Multi-Level Metric Learning via Smoothed Wasserstein Distance

Jie Xu[1,2], **Lei Luo**[2], **Cheng Deng**[1], **Heng Huang**[2,1*]
[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2] Electrical and Computer Engineering, University of Pittsburgh, PA, 15261, USA

Abstract

Traditional metric learning methods aim to learn a single Mahalanobis distance matrix $\mathbf{M}$, which, however, is not discriminative enough to characterize the complex and heterogeneous data. Besides, if the descriptors of the data are not strictly aligned, Mahalanobis distance would fail to exploit the relations among them. To tackle these problems, in this paper, we propose a multi-level metric learning method using a smoothed Wasserstein distance to characterize the errors between any two samples, where the ground distance is considered as a Mahalanobis distance. Since smoothed Wasserstein distance provides not only a distance value but also a flow-network indicating how the probability mass is optimally transported between the bins, it is very effective in comparing two samples whether they are aligned or not. In addition, to make full use of the global and local structures that exist in data features, we further model the commonalities between various classification through a shared distance matrix and the classification-specific idiosyncrasies with additional auxiliary distance matrices. An efficient algorithm is developed to solve the proposed new model. Experimental evaluations on four standard databases show that our method obviously outperforms other state-of-the-art methods.

1 Introduction

Metric learning problem targets at learning an optimal distance matrix $\mathbf{M}$ via minimizing the distance function, such as Mahalanobis distance

$$d_{\mathbf{M}(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}, \tag{1}$$

to capture the important relationships among data for a given task. It has been shown to be indispensable when used in conjunction with many computer vision techniques that heavily rely on distances or similarities [Roth *et al.*, 2014; Hu *et al.*, 2014].

To obtain a well-learned $\mathbf{M}$, various kinds of metric learning algorithms have been proposed, *e.g.*, large-margin nearest neighbors (LMNN) [Weinberger and Saul, 2009], information-theoretic metric learning (ITML) [Davis *et al.*, 2007], logistic discriminant metric learning (LDML) [Guillaumin *et al.*, 2009], robust metric learning [Luo and Huang, 2018; Wang *et al.*, 2014], *etc*. Among all these works, LMNN is widely used and generalized. If we denote the similar sample pairs by $\mathcal{S}$ and triplet constraint by $\mathcal{R}$ as:

$$\begin{aligned}&\mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are similar}\},\\ &\mathcal{R} = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : \mathbf{x}_i \text{ is more similar to } \mathbf{x}_j \text{ than to } \mathbf{x}_k\},\end{aligned} \tag{2}$$

then LMNN model can be formulated as:

$$\begin{aligned}\min_{\mathbf{M}\in\mathbb{S}_+} \quad & (1-\mu) \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) + \mu \sum_{(i,j,k)\in\mathcal{R}} \xi_{ijk}\\ s.t. \quad & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k) - d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \geq 1 - \xi_{ijk},\\ & \forall(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k) \in \mathcal{R},\end{aligned} \tag{3}$$

where $\mu \in [0, 1]$ controls relative weight between two terms, and ξ_{ijk} is a safety margin distance for each triplet. Although LMNN is proved to be very effective in learning a good Mahalanobis distance for practical problems, it is sensitive to Euclidean distance when it computes neighbors of each sample at the beginning.

As with LMNN, most metric learning works mainly focus on improving the discriminability of the distance matrix

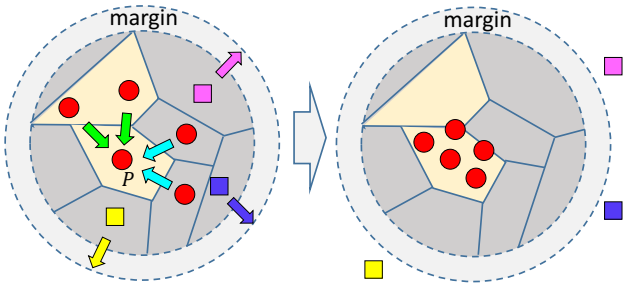


Figure 1: An illustration of the proposed method. Red filled circles are target neighbors lied in different regions of sample P's local neighborhood. Solid squares are differently labeled data which also lied in P's local neighborhood. During training process, a global distance matrix and several local distance matrices are learned with respect to different regions. After training, similarly labeled samples get closer while differently labeled samples are pulled farther away from sample P.

*To whom all correspondence should be addressed.

