Large Margin Multi-Modal Triplet Metric Learning

Xing Di and Vishal M. Patel
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
94 Brett Road, Piscataway, NJ 08854
xd55@scarletmail.rutgers.edu, vishal.m.patel@rutgers.edu

***Abstract*— Distance metric learning is a significant technique that can improve the similarity accuracy in verification systems. In this paper, we propose a multi-metric learning algorithm with the triplet distance constraints for multi-modal verification problems. The main feature of our algorithm is that when learning multi-metric, we not only enforce the distance between the anchor and the positive samples to be less than the distance between the anchor and the negative samples but we also make the distance between the anchor and the positive samples as small as possible. A simple iterative procedure is introduced to solve the proposed optimization problem. Extensive experiments on three publicly available multi-modal datasets show that our method can perform significantly better than many state-of-the-art multi-modal metric learning methods.**

I. Introduction

Due to recent advances in sensing, communication and storage technologies, we have seen an explosion in the availability of visual data from multiple sources and modalities in recent years. Millions of cameras have been installed in buildings, streets, and airports around the world. Furthermore, people are using billions of handheld devices that are capable of capturing multi-modal information such as light, heat and depth. Multi-modal data is drastically increasing with the significant use of social media. For instance, in many social networking sites, images and videos are often described by user comments, image contents, audio, tags and meta data information such as albums and groups. These information from different sources and modalities such as text, audio and image frames, can be used to develop better detection, classification, and retrieval algorithms.

One of the biggest challenges in designing classification or retrieval algorithms for unimodal as well as multi-modal data is to choose a proper similarity or distance measure function. Various metric learning algorithms have been developed in the literature for unimodal data [23], [21], [8], [1], [3], [10], [20] [6]. The main idea is to learn an optimal metric which minimizes the distance between similar data and simultaneously maximizes the distance between dissimilar data [14], [24]. Some of these algorithms learn the metric from data pairs and side information indicating the relationship of the data pairs [23]. The constraints used by these methods are either pairwise constraints or triplet constraints. In the pairwise constraint [23], similar set S and dissimilar set D are used to train a metric. In other words, a pair of data $(\mathbf{x}_i, \mathbf{x}_j) \in S$ if $\mathbf{x}_i$ is similar to $\mathbf{x}_j$ otherwise $(\mathbf{x}_i, \mathbf{x}_j) \in D$. In the case of triplet constraint [21], a triplet set $T = (\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)$ is given, where $(\mathbf{x}_i^a, \mathbf{x}_i^p) \in S$ and $(\mathbf{x}_i^a, \mathbf{x}_i^n) \in D$ for $i = 1, \ldots, L$. In other words, the anchor sample, $\mathbf{x}_i^a$, is similar to the positive sample, $\mathbf{x}_i^p$, and the anchor sample, $\mathbf{x}_i^a$, is dissimilar to the negative sample, $\mathbf{x}_i^n$. Here, L denotes the total number of triplets in T. Recent metric learning algorithms also explore the structure of the metric by enforcing sparse and/or low-rank constrains [7], [17], [25], [19], [16], [15]. Some of the other unimodal metric learning algorithms include joint Bayesian metric leaning [2],[3] and Discriminative Deep Metric Learning (DDML) [10].

One way to extend these unimodal metric learning algorithms for multi-modal data is to simply concatenate the feature vectors from different modalities into a long vector and then feed them directly to one of the unimodal metric learning algorithms. However, this simple method of learning multi-modal metric suffers from the following two limitations [22]: 1) Some features often dominate in the final concatenated feature. As a result, they are biased and can weaken the potential of all the other features. 2) Since the dimensionality of the resulting concatenated feature vector can be very large, it makes the overall metric learning algorithm computationally very expensive. In fact, our experiments indicate that in some cases when the metric is learned by simply concatenating multi-modal features, we often get a metric whose performance is much worse than when a single feature is used for metric learning.

In [27], a heterogeneous multi-metric learning algorithm was proposed which essentially extends the Large Margin Nearest Neighbor (LMNN) algorithm [21] for multi-metric leaning. Similarly, in [11] a Large Margin Multi-Metric Learning (LM3L) was proposed for face and kinship verification which learns multiple distance metrics under which the correlations of different feature representations of each sample are maximized. Some of the other multi-modal metric learning algorithms include Pairwise-constrained Multiple Metric Learning (PMML) [5]. More recently, a triplet constraint-based online multi-modal distance metric learning algorithm was proposed for image retrieval in [22].

Inspired by the LMNN formulation in [21], in this paper, we propose a novel multi-modal metric learning algorithm for the multi-modal verification problems using the triplet

This work was supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

978-1-5090-4023-0/17/$31.00 ©2017 IEEE

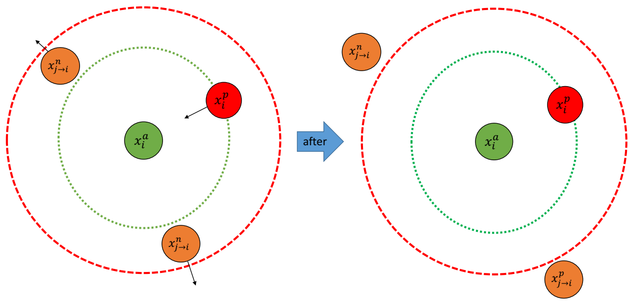

Fig. 1: An overview of the proposed LMMTML method. The dotted green circles refer to the distance between the anchor $\mathbf{x}_i^a$ and the positive sample $\mathbf{x}_i^p$. The distance between the green circle and the red circle is the margin distance.

