New Robust Metric Learning Model Using Maximum Correntropy Criterion

Jie Xu
School of Electronic Engineering
Xidian University
Xi'an, Shaanxi, China
jie.xu@pitt.edu

Lei Luo
Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA, USA
luolei2017@pitt.edu

Cheng Deng
School of Electronic Engineering
Xidian University
Xi'an, Shaanxi, China
chdeng.xd@gmail.com

Heng Huang*
Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA, USA
heng.huang@pitt.edu

ABSTRACT

Metric learning has recently become an active data mining research topic with many real-world applications. Most existing metric learning methods aim to learn an optimal Mahalanobis distance matrix $\mathbf{M}$, under which data samples from the same class are forced to be close to each other and those from different classes are pushed far away. The Mahalanobis distance matrix $\mathbf{M}$ can be factorized as $\mathbf{M} = \mathbf{L}^\top\mathbf{L}$, and the Mahalanobis distance induced by $\mathbf{L}$ is equivalent to the Euclidean distance after linear projection of the feature vectors on the rows of $\mathbf{L}$. However, the Euclidean distance is only suitable for characterizing Gaussian noise, thus the traditional metric learning algorithms are not robust to achieve good performance when they are applied to the occlusion data, which often appear in image and video data mining applications.

To overcome this limitation, we propose a new robust metric learning approach by introducing the maximum correntropy criterion to deal with real-world malicious occlusions or corruptions. In our new model, we enforce the intra-class reconstruction residual of each sample to be smaller than the inter-class reconstruction residual by a large margin. Meanwhile, we employ correntropy induced metric to fit the reconstruction residual, which has been proved to be useful in non-Gaussian data processing. Leveraging the half-quadratic optimization technique, we derive an efficient algorithm to solve the proposed new model and provide its convergence guarantee as well. Extensive experiments on various occluded data sets indicate that our proposed model can achieve more promising performance than other related methods.

*To whom all correspondence should be addressed. This work was partially supported by U.S. NIH R01 AG049371, NSF IIS 1302675, IIS 1344152, DBI 1356628, IIS 1619308, IIS 1633753.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220016

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**;

KEYWORDS

Robust Metric Learning, Maximum Correntropy Criterion.

ACM Reference Format:
Jie Xu, Lei Luo, Cheng Deng, and Heng Huang. 2018. New Robust Metric Learning Model Using Maximum Correntropy Criterion. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220016

1 INTRODUCTION

In recent data mining and machine learning research, metric learning models have been successfully applied to address many image or video data analysis applications that heavily rely on distances or similarities, *e.g.*, face verification [9, 14, 32], image/video classification [20, 21, 29], person re-identification [27], *etc.* Metric learning methods mainly focus on learning an optimal distance (such as Mahalanobis distance) matrix $\mathbf{M}$ that captures the important relationships among data for a given task, *i.e.*, assigning smaller distances for similar data samples and larger distances for dissimilar ones.

To capture the relationships between data samples as precise as possible, various kinds of metric learning algorithms have been proposed in literature, *e.g.*, large-margin nearest neighbors (LMNN) [30], information-theoretic metric learning (ITML) [5], logistic discriminant metric learning (LDML) [7], *etc.* Although these metric learning methods were successful in solving many problems, they cannot handle the noisy data well, which has become the bottleneck for them to achieve good performance in real-world applications. To address this challenging issue, different aspects of efforts have been made to improve the metric learning models, such as: designing effective choices for model parameters; deriving the proper regularizations [9, 14] to fit distance matrix $\mathbf{M}$; using the structure of data in the learning process, *e.g.*, pairs [22], triplets [28] or quadruplets [13] models. Although notable improvements have been achieved by these methods, they ignore the effect of real-world malicious occlusions or corruptions on the model performance.

Images and videos captured in real-world conditions include large variations in shape and occlusions, *e.g.*, human faces may be partially occluded due to the use of accessories such as hats and sunglasses, and scenery images may be corrupted because of rain drops and heavy fog. To address these occlusion or corruption problems, plenty of strategies have been given in previous works. The most popular way to deal with the data occlusions is to locally analyze the occluded test images by partially matching them with unoccluded training images [10]. However, this method only operates on the non-occluded part, which overlooks occluded regions of an image. In the other related work, Wright *et al.* [31] proposed a sparse representation classifier (SRC) to eliminate the occlusion in the face image. Although some good results in coping with sparse noise have been reported, the SRC algorithm is still sensitive to contiguous occlusion. To overcome this shortcoming, He *et al.* [8] learned a robust sparse representation based on the correntropy [16] along with the use of an l_1-norm penalty, and obtained encouraging results.