$\mathbf{M}$ via minimizing the squared Mahalanobis distance from similar sample pairs. Especially, the descriptors of the sample pairs they address are usually assumed to be aligned in advance. However, in practice, such an assumption is often violated due to some unavoidable factors like geometrical deformation, non-linear lighting changes and heavily intensity noise. In this case, Mahalanobis distance is not applicable due to its limitation on descriptors of sample data.

Fortunately, the Wasserstein distance, as a cross-bin distance function, is good at dealing with such kind of hard-to-align problem. It at present plays an important role in many practical applications, including computer vision, statistics and *etc* [Bisot *et al.*, 2015]. The original Wasserstein distance is motivated by the theory of optimal transport and some related methods used in image processing and compressed sensing [Rubner *et al.*, 2000]. Due to its high time complexity, many researchers try to improve the efficiency by simplifying the linear programming formulation with regard to Wasserstein distance or designing a parallel method for Wasserstein distance related problems [Ling and Okada, 2007; Li *et al.*, 2017].

In this paper, we proposed a novel multi-level metric learning method using smoothed Wasserstein distance as a measure of errors between samples. To enhance the discriminability of model, we choose the ground distance as a Mahalanobis distance. Meanwhile, to take advantage of the crucial structured information that exist in data space, we learn a global smoothed Wasserstein distance through a shared distance matrix and local smoothed Wasserstein distance with additional auxiliary distance matrices, where the former reflects commonalities between various classification and the later captures classification-specific idiosyncrasies. Smoothed Wasserstein distance yields not only a distance value but also a optimal transformation plan. Thus, our method is more stable in characterizing the differences between two samples with noise than traditional Euclidean or Mahalanobis distance. We propose to jointly optimize the Mahalanobis distance matrix in the ground distance and the Wasserstein distance flow-network by using the alternative iterative strategy. The effectiveness of the proposed method is demonstrated by a series of computer vision tasks.

For a given vector $\mathbf{a} = (a_1, a_2, \cdots, a_d)^\top$, $\mathrm{diag}(\mathbf{a}) = \mathbf{A}$ corresponds to a square diagonal matrix such that $\forall i, A_{i,i} = a_i$. $\mathbf{e}$ represents a unit vector, and $\mathbf{I}$ is the unit matrix.

2 Multi-Level Metric Learning via Smoothed Wasserstein Distance

We start this section by revisiting the Smooth Wasserstein distance, and then propose our multi-level metric learning model.

2.1 Smoothed Wasserstein Distance

Wasserstein distance is actually the optimal solution of transportation problem in linear programming. It can be viewed as a minimum amount of work required to move all the earth from the source to the destination [Sandler and Lindenbaum, 2011]. Given two signatures $P = \{(\mathbf{p}_1, w_{p_1}), (\mathbf{p}_2, w_{p_2}), \cdots, (\mathbf{p}_m, w_{p_m})\}$ and $Q = \{(\mathbf{q}_1, w_{q_1}), (\mathbf{q}_2, w_{q_2}), \cdots, (\mathbf{q}_n, w_{q_n})\}$, the Wasserstein distance between them is defined as

$$W(P,Q) = \min_{\mathbf{F}\in\mathbb{F}(P,Q)} Tr(\mathbf{D}^\top \mathbf{F}) \tag{4}$$

where $\mathbf{D} = \{d(i,j)\}, i = 1, \cdots, m, j = 1, \cdots, n$ is the ground distance matrix, and $d(i,j)$ defines the cost of moving one unit of earth from the source $\mathbf{p}_i$ to the target $\mathbf{q}_j$. $\mathbf{F} = \{f(i,j)\}, i = 1, \cdots, m, j = 1, \cdots, n$ is a flow-network matrix, and $f(i,j)$ denotes the amount of earth moved from the source $\mathbf{p}_i$ to the target $\mathbf{q}_j$.

Let $\mathbf{w_p} = [w_{p_1}, w_{p_2}, \cdots, w_{p_m}] \in \mathbb{R}^m$, $\mathbf{w_q} = [w_{q_1}, w_{q_2}, \cdots, w_{q_n}] \in \mathbb{R}^n$, then $\mathbb{F}(P,Q)$ can be written as

$$\begin{aligned}\mathbb{F}(P,Q) = \{\mathbf{F} | \mathbf{F} \in \mathbb{R}^{m\times n}, \mathbf{F}^\top \mathbf{1}_m = \mathbf{w_q}, \\ \mathbf{F}\mathbf{1}_n = \mathbf{w_p}, \mathbf{F}_{ij} \geq 0, \forall i,j\}.\end{aligned} \tag{5}$$