constraint. Note that triplet constraint-based multi-modal metric learning algorithms such as [5] [22] learn metric by enforcing the distance between the anchor and the positive samples to be less than the distance between the anchor and the negative samples. However, in some cases, the distance between the anchor and the positive pairs in different triplet pairs may have a large intra-class distance [4]. Thus, to learn a better metric, the loss function based on the conventional triplet constraint should not only keep its original feature, but also should make the distance between the anchor and the positive samples as small as possible (see Figure 1). This is essentially the main motivation behind the proposed Large Margin Multi-modal Triplet Metric Learning (LMMTML) algorithm.

This paper is organized as follows. In Section II, we give a brief background on distance metric learning. The proposed LMMTML algorithm is presented in Section III. Experimental results are presented in Section IV. Finally, Section V concludes the paper with a brief summary.

II. Background

In this section, we give a brief background on distance metric learning. A number of different metric leaning algorithms can be developed by using a linear projection $\mathbf{W}$ of data $\mathbf{x} \in \mathbb{R}^n$, as $\hat{\mathbf{x}} = \mathbf{W}\mathbf{x}$. The squared Euclidean distance between two samples $\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j$ in the transformed space can be calculated as

$$\begin{aligned} d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) &= (\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j)^\top (\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{W}^\top \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \end{aligned} \qquad (1)$$

where $\mathbf{M} = \mathbf{W}^\top\mathbf{W}$. As a result, a linear projection $\mathbf{W}$ results in a Mahalanobis like metric $\mathbf{M}$ in the original space. Hence, $d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = d_{Mahal}(\mathbf{x}_i, \mathbf{x}_j)$. Note that often $\mathbf{M}$ is required to be a positive semi-definite (PSD) matrix when leaning this metric. In the case when $\mathbf{W}$ is chosen to be the identity matrix $\mathbf{I}$, the above formulation reduces to squared Euclidean distance. When learning this metric, side information based on the pairwise constraints is often used. In other words, the metric should be leaned such that the distance between similar pairs $(\mathbf{x}_i, \mathbf{x}_j) \in S$ should be less than the distance between dissimilar pairs $(\mathbf{x}_i, \mathbf{x}_k) \in D$.

One can also lean the above metric by enforcing triplet constrains while training the metric. In this case, (1) can be rewritten in terms of triplet loss as

$$\begin{aligned} E &= \sum_{\mathbf{x}^a, \mathbf{x}^p, \mathbf{x}^n} \left[\|\mathbf{W}\mathbf{x}^a - \mathbf{W}\mathbf{x}^p\|_2^2 - \|\mathbf{W}\mathbf{x}^a - \mathbf{W}\mathbf{x}^n\|_2^2 + \alpha \right]_+ \\ &= \sum_{\mathbf{x}^a, \mathbf{x}^p, \mathbf{x}^n} \left[(\mathbf{x}^a - \mathbf{x}^p)^\top \mathbf{M} (\mathbf{x}^a - \mathbf{x}^p) - (\mathbf{x}^a - \mathbf{x}^n)^\top \mathbf{M} (\mathbf{x}^a - \mathbf{x}^n) + \alpha \right]_+, \end{aligned} \qquad (2)$$

where $[\cdot]_+ = \max(.,0)$ donates the hinge loss and α is a margin parameter. The first term of the above formulation captures the distance between the anchor and the positive samples while the second term captures the distance between the anchor and the negative samples. As before, the purpose of this loss function is to make the distance between similar samples smaller than the distance between dissimilar samples.

The distance metric learning method for LMNN classification is an algorithm to learn a Mahalanobis distance metric for kNN classification from labeled examples [21]. In their formulation, two kinds of for energies were considered - $\varepsilon_{pull}(\mathbf{W})$ and $\varepsilon_{push}(\mathbf{W})$. They are defined as follows

$$\begin{aligned} \varepsilon_{pull}(\mathbf{W}) &= \sum_{j->i} \|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)\|^2, \\ \varepsilon_{push}(\mathbf{W}) &= \sum_{i,j->i} \sum_{l} (1 - y_{il})[1 + \|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \\ &\quad - \|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_l)\|^2]_+, \end{aligned} \qquad (3)$$

where $\mathbf{W}$ is the linear projection, $j->i$ denotes the set of target neighbors of $\mathbf{x}_i$ (i.e. k nearest neighbors with the same label as $\mathbf{x}_i$) and $y_{il} \in \{0, 1\}$ is a binary number indicating whether $\mathbf{x}_i$ and $\mathbf{x}_l$ correspond to the same class or not. The overall energy function is defined as the linear combination of the two enrages as

$$\varepsilon(\mathbf{W}) = (1 - \mu)\varepsilon_{pull}(\mathbf{W}) + \mu\varepsilon_{push}(\mathbf{W}), \qquad (4)$$

where $\mu \in [0, 1]$ is the parameter balancing the two energies. Note that $\varepsilon_{push}(\mathbf{W})$ quantifies the energy between samples from different classes which gives large energy to the small distance kNN samples from a different class. On the other hand, $\varepsilon_{pull}(\mathbf{W})$ gives large energy to the large distances of the kNN samples belonging to the same class. In contrast to the other metric learning algorithms, LMNN is coupled with the kNN classifier and hence it is specifically designed for the classification problems. In order to solve the above problem efficiently, one can reformulate it into an SDP problem as follows