In this paper, we introduce the maximum correntropy criterion into the metric learning model to deal with real-world malicious occlusions or corruptions. Unlike conventional methods, we consider the metric learning model as a reconstruction problem, and enforce the intra-class reconstruction residual of each sample to be smaller than the inter-class reconstruction residual by a large margin. Since the distance matrix $\mathbf{M}$ is positive semi-definite, it can be factorized as $\mathbf{M} = \mathbf{L}^\top\mathbf{L}$. Then the reconstruction residual can be viewed as implementing a linear transformation with projection $\mathbf{L}$. To effectively cope with the noise caused by occlusions or corruptions, we utilize correntropy induced metric to characterize this projected reconstruction residual. Therefore, our model not only can handle the occlusion problem in real data but also absorbs the advantages of conventional metric learning methods, *i.e.*, the effectiveness in the image alignment. In addition, an efficient optimization algorithm is derived to solve the proposed new model with the proof of its convergence. Extensive experiments on various occluded data sets indicate that the proposed model can achieve more promising performance than the other related methods.

Notations: For matrices $\mathbf{A}$ and $\mathbf{B}$, denote the Frobenius inner product by $\langle \mathbf{A}, \mathbf{B}\rangle = Tr(\mathbf{A}^\top\mathbf{B})$, where '$Tr$' denotes the trace of a matrix. For a given vector $\mathbf{a} = (a_1, a_2, ..., a_d)^\top$, $diag(\mathbf{a}) = \mathbf{A}$ corresponds to a squared diagonal matrix such that $\forall i, A_{i,i} = a_i$. $\mathbf{e}$ represents a unit vector, and $\mathbf{I}$ is a unit matrix. Finally, for $x \in \mathbb{R}$, let $[x]_+ = \max(0, x)$.

2 MAXIMUM CORRENTROPY CRITERION FOR RECONSTRUCTION-BASED LMNN

In this section, we first briefly revisit the Large Margin Nearest Neighbor (LMNN) algorithm and then describe the reconstruction-based LMNN [18]. After that, we introduce the maximum correntropy criterion into the reconstruction-based LMNN model and propose our new robust metric learning model.

2.1 Large Margin Nearest Neighbor

Most metric learning methods aim to learn a Mahalanobis distance matrix $\mathbf{M}$, under which samples from the same class are forced to be close to each other and those from different classes are pushed far away. Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{m\times n}$ be the training set, where $\mathbf{x}_i = (x_{i1}, x_{i2}, ...x_{im})^\top \in \mathbb{R}^m, i = 1, 2, ..., n$, m is the feature dimension, and n is the number of training sample, then the Mahalanobis distance between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as:

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

Large Margin Nearest Neighbor (LMNN) [30], as one of the most widely used metric learning methods, uses triplet constraints on training examples. If we denote the similar 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. ξ_{ijk} is a safety margin distance for each triplet. Let y_i denotes the label of $\mathbf{x}_i$, then we have:

$$\begin{aligned}\mathcal{S} = \{&(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \mathbf{x}_j \text{ belongs}\\ &\text{to the } k\text{-neighborhood of } \mathbf{x}_i\},\\ \mathcal{R} = \{&(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}, y_i \neq y_k\}.\end{aligned} \tag{4}$$

LMNN has been proved to be very effective in learning a good Mahalanobis distance for practical problems. Meanwhile, it can be easily integrated into many other approaches such as deep learning [17], multi-task learning [24] and transferring learning [6]. However, LMNN cannot use the geometric structure of samples well for a joint classification task, since the learned distance metric only characterizes the point-to-point distance. Also, it is sensitive to Euclidean distance when it computes neighbors of each sample at the beginning.

2.2 Reconstruction-Based LMNN

To overcome the limitation that LMNN ignores the relationship of samples in the training set, Lu *et al.* proposed to learn distance metrics under the sparse representation-based classification framework [18].

For each training example $\mathbf{x}_i = (x_{i1}, x_{i2}, ...x_{im})^\top \in \mathbb{R}^m, i = 1, 2, ..., n$, the remaining training examples are used to construct two sample sets $\mathbf{A}^i$ and $\mathbf{B}^i$, where $\mathbf{A}^i = [a_{ij}]^i \in \mathbb{R}^{m\times n_1}$ and $\mathbf{B}^i = [b_{ik}]^i \in \mathbb{R}^{m\times n_2}$ denote the intra-class and inter-class sample sets of $\mathbf{x}_i$, respectively. The intra-class and inter-class sparse reconstruction squared Mahalanobis distances ($d^2_{1,\mathbf{M}}$ and $d^2_{2,\mathbf{M}}$) under the distance matrix $\mathbf{M}$ are defined as follows:

$$\begin{aligned}\min_{\beta^i} d^2_{1,\mathbf{M}}(\mathbf{x}_i, \mathbf{A}^i) &= (\mathbf{x}_i - \mathbf{A}^i\beta^i)^\top\mathbf{M}(\mathbf{x}_i - \mathbf{A}^i\beta^i) + \lambda||\beta^i||_1\,,\\ \min_{\gamma^i} d^2_{2,\mathbf{M}}(\mathbf{x}_i, \mathbf{B}^i) &= (\mathbf{x