Optimizing Wasserstein distance problem is actually solving several costly optimal transport problems. Furthermore, the Wasserstein itself is not a smoothed function of its arguments because of a minimum of affine function, which limits the application of Wasserstein distance. To overcome the above problems, some researchers proposed to smooth the optimal transport problem with an entropic term [Cuturi, 2013]

$$W_\gamma(P,Q) = \min_{\mathbf{F}\in\mathbb{F}(P,Q)} Tr(\mathbf{D}^T\mathbf{F}) - \gamma h(\mathbf{F}), \tag{6}$$

where h is the (strictly concave) entropy function

$$h(\mathbf{F}) = -\langle \mathbf{F}, \log \mathbf{F} \rangle. \tag{7}$$

and $\gamma > 0$ is a balance parameter. In this paper, we call Eq. (6) as smoothed Wasserstein distance.

2.2 New Multi-Level Metric Learning Model Using Smoothed Wasserstein Distance

The ground distance in Eq. (6) is usually Euclidean, cosine or Sparse L_1-norm distances. However, these distances can neither admit arbitrary linear scalings and rotations of the feature space nor exploit the discriminative information that exists in data space. Therefore, in this paper, we use the Mahalanobis distance as the ground measurement to improve the discriminative capability of W_γ. The squared Mahalanobis distance between the i-th bin of P and the j-th bin of Q can be expressed as

$$d_{\mathbf{M}}(i,j) = (\mathbf{p}_i - \mathbf{q}_j)^\top \mathbf{M} (\mathbf{p}_i - \mathbf{q}_j), \tag{8}$$

where $\mathbf{M}$ is a global linear transformation of the underlying space, and $\mathbf{D}_{\mathbf{M}} = \{d_{\mathbf{M}}(i,j)\}$ is the ground distance.

On the basis of Eq. (8), we can construct a Mahalanobis smoothed Wasserstein distance ($W_{\gamma,\mathbf{M}}$ for short). Let the smoothed Wasserstein distance between signatures P and Q be $W_{\gamma,\mathbf{M}}(P,Q)$ and the triplet be $(P,Q,S) \in \mathcal{R}$. Then, replacing the Mahalanobis distance in Eq. (3) with $W_{\gamma,\mathbf{M}}$, a novel LMNN model based on $W_{\gamma,\mathbf{M}}$ is obtained as follows

$$\begin{aligned}\min_{\mathbf{M}\in\mathbb{S}_+} (1-\mu) \sum_{(P,Q)\in\mathcal{S}} W_{\gamma,\mathbf{M}}(P,Q) + \mu \sum_{(i,j,k)\in\mathcal{R}} \xi_{ijk} \\ \text{s.t.} \quad W_{\gamma,\mathbf{M}}(P,R) - W_{\gamma,\mathbf{M}}(P,Q) \geq 1 - \xi_{ijk}, \\ \forall (P,Q,R) \in \mathcal{R}.\end{aligned} \tag{9}$$

As we can see, model (9) only uses a global $W_{\gamma,\mathbf{M}}$, thus it is hard to capture the local discriminative information of samples. To further improve the performance of (9), we model the commonalities between various classification through a shared Mahalanobis distance matrix $\mathbf{M}_0 \succeq 0$ and the classification-specific idiosyncrasies with additional matrices $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_C \succeq 0$ in the ground distance, where C denotes the number of classes in samples. As a result, we can rephrase Eq. (8) as

$$d_{\mathbf{M},t}(i,j) = (\mathbf{p}_i - \mathbf{q}_j)^\top (\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{p}_i - \mathbf{q}_j), \quad (10)$$

where $t = 1, 2, \cdots, C$.

Additionally, we can easily prove that each $\sqrt{d_{\mathbf{M},t}(i,j)}$ is indeed a pseudo-metric. Using Eq. (10) as the ground distance, problem (6) can be rewritten as

$$W_{\gamma,\mathbf{M},t} = \min_{\mathbf{F}\in\mathbb{F}(P,Q)} Tr(\mathbf{D}_{\mathbf{M},t}^T \mathbf{F}) - \gamma h(\mathbf{F}). \quad (11)$$

where $\mathbf{D}_{\mathbf{M},t} = \{d_{\mathbf{M},t}(i,j)\}, \forall i,j$.