Minimize

$$
\begin{aligned}
&(1-\mu)\sum_{i,j->i}(\mathbf{x}_i-\mathbf{x}_j)^\top\mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)+\\
&\mu\sum_{i,j->i,l}(1-y_{il})\epsilon_{i,j,l}
\end{aligned}
\tag{5}
$$

s.t.

$$
\begin{aligned}
&(\mathbf{x}_i-\mathbf{x}_l)^\top\mathbf{M}(\mathbf{x}_i-\mathbf{x}_l)-(\mathbf{x}_i-\mathbf{x}_j)^\top\mathbf{M}\\
&(\mathbf{x}_i-\mathbf{x}_j)\geq 1-\epsilon_{i,j,l}\\
&\epsilon_{i,j,l}\geq 0, \qquad \mathbf{M}\succeq 0.
\end{aligned}
$$

A multi-modal extension of this algorithm for multi-modal sensor classification problems was recently proposed in [27].

III. Large Margin Multi-Modal Triplet Matrices Learning (LMMTML)

Inspired by the elegant framework of LMNN algorithm, we propose a multi-modal extension of this algorithm for multi-modal verification problems. In the biometrics context, a verification problem is the one in which given two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, we have to determine whether they belong to the same person or not. In the multi-modal verification problem, we essentially do the same but now based on multi-modal data. Multi-modal verification is very important in various biometrics and computer vision problems. Note that the verification problem is different than the classification (or the identification problem). In the classification problem we identify the class label of a test sample based on a classifier and the training samples. This is clearly not the case in verification. In verification, receiver operating curve (ROC) is normally used to evaluate the performance of a matching algorithm. Before we define our LMMTML problem, we briefly define the notation used in this paper.

A. Notation

We use bold upper case letters and bold lower case letters to denote matrices and vectors, respectively. The terms and operates used in this paper are defined as follows.

- s: the total number of modalities.
- N: the total number of similar (anchor and positive) pairs.
- n_k: the dimensionality of the kth modality feature, $k = 1,2,3,...,s$.
- S: a positive constraint set, where $(\mathbf{x}_i,\mathbf{x}_j)\in S$ if and only if $\mathbf{x}_i$ is similar to $\mathbf{x}_j$.
- D: a negative constraint set, where $(\mathbf{x}_i,\mathbf{x}_j)\in D$ if and only if $\mathbf{x}_i$ is dissimilar to $\mathbf{x}_j$.
- Triplets $T_{i,j} = (\mathbf{x}_i^a,\mathbf{x}_i^p,\mathbf{x}_{j->i}^n), j = 1,2,3,...,L, i = 1,2,3,...,N$, where L is the number of negatives corresponding to the ith anchor, $(\mathbf{x}_i^a,\mathbf{x}_i^p)\in S$, and $(\mathbf{x}_i^a,\mathbf{x}_{j->i}^n)\in D$.
- $\{\mathbf{x}_i^a\}^k\in\mathbf{R}^{n_k}$: the kth type of modality feature of the anchor image in the Triplets $T_{i,j}$.
- $\{\mathbf{x}_i^p\}^k\in\mathbf{R}^{n_k}$: the kth type of modality feature of the positive image in the Triplets $T_{i,j}$.
- $\{\mathbf{x}_{j->i}^n\}^k\in\mathbf{R}^{n_k}$: the kth type of modality feature in the jth negative image in Triplets $T_{i,j}$.
- $\mathbf{W}_k\in\mathbf{R}^{r_k\times n_k}$: the linear projection matrix.
- $\mathbf{M}_k$: the optimal Mahalanobis matrix we want to learn, such that $\mathbf{M}_k = \mathbf{W}_k^\top\mathbf{W}_k$.
- c_k: the parameter to balance the influence caused by the unequal scale of the kth modality.
- $d_k^2(\mathbf{p},\mathbf{q})$: the squared distance between two samples $\mathbf{p}$ and $\mathbf{q}$ from the kth modality.
- α: the margin for the triplet constraint.
- N_t: a set of triplets such that the jth negative in $T_{i,j}$: $(\mathbf{x}_i^a,\mathbf{x}_i^p,\mathbf{x}_{j->i}^n)$ triggers the triplet constraint at the tth iteration.

$$\{\mathbf{C}_i^{a,p}\}^k = (\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)^\top$$

$$\{\mathbf{C}_{j->i}^{a,n}\}^k = (\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_{j->i}^n\}^k)(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_{j->i}^n\}^k)^\top.$$

B. Problem Formulation

The basic idea of our algorithm is that we want to learn a metric that can pull the similar samples (anchor and positive) closer while pushing away the dissimilar pairs (anchor and negative) whose distance is apparently closer than the similar pairs (see Figure 1). The proposed LMMTML formulation is as follows

$$
\begin{aligned}
\{\mathbf{M}_k\} = \underset{\{\mathbf{M}_k\}}{\arg\min}&(1-\mu)\sum_{i=1}^{N}\sum_{k=1}^{s}c_k(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)^\top\mathbf{M}_k\\
&(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)+\mu\sum_{i=1}^{N}\sum_{j->i}^{L}[\sum_{k=1}^{s}c_k[(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)^\top\\
&\mathbf{M}_k(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_i^p\}^k)-(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_{j->i}^n\}^k)^\top\mathbf{M}_k\\
&(\{\mathbf{x}_i^a\}^k-\{\mathbf{x}_{j->i}^n\}^k)]+\alpha]_+.
\end{aligned}
\tag{6}
$$

