include computer vision, video analytics, gesture and action analysis, large-scale visual search and mining. He received 2016 Best Paper Award from IEEE Trans. on Multimedia, Doctoral Spotlight Award from IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'09), Nanyang Assistant Professorship from NTU, Outstanding EECS Ph.D. Thesis award from Northwestern University, and National Outstanding Student from Ministry of Education, P.R.China. He is currently an Associate Editor of IEEE Trans. on Image Processing (T-IP), IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT) and The Visual Computer journal (TVC), and served as Guest Editor of International Journal of Computer Vision (IJCV). He is Area Chair of CVPR'17, ICIP'17, ICPR'16, ICME'15'14, ACCV'14, and WACV'14. He is also in the organizing committee of CVPR'17, ICME'16'18, VCIP'15, and ACCV'14.

Jie Zhou (M'01-SM'04) received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as PAMI, TIP and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the International Journal of Robotics and Automation and two other journals.