In modeling, we have to ensure that the learning algorithm does not put too much emphasis onto the shared parameters $\mathbf{M}_0$ or the individual parameters $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_C$. To ensure this balance, we use the regularization term stated below

$$\min_{\mathbf{M}_0,\cdots,\mathbf{M}_C} \rho_0 \|\mathbf{M}_0 - \mathbf{I}\|_F^2 + \sum_{t=1}^{C} \rho_t \|\mathbf{M}_t\|_F^2, \quad (12)$$

where $\mathbf{I}$ is an identity matrix and the trade-off parameter ρ_t controls the regularization of $\mathbf{M}_t$ for all $t = 0, 1, \cdots, C$. Therefore, model (9) is further rewritten as

$$\begin{aligned} &\min_{\mathbf{M}\in\mathbb{S}_+} \rho_0 \|\mathbf{M}_0 - \mathbf{I}\|_F^2 + \sum_{t=1}^{C} (\rho_t \|\mathbf{M}_t\|_F^2 \\ &+ \sum_{(P,Q)\in\mathcal{S}} (1-\mu) W_{\gamma,\mathbf{M},t}(P,Q) + \mu \sum_{(i,j,k)\in\mathcal{R}} \xi_{ijk} \quad (13) \\ &\text{s.t.} \quad W_{\gamma,\mathbf{M},t}(P,R) - W_{\gamma,\mathbf{M},t}(P,Q) \geq 1 - \xi_{ijk}, \\ &\qquad\qquad \forall (P,Q,R) \in \mathcal{R}. \end{aligned}$$

Different from the existing metric learning methods, which use a global Mahalanobis metric to measure the differences between samples, we use multiple smoothed Wasserstein distances to adaptively learn global and local structures of data space, which is helpful for practical multi-classification problems. However, solving problem (13) is very challenging since it needs to optimize multiple Wasserstein distance problems. This may lead to high computation complexity. In the next section, we will propose an efficient algorithm to solve model (13).

3 Optimization Algorithm

There are two groups of variables that need to be learned in model (13), *i.e.*, flow-network $\mathbf{F}$ and matrices $\mathbf{M}_0, \mathbf{M}_1, \cdots, \mathbf{M}_C$. When flow-network $\mathbf{F}$ is fixed, model (13) turns into the Mahalanobis-like metric learning problem, *i.e.*,

$$\begin{aligned} &\min_{\mathbf{M}\in\mathbb{S}_+} \rho_0 \|\mathbf{M}_0 - \mathbf{I}\|_F^2 + \sum_{t=1}^{C} (\rho_t \|\mathbf{M}_t\|_F^2 \\ &+ \sum_{(P,Q)\in\mathcal{S}} (1-\mu) Tr(\mathbf{D}_{\mathbf{M},t}(P,Q)^\top \mathbf{F}) + \mu \sum_{(i,j,k)\in\mathcal{R}} \xi_{ijk}) \\ &\text{s.t.} \quad Tr(\mathbf{D}_{\mathbf{M},t}(P,R)^\top \mathbf{F}) - Tr(\mathbf{D}_{\mathbf{M},t}(P,Q)^\top \mathbf{F}) \geq 1 - \xi_{ijk}, \\ &\qquad\qquad \forall (P,Q,R) \in \mathcal{R}. \end{aligned} \quad (14)$$

Though the above problem is non-convex, we can compute its sub-gradient with

$$\begin{aligned} \nabla_{\mathbf{M}_0} &= 2\rho_0(\mathbf{M}_0 - \mathbf{I}) + C((1-\mu) \sum_{(P,Q)\in\mathcal{S}} G_{P,Q} \\ &\quad + \mu \sum_{(P,Q,R)\in\mathcal{R}'} (G_{P,Q} - G_{P,R})), \\ \nabla_{\mathbf{M}_t} &= \sum_{t=1}^{C} (2\rho_t \mathbf{M}_t + (1-\mu) \sum_{(P,Q)\in\mathcal{S}} G_{P,Q} \\ &\quad + \mu \sum_{(P,Q,R)\in\mathcal{R}'} (G_{P,Q} - G_{P,R})), \\ \mathcal{R}' &= \mathcal{R} - \{(P,Q,R) | tr(G_{P,R}^\top \mathbf{M}) - tr(G_{P,Q}^\top \mathbf{M}) > \nu\} \\ G_{P,Q} &= G_{P,Q}^1 - G_{P,Q}^0, \\ G_{P,Q}^1 &= P\text{diag}(\mathbf{F}\mathbf{e})P^\top + Q\text{diag}(\mathbf{e}^\top \mathbf{F})Q^\top, \\ G_{a,b}^0 &= P\mathbf{F}Q^\top + Q\mathbf{F}^\top P^\top. \end{aligned} \quad (15)$$

where $t = 1, 2, \cdots, C$.