The above formulation can be rewritten as

$$
\begin{aligned}
&\{\mathbf{M}_k\} = \underset{\{\mathbf{M}_k\}}{\arg\min}(1-\mu)\sum_{i=1}^{N}\sum_{k=1}^{s}c_k d_k^2(\mathbf{x}_i^a,\mathbf{x}_i^p)\\
&+\mu\sum_{i=1}^{N}\sum_{j->i}^{L}\left[\sum_{k=1}^{s}c_k[d_k^2(\mathbf{x}_i^a,\mathbf{x}_i^p)-d_k^2(\mathbf{x}_i^a,\mathbf{x}_{j->i}^n)]+\alpha\right]_+,
\end{aligned}
\tag{7}
$$

where μ is a parameter. The first term in (7) corresponds to the pull energy between the anchors and positive samples. The second term corresponds to the push energy which essentially is the energy of the triplet loss constraint. From the overall energy function, we see that the first term enforces the sum of the distance between anchor and positive samples to be as small as possible, while the second term makes the distance between anchors and their corresponding negatives larger than the distance between anchors and positives. The overall problem (7) is not convex. We formulate it into a

PSD problem by using a slack variable as follows

$$\begin{aligned} &\arg\min_{\mathbf{M}_k}(1-\mu)\sum_{i=1}^{N}\sum_{k=1}^{s} c_k d_k^2(\mathbf{x}_i^a,\mathbf{x}_i^p)+\mu\sum_{i=1}^{N}\sum_{j->i}^{L}\epsilon_{ij} \\ &\text{s.t.} \sum_{k=1}^{s} c_k d_k^2(\mathbf{x}_i^a,\mathbf{x}_i^p)+\alpha-\epsilon_{ij}\leq\sum_{k=1}^{s} c_k d_k^2(\mathbf{x}_i^a,\mathbf{x}_{j->i}^n), \\ &\epsilon_{ij}\geq 0, \qquad \mathbf{M}_k\succeq 0, \end{aligned} \tag{8}$$

where $\mathbf{M}\succeq 0$ denotes that $\mathbf{M}$ is a PSD matrix.

C. Optimization

The SPD optimization problem (8) can be solved iteratively by first taking the gradient descent of the metrics and then projecting the metrics onto the SPD cone. In what follows, we describe these individual steps in detail.

1) Gradient Direction Computation: Using the notations $\{\mathbf{C}_i^{a,p}\}^k$ and $\{\mathbf{C}_{j->i}^{a,n}\}^k$ into (8), we can rewrite it at the tth iteration as

$$\begin{aligned} &\arg\min_{\mathbf{M}_k} \quad (1-\mu)\sum_{i=1}^{N}\sum_{k=1}^{s} c_k tr(\mathbf{M}_k^t\{\mathbf{C}_i^{a,p}\}^k)+ \\ &\mu\sum_{i=1}^{N}\sum_{j->i}^{L}[\sum_{k=1}^{s} c_k[tr(\mathbf{M}_k^t\{\mathbf{C}_i^{a,p}\}^k)-tr(\mathbf{M}_k^t\{\mathbf{C}_{j->i}^{a,n}\}^k)]+\alpha]_+. \end{aligned} \tag{9}$$

The gradient of (9) with respect to $\mathbf{M}_k^t$ is

$$\mathbf{G}_k^t=(1-\mu)c_k\sum_{i=1}^{N}\{\mathbf{C}_i^{a,p}\}^k+\mu\, c_k\sum_{(i,j)\in N_t}(\{\mathbf{C}_i^{a,p}\}^k-\{\mathbf{C}_{j->i}^{a,n}\}^k). \tag{10}$$

In order to save the computational expense, we reformulate (10) as [21]

$$\begin{aligned} \mathbf{G}_k^{t+1}=\mathbf{G}_k^t &-\mu\, c_k\sum_{(i,j)\in(N_t-N_{t+1})}(\{\mathbf{C}_i^{a,p}\}^k-\{\mathbf{C}_{j->i}^{a,n}\}^k) \\ &+\mu\, c_k\sum_{(i,j)\in(N_{t+1}-N_t)}(\{\mathbf{C}_i^{a,p}\}^k-\{\mathbf{C}_{j->i}^{a,n}\}^k), \end{aligned} \tag{11}$$

where N_t-N_{t+1} donates the samples in N_t but not in N_{t+1} and similarly $N_{t+1}-N_t$ donates the samples in N_{t+1} but not in N_t.

2) PSD Projection: The minimization of (8) must ensure that the metric $\mathbf{M}_k$ is PSD. This can be achieved by projecting the current estimate onto the cone of all positive semidefinite matrices. We perform the eigen decomposition of the current estimate $\mathbf{M}_k^t$ as

$$\mathbf{M}_k^t=\mathbf{V}\mathbf{\Delta}\mathbf{V}^\top, \tag{12}$$

where $\mathbf{V}$ consists of the eigenvectors of $\mathbf{M}_k^t$ and $\mathbf{\Delta}$ is the diagonal matrix with the corresponding eigen values. Then, $\mathbf{M}_k^t$ can be projected onto the PSD cone as

$$\text{PSD}(\mathbf{M}_k^t)=\mathbf{V}\mathbf{\Delta}^+\mathbf{V}^\top, \tag{13}$$

where $\mathbf{\Delta}^+=\max(0,\mathbf{\Delta})$. The overall LMMTML learning procedure is described in Algorithm 1.

Algorithm 1 Large Margin Multi-Modal Triplet Matrices Learning (LMMTML) Algorithm

```
Input: similar pairwise $\{\mathbf{x}_i^a\}^k, \{\mathbf{x}_i^p\}^k$, and parameters $\alpha, \mu, c_k$, and learning rate $\theta$
Output: learned metrics $\mathbf{M}_k,\ k=1,\dots,s$.
Initialize $\mathbf{M}_k^0=\mathbf{I}, \mathbf{G}_k^0=(1-\lambda)c_k\sum_{i=1}^{i=N}\{\mathbf{C}_{i,}^{a,p}\}^k$, $t\leftarrow 0$, $N_t=\{\}$
while iteration $t\leq$ max iteration
for $i=1,2,3,...,N$
    Find all the dissimilar pairwise in triplets $T_{i,j}=(\mathbf{x}_i^a,\mathbf{x}_i^p,\mathbf{x}_{j->i}^n)$, which violate the hinge loss in (6), in order to update the set $N_{t+1}$.
end
for $k=1,2,3,...,s$
Compute the gradient of $\mathbf{M}_k^t$
    $\mathbf{G}_k^{t+1}\leftarrow\mathbf{G}_k^t-\mu\ c_k\sum_{(i,j)\in N_t-N_{t+1}}(\{\mathbf{C}_i^{a,p}\}^k-\{\mathbf{C}_{j->i}^{a,n}\}^k)$
        $+\mu\ c_k\sum_{(i,j)\in N_{t+1}-N_t}(\{\mathbf{C}_i^{a,p}\}^k-\{\mathbf{C}_{j->i}^{a,n}\}^k)$
Project onto the PSD cone and update
    $\mathbf{M}_k^{t+1}=\text{PSD}(\mathbf{M}_k^t-\theta\mathbf{G}_k^{t+1})$
end
$t\leftarrow t+1$
end
```

IV. Experimental Results

To illustrate the effectiveness of our method, we present experimental results on three publicly available multi-modal datasets: Long Distance Heterogeneous Face Database consisting of visible (VIS) and near infrared (NIR) face images [13], Multi-Modal UMD Active Authentication Dataset UMDAA-01 [9], [26], ARL Multi-Modal Visible and Polarimetric Face Database [12]. We compare the performance of our method with several recently introduced multi-modal metric leaning algorithms. We also compare the performance of our method with many state-of-the-art unimodal metric leaning algorithms where we simply concatenate the features from different modalities and feed them into the unimodal metric leaning algorithms (i.e. feature level fusion). The algorithms used in this paper for comparisons are described as follows:

- LMMTML (multi-metric): This is the proposed LMMTML algorithm for multi-modal verification. We learn $\mathbf{M}_k,\quad k=1,\dots,s$ using Algorithm 1. Then, the squared distance between two multi-modal feature vectors is calculated as
$$d(\mathbf{x}_i,\mathbf{x}_j)=\sum_{k=1}^{s}(\mathbf{x}_i-\mathbf{x}_j)^\top\mathbf{M}_k(\mathbf{x}_i-\mathbf{x}_j). \tag{14}$$
- LMMTML (fusion): We set the number of modalities s equal to one in our metric learning algorithm. Features from different modalities are concatenated into a long vector. Then this multiple modality feature vector is fed

into our LMMTML algorithm with $s = 1$ for learning a single metric. Finally, (1) is used to determine the distance between two vectors.

- LMMTML (feature): The objective here is to study how much information each feature vector brings in the final verification. We learn only one metric based on a certain individual modality feature vector. This corresponds to the LMTML when $s = 1$ and only a single modality is used for learning a metric. (1) is used to determine the distance between two vectors.
- JB: We implemented the joint Bayesian unimodal metric learning algorithm proposed in [3] for face verification. In this algorithm, the concatenated features are treated as a single fusion-feature, and learn the metric based on that fusion-feature.
- TDE: We implemented the triplet distance embedding metric learning algorithm proposed [20] for face verification. In this algorithm, the concatenated features are treated as a single fusion-feature, and learn the metric based on that fusion-feature.
- LM3L: We implemented the large margin multi-metric learning algorithm proposed in [11]. In this algorithm, multiple metrics are learned for each individual modality features.
- PMML: We implemented the pairwise constraint-based multiple metric learning algorithm proposed in [5]. In this algorithm, multiple metrics are learned for each individual modality features.
- SML: We implemented the similarity metric learning algorithm proposed in [1]. In this algorithm, the concatenated features are treated as a single fusion-feature, and learn the metric based on that fusion-feature.
- DDML: We implemented the discriminative deep metric learning algorithm proposed in [10]. In this algorithm, the concatenated features are treated as a single fusion-feature, and learn the deep neural network-based metric on that fusion-feature.