Next, we can update each $\mathbf{M}_t (t = 0, 1, \cdots, C)$ using sub-gradient method, *i.e.*,

$$\mathbf{M}_t \Leftarrow \mathcal{P}_{\mathbb{S}_+}(\mathbf{M}_t - \tau \nabla_{\mathbf{M}_t}), \quad (16)$$

where $\mathcal{P}_{\mathbb{S}_+}(\cdot)$ denotes the projection operator and $\tau > 0$ is a step size. In our method, we set $\rho_0 = 1$, and $\rho_t = \frac{1}{C}, t = 1, \cdots, C$. We can use the similar technique as in [Weinberger and Saul, 2009] to reduce the computation complexity of (16).

When $\mathbf{M}_t, t = 0, 1, \cdots, C$ are fixed, problem (13) can be splited into some independent traditional smoothed Wasserstein distance sub-problems, which can be solved by the method in [Rolet *et al.*, 2016]. Here we omit the detailed process for solving smoothed Wasserstein distance sub-problems. Algorithm 1 summarizes the main steps of the proposed method.

4 Experimental Results

In this section, to prove the effectiveness of the proposed method, we apply it to three kinds of computer vision applications, *i.e.*, person re-identification, facial kinship verification and video classification.

4.1 Person Re-identification

During this part, we evaluate the proposed method on person re-identification task, which aims to match a given probe

Algorithm 1 Optimization Algorithm for solving Problem (13)

Input: Training data set, and parameters k_g, k_i, ρ_0, $\{\rho_t, t = 1, \cdots, C\}$, γ, λ and μ
Output: Global distance matrix $\mathbf{M}_0$ and local distance matrices $\mathbf{M}_t, t = 1, \cdots, C$
Initialize: $\mathbf{M}_t = \mathbf{I}, t = 0, \cdots, C$
Calculate the smoothed Wasserstein distance between each pair of signatures.
Construct the set $\mathcal{S}$ and $\mathcal{R}$ by selecting k_i target neighbors and k_g imposters for each training instance.
Initialize the learning rate α
repeat
    Fix $\mathbf{M}_t, t = 0, \cdots, C$, solve for the Wasserstein distance flow-network $\mathbf{F}$.
    Fix $\mathbf{F}$.
    repeat
        Compute the gradient $\nabla \mathbf{M}_t, t = 0, \cdots, C$ by Eq. (15).
        Update $\mathbf{M}_t, t = 0, \cdots, C$ by Eq. (16).
    until Converge
until Converge

image against a collection of gallery images. We use two challenge person re-identification datasets at multi-shot scenario, *i.e.*, PRID 2011 dataset [Hirzer *et al.*, 2011] and iLIDS-VID dataset [Office, 2008]. Some example pairs appeared in different camera views from these two datasets are shown in Fig. 2.

- **PRID 2011 dataset** - The PRID 2011 dataset includes 385 persons showed in camera view A and 749 persons in camera view B, with 200 persons of them appeared in both camera views [Hirzer *et al.*, 2011].
- **iLIDS-VID dataset** - The iLIDS-VID dataset involves 300 different pedestrians observed across two disjoint camera views in public open space [Office, 2008].

Compared Methods: We compute Wasserstein distance, also called Earth Mover's Distance (EMD) directly as a baseline. Besides, we also compare the proposed method against four representative methods, including Pairwise Constrained Component Analysis (PCCA) [Mignon and Jurie, 2012], KISSME [Koestinger *et al.*, 2012], Local Fisher Discriminant Analysis (LFDA) [Pedagadi *et al.*, 2013] and Marginal Fisher Analysis (MFA) [Yan *et al.*, 2007].

Experiment Settings: In the experiment, we split each dataset into two folds. In each time, one fold of data is for training and the other fold is used as testing data. We randomly generate five splits, and the average results are as final performance. Specifically, for PRID 2011 dataset, there are 200 person image pairs. Half of them are randomly selected for training and the rest are for testing. To evaluate the test set, we follow the same procedure described in [Hirzer *et al.*, 2011], *i.e.*, camera A is used as probe set and camera B is as gallery set. Thus, each of the 100 persons in the probe set is searched in a gallery set of 649 persons. For iLIDS-VID dataset, there are 300 person image pairs. All these pairs are randomly divided into two folds of the same size.