Equal error rate (EER) and the ROC curves are used to measure the performance of different methods. All the parameters of our algorithm are obtained by cross validation.

A. *Heterogeneous NIR-VIS Face Dataset*

In the first set of experiments, we use visible and near infrared faces as different modalities. Long Distance Heterogeneous Face Database (LDHF) database [13] consists of visible and near-infrared face images of 100 individuals (70 males and 30 females). The face images were captured in both daytime and nighttime at different standoffs (e.g., 1m, 60m 100m, and150m) resulting in four VIS-NIR pairs per subject. Sample image pairs from this dataset are shown in Figure 3. The face area is cropped and resized to a fixed size of 64 × 56 pixels. We reduce the dimensionality of the VIS and NIR image features to 59 and 95, respectively using principal component analysis (PCA) by capturing 95% energy in the corresponding principle components. We randomly split 60 subjects' data as training, 20 subjects' data as validation set and 20 subjects data as the test set with no overlap among them. We repeat this process three times and obtain the average ROC curves. In order to generate triplets, we randomly choose two templates corresponding to the same subject from the training set as positive and anchor and another template from a different subject as negative sample. This way, many triplets can be generated from the training set.

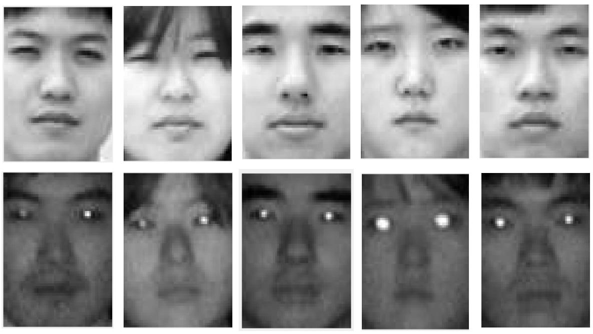

Fig. 3: Sample NIR-VIS image pairs from the LDHF database. The first row shows the VIS images and the second row shows the NIR images.

The parameters we set for this experiment are: $\alpha = 0.1, \mu = 0.5, \theta = 0.01, \text{maxiter} = 51$. The parameter c_k is uniformly chosen between $[0, 1]$. Here they are 0.5 and 0.5 for those two unit normalized features. The average ROC curves and the EER values corresponding to this experiment are shown in the first row of Figure 2, and Table I, respectively. From the results shown on the left column, we see that simple concatenation of visible and near infrared features produce better results than using the individual features. However, when the multi-modal metrics are learned directly from multi-modal data using our method, we achieve the best results.

From the ROC curves on the left column of the first raw of this figure, we see that our method performs the best on this data compared to some recent multi-modal metric learning methods. Feature concatenation-based joint Bayesian metric seems to produce comparable results. The PMML method performs poorly on this dataset. This may be due to the small data size problem in this dataset.

Also we plot the average imposter number (the negative, $x^n_{j->i} \in T_{i,j}$, which violate the hinge condition in (7)) vs the number of iterations in Figure 4 corresponding the experiments with the NIR-VIS data. As can be seen from this figure, that our method is able to decrease the number of imposters in a few iterations, which means that our algorithm pushes away the dissimilar pairs whose distance is apparently closer than the similar pairs. As a result, our method converges in a few iterations.

B. *UMDAA-01 Multi-Modal Active Authentication Dataset*

In mobile active authentication systems, users are continuously monitored after the initial access to the mobile device by making use of their physiological and behavioral biometrics captured by the built-in sensors and accessories