Method	PUR	rank=1	rank=10	rank=20
PCCA	21.69	4.88	24.68	35.06
KISSME	27.91	15.62	38.16	48.52
LFDA	21.30	16.94	48.64	59.88
MFA	20.84	16.26	48.44	59.00
EMD	31.81	10.80	31.20	39.60
Proposed	**46.58**	**43.80**	**78.20**	**85.60**

Table 1: CMC at $rank = 1, 10, 20$ and PUR scores on PRID 2011 dataset with 100 test individuals searched in a gallery set of 649 individuals. Red and blue numbers are the best and second best results, respectively.

Method	PUR	rank=1	rank=10	rank=20
PCCA	14.08	9.36	42.88	61.64
KISSME	8.05	7.22	35.22	50.72
LFDA	22.01	16.28	58.22	74.80
MFA	22.86	16.06	58.86	75.36
EMD	12.31	6.93	23.73	33.20
Proposed	**43.18**	**35.47**	**79.60**	**90.93**

Table 2: CMC at $rank = 1, 10, 20$ and PUR scores on iLIDS-VID dataset with 300 test individuals searched in a gallery set of 300 individuals. Red and blue numbers are the best and second best results, respectively.

We focus on multi-shot scenario. 26960-d Local Maximal Occurrence (LOMO) feature is extracted for each frame of both datasets [Liao *et al.*, 2015]. PCA is further applied to reduce the feature dimension to 100-d for our method and 65-d for KISSME method. In addition, to decrease the quantity of calculation of flows in our method, we cluster each sequence into several frames and use clustering centers to represent the corresponding sequence. For those only applicable to single-shot scenario methods, we average the centers as features. We choose linear kernel for MFA method.

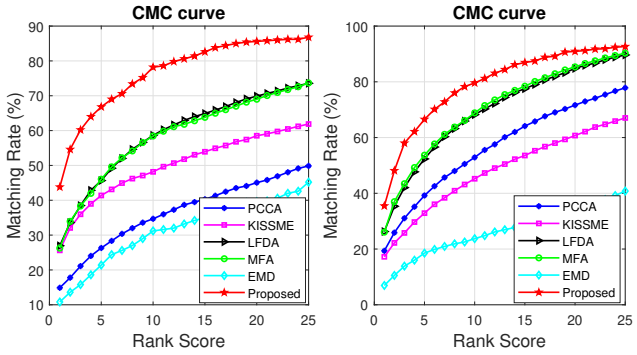
Experiment Results: To measure the performance of person re-identification task, we report the widely used Cumulative Match Characteristic (CMC) performance curves averaged across the experiments in Fig. 3. It is obvious that our method always outperforms other compared methods, especially for PRID 2011 dataset. Because our method not only tries to learn a global distance matrix, but also takes advantage the local information. LFDA also achieves good performance because it does a better job at selecting the features with no need of PCA pre-processing step while making full use of the locally scaled affinity matrix. MFA obtains comparable results to LFDA. In contrast to LFDA, the advantage of MFA is its ability to maximize the marginal discriminant even when the assumption of a Gaussian distribution for each class is not true. With the increase of the rank value, LFDA and MFA catch up the proposed method. Simply use Wasserstein distance without metric learning couldn't take advantage of the prior knowledge in training data, thus performs worse in all compared methods.

In addition, we also report the Proportion of Uncertainty Removed (PUR) scores [Pedagadi *et al.*, 2013]:

$$\text{PUR} = \frac{\log(N) + \sum_{r=1}^{N} M(r) \log(M(r))}{\log(N)}, \qquad (17)$$

(a) PRID 2011 (b) iLIDS-VID

Figure 2: Example pairs of image sequences of the same person appearing in different camera views from (a) the PRID 2011 dataset, (b) the iLIDS-VID dataset.



(a) PRID 2011 (b) iLIDS-VID

Figure 3: CMC curves for each method on two datasets: (a) PRID 2011 dataset; (b) iLIDS-VID dataset.

where N is the size of the gallery set, r is the rank. The rank of the correct match was recorded and accumulated to generate the match characteristic $M(r)$. The PUR scores of two datasets along with the CMC values at $rank = 1, 10, 20$ are summarized in Table 1 and Table 2. The highest and second highest CMC and PUR scores in every experiment at every ranking were highlighted in red and blue, respectively. As shown in the table, the overall PUR score is higher for the PRID 2011 dataset, probably because the iLIDS-VID dataset is very challenging due to clothing similarities among people, cluttered background, viewpoint and lighting variations across camera views as shown in Fig. 2.

4.2 Facial Kinship Verification

In this section, we evaluate our methods on facial kinship verification task, which is to determine whether there is a kin relation between a pair of given face images. We use KinFaceW-II dataset, and some example pairs are shown in Fig. 4 [Lu *et al.*, 2014].