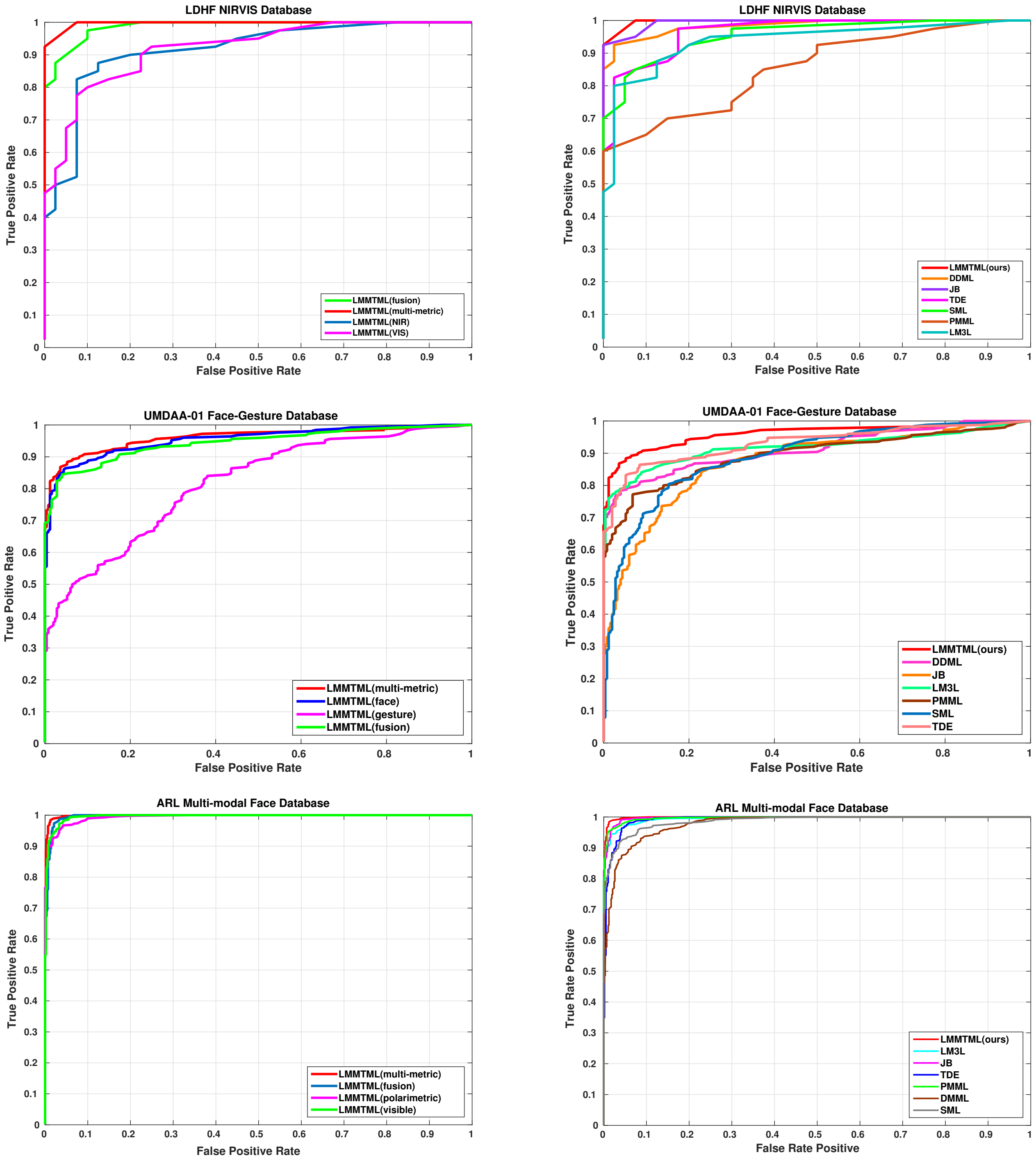

Fig. 2: The average ROC curves correspond to different methods. The plots on the first column show the multiple-modality vs single-modality results corresponding to our method. The plots on the second column show the comparison of our method with other metric learning algorithms. These results are the average of three random trials.

	LMMTML (ours)	LMMTML (fusion)	LM3L [11]	JB [3]	TDE [20]	PMML [5]	DDML [10]	SML [1]
EER-NIRVIS	**0.0500**±0.0661	0.0750±0.1041	0.1250±0.0750	0.0750±0.0250	0.1500±0.0500	0.3000±0.0722	0.0750±0.0144	0.1250± 0.0520
EER-UMDAA-01	**0.0920**±0.0356	0.1320±0.0485	0.1360±0.0363	0.2040±0.0220	0.1320±0.0635	0.1840±0.0583	0.1640±0.1100	0.1840±0.0201
EER-ARL	**0.0156**±0.0184	0.0243±0.0076	0.0417±0.0148	0.0296±0.00460	0.0417±0.0020	0.0330±0.0490	0.0833±0.0135	0.0642±0.0096

TABLE I: Average (mean ± std) EER values corresponding to different methods.

such as gyroscope, front-facing camera, accelerometer, and

pressure sensor [18]. In the second set or experiments, we use

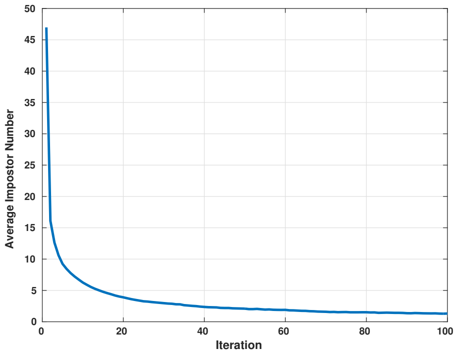

Fig. 4: The number of average negatives which violate the hinge loss in (7).

Fig. 5: Samples face images from the UMDAA-01 active authentication dataset.

the UMDAA-01 multi-modal active authentication dataset [9], [26] consisting of face images and touch gestures of 50 individuals collected from an iPhone 5 device. We extract the Local Binary Patterns (LBP) features from the face images and use PCA to reduce their dimension to 121 by keeping only 72% energy in the corresponding principle components. From each touch gesture, we extract a 27-dimensional feature vector using the method described in [26]. Sample face images from the UMDAA-01 are shown in Figure 5.

As before, we split the data by subject into $60\%, 20\%$ and 20% for training, validating and testing without overlaps among them. The parameters we set for this experiment are: $\alpha = 0.1, \mu = 0.5, \theta = 0.1, \text{maxiter} = 51$. The parameter c_k is uniformly chosen between $[0, 1]$. Here they are 0.5 and 0.5 for those two unit normalized features. The average ROC curves and the EER values corresponding to this experiment are shown in the second row of Figure 2, and Table I, respectively.

As can be seen from these results that our method performs the best compared to the other metric learning methods. Due to the weak nature of touch gestures, they perform worse than faces. Furthermore, a single metric learning algorithm based on the concatenated features seems to perform well on this dataset. This makes sense because the face modality introduces a significant amount of bias in the concatenated feature compared to touch gesture features. As a result, face modality dominates in the concatenated feature. This can be clearly seen by comparing the blue and the green curves on the left sub-figure.