- **KinFaceW-II dataset** - KinFaceW-II dataset consists of four representative types of kin relations: Mother-Son (M-S), Father-Son (F-S), Mother-Daughter (M-D) and Father-Daughter (F-D), respectively. Each relation contains 250 pairs of kinship images.

Compared Methods: We compute Euclidean and EMD distance between a pair of faces directly as a baseline. Also, traditional Mahalanobis distance between images of a pair is computed, where the metric matrix is inverse of covariance between two vectors. We compare the proposed methods against two representative metric learning methods, *i.e.*, KISSME [Koestinger *et al.*, 2012] and LDML [Guillaumin *et al.*, 2009].

Experiment Settings: As a benchmark for comparison, we use the pre-specified training/testing split, which is generated for 5-fold cross validation [Lu *et al.*, 2014]. We use the given Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) from image blocks. Each kind of feature is as one feature vector of a signature. For other methods, we simply concatenate these two features. PCA is further employed to reduce dimensionality of each vector to 100 dimension.

Experiment Results: To measure the kinship verification accuracy for all these compared methods, we report a Receiver Operator Characteristic (ROC) curve in Fig. 5. It seems that our method achieves best results on M-S and M-D relations. For other two relations, our method can get comparable results. Since it is difficult to distinguish which performs better, we compute Equal Error Rate (EER) of the respective method, and use $1 - EER$ as evaluation criterion, as shown in Fig. 5.

4.3 Video Classification

In this section, we evaluate our methods on video classification task, which is assigning an video to a predefined class. We use traffic video database. Several frames of different videos from the database are shown in Fig. 6.

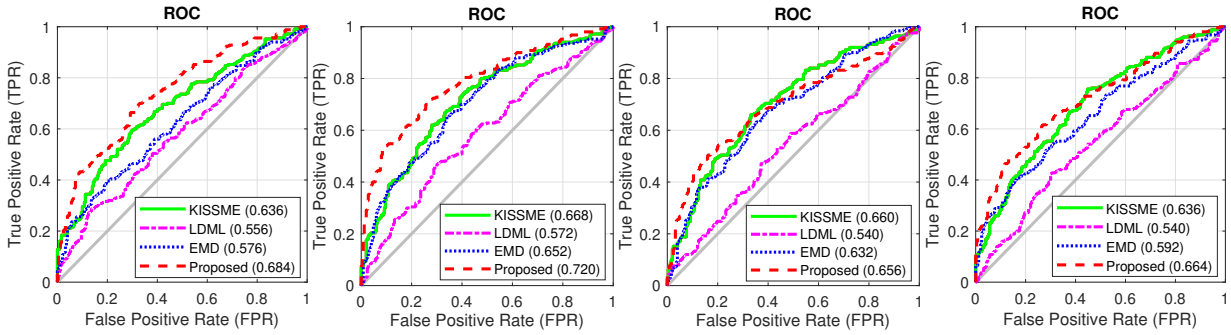
- **Traffic video database** - Traffic video database consists of 254 video sequences of highway traffic in Seattle [Chan and Vasconcelos, 2005]. As in [Chan and Vasconcelos, 2005], each sequence was converted to grayscale, resized to 80×60 pixels, and then clipped to a 48×48 window over the area with the most total motion. For each frame, we use gray values as feature representation

Compared Methods: We compare our method with five widely used classification methods, including K-nearest Neighbor method (KNN) [Peterson, 2009], Support Vector Machine (SVM) [Fan *et al.*, 2008], LMNN [Weinberger and Saul, 2009], and LMNN with capped trace norm [Huo *et al.*, 2016] and Fantope norm [Law *et al.*, 2014].

Experiment Settings: For our method, we simply use pixel intensities of each frame as feature vectors. Since each video sequence contains 42 to 52 frames, to reduce the amount of

(a) M-S (b) F-S (c) F-D (d) M-D

Figure 4: Example pairs of images from the KinFaceW-II database.



(a) M-S (b) F-S (c) F-D (d) M-D

Figure 5: Facial kinship verification results on all kinship relations of KinFaceW-II dataset: (a) Mother-Son kinship relation; (b) Father-Son kinship relation; (c) Father-Daughter kinship relation; (d) Mother-Daughter kinship relation.

Figure 6: Examples from the traffic video database.

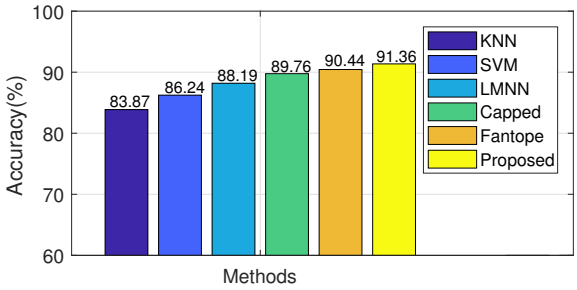


Figure 7: Classification accuracy of different methods on traffic video database.

computation, we cluster every sequence into several frames. We treat each video as a signature, and the cluster centers are its feature vectors. The weights of each signature are determined by the proportion of samples in each cluster. For other methods, we follow the same procedure described in [Chan and Vasconcelos, 2005], *i.e.*, for each video clip, the mean image is subtracted and the pixel intensities were normalized to have unit variance. For LMNN with capped trace norm and Fantope norm methods, the regularization parameters are tuned from range $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$, and parameter rank of matrix $\mathbf{M}$ is from $[30 : 5 : 70]$.

Experiment Results: We compute classification accuracy for each method as evaluation criterion. As shown in Fig. 7, KNN method using Euclidean distance works worst on this task. SVM method increases the performance of KNN method by about $2 \sim 3\%$. When we use LMNN method with or without norm to learn Mahalanobis distance, we choose KNN classifier ($K = 1$). The proposed method does a better job than other compared methods in this task, which proves the effectiveness of using smoothed Wasserstein distance.

5 Conclusion

In this paper we proposed a novel multi-level metric learning algorithm. In our method, the errors between two samples are characterized using the smoothed Wasserstein distance. We consider the ground distance as a Mahalanobis distanc and automatically learn the corresponding matrices by the alternative iterative approach. To enhance the robustness of model, we further learn the global distance with regard to all samples and the local distance with regard to samples with specific class label. It is noteworthy that our method can be extended to other metric learning models. We verify the abilities of our method on several real-world datasets. The experimental results show that the proposed method consistently outperforms some related methods and obtains better classification results than traditional Mahalanobis metric learning approaches.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grants 61572388,

61703327, the National Key Research and Development Program of China (2017YFE0104100), and in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02 and Grant 2017ZDCXL-GY-05-04-02.

References

[Bisot *et al.*, 2015] Victor Bisot, Slim Essid, and Gaël Richard. Hog and subband power distribution image features for acoustic scene classification. In *EUSIPCO*, pages 719–723. IEEE, 2015.

[Chan and Vasconcelos, 2005] Antoni B Chan and Nuno Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *CVPR*, pages 846–851, 2005.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505. IEEE, 2009.

[Hirzer *et al.*, 2011] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[Hu *et al.*, 2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014.

[Huo *et al.*, 2016] Zhouyuan Huo, Feiping Nie, and Heng Huang. Robust and effective metric learning using capped trace norm: Metric learning via capped trace norm. In *SIGKDD*, pages 1605–1614. ACM, 2016.

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012.

[Law *et al.*, 2014] Marc T Law, Nicolas Thome, and Matthieu Cord. Fantope regularization in metric learning. In *CVPR*, pages 1051–1058, 2014.

[Li *et al.*, 2017] WUCHEN Li, ERNEST K Ryu, STANLEY Osher, WOTAO Yin, and WILFRID Gangbo. A parallel method for earth mover's distance. *UCLA Comput. Appl. Math. Pub.(CAM) Rep*, pages 17–12, 2017.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Ling and Okada, 2007] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5):840–853, 2007.

[Lu *et al.*, 2014] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):331–345, 2014.

[Luo and Huang, 2018] Lei Luo and Heng Huang. Matrix variate gaussian mixture distribution steered robust metric learning. In *AAAI*, pages 933—-940, 2018.

[Mignon and Jurie, 2012] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672. IEEE, 2012.

[Office, 2008] UK Home Office. i-lids multiple camera tracking scenario definition. 2008.

[Pedagadi *et al.*, 2013] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, pages 3318–3325, 2013.

[Peterson, 2009] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[Rolet *et al.*, 2016] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638, 2016.

[Roth *et al.*, 2014] Peter M Roth, Martin Hirzer, Martin Koestinger, Csaba Beleznai, and Horst Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[Sandler and Lindenbaum, 2011] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

[Wang *et al.*, 2014] Hua Wang, Feiping Nie, and Heng Huang. Robust distance metric learning via simultaneous l1-norm minimization and maximization. In *ICML*, pages 1836–1844, 2014.

[Weinberger and Saul, 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[Yan *et al.*, 2007] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.