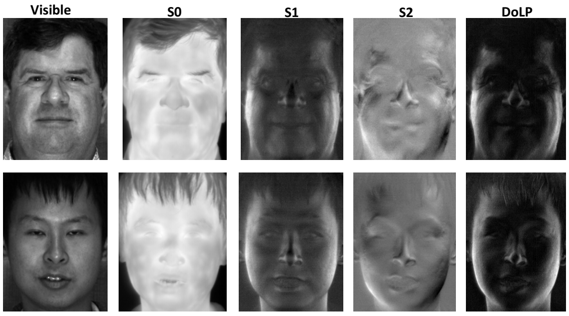

Fig. 6: Sample images corresponding to two subjects from the ARL multi-modal face database [12].

C. ARL Multi-Modal Visible and Polarimetric Face Database

In the final set of experiments, we use the ARL multi-modal visible and polarimetric face dataset [12] which consists of visible and polarimetric face images from 60 subjects. In particular, for each subject there are visible and S0, S1,S2, and DoLP (4 stroke parameters) images. These images were taken at different ranges (i.e. 31m, 44m, and 87m). Sample images from this dataset are shown in Figure 6. In this experiment, we use visible images as one modality and concatenating S0, S1,S2 as the polarimetric modality. We extracted 4640 dimension LBP features from all the visible and polarimetric images and use PCA to reduce their dimension to 54 for both.

As before, we split the data by subject into $60\%, 20\%$ and 20% for training, validating and testing without overlaps among them. The parameters we set for this experiment are: $\alpha = 0.1, \mu = 0.5, \theta = 0.01, \text{maxiter} = 101$. The parameter c_k is uniformly chosen between $[0, 1]$. Here they are 0.5 and 0.5 for those two unit normalized features. The average ROC curves and the EER values corresponding to this experiment are shown in the last row of Figure 2, and Table I, respectively. As can be seen from these results, our method performs comparably to some of the recent metric learning algorithms such as Joint Bayesian, LM3L and PMML. From the left-subfigure, we see that polarimetric features perform slightly worse than the original visible features. However, when they are combined, their performance increases significantly.

These experiments clearly show the significance of using our proposed LMMTML algorithm for multi-modal verification problems. In particular, it shows that when learning a metric based on a triplet loss, it is important to not only enforce the distance between the anchor and the positive samples to be less than the distance between the anchor and the negative samples but also make the distance between the anchor and the positive samples as small as possible.

V. Conclusion

We presented a multi-modal extension of the large margin nearest neighbor algorithm for multi-metric leaning. A simple two step iterative procedure was developed to solve the the proposed optimization problem. Extensive experiments on three real-world verification datasets demonstrate that the proposed method is very effective for multi-modal verification compared to some of the recent state-of-the-art metric learning methods.

In the future, we will investigate the possibility of applying our LMMTML algorithm on various biometrics verification problems such as video-based face verification and multi-modal biometrics fusion.

References

[1] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2408–2415, 2013.

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579. Springer, 2012.

[3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3554–3561, 2013.

[6] S. Dai and H. Man. Statistical adaptive metric learning in visual action feature set recognition. *Image and Vision Computing*, 55, Part 2:138 – 148, 2016. Handcrafted vs. Learned Representations for Human Action Recognition.

[7] J. V. Davis and I. S. Dhillon. Structured metric learning for high dimensional problems. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 195–203, New York, NY, USA, 2008. ACM.

[8] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine Learning*, pages 209–216, June 2007.

[9] M. E. Fathy, V. M. Patel, and R. Chellappa. Face-based active authentication on mobile devices. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1687–1691. IEEE, 2015.

[10] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[11] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Asian Conference on Computer Vision*, pages 252–267. Springer, 2014.

[12] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan. A polarimetric thermal database for face recognition research. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 119–126, 2016.

[13] D. Kang, H. Han, A. K. Jain, and S.-W. Lee. Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, 47(12):3750–3766, 2014.

[14] B. Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364.

[15] D. K. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In *International Conference on Machine Learning*, 2013.

[16] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu. Semi-supervised sparse metric learning using alternating linearization optimization. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1139–1148, 2010.

[17] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S. Chang. Low-rank similarity metric learning in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2792–2799, 2015.

[18] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous user authentication on mobile devices: Recent progress and remaining challenges. *IEEE Signal Processing Magazine*, 33(4):49–61, July 2016.

[19] R. Rosales and G. Fung. Learning sparse metrics via linear programming. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 367–373, 2006.

[20] S. Sankaranarayanan, A. Alavi, and R. Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[21] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[22] P. Wu, S. C. Hoi, P. Zhao, C. Miao, and Z.-Y. Liu. Online multimodal distance metric learning with application to image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):454–467, 2016.

[23] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 521–528. MIT Press, 2003.

[24] L. Yang. Distance metric learning: A comprehensive survey, 2006.

[25] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2214–2222. Curran Associates, Inc., 2009.

[26] H. Zhang, V. M. Patel, M. Fathy, and R. Chellappa. Touch gesture-based active user authentication using dictionaries. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 207–214. IEEE, 2015.

[27] Y. Zhang, H. Zhang, N. M. Nasrabadi, and T. S. Huang. Multi-metric learning for multi-sensor fusion based classification. *Information Fusion*, 14(4):431–440, 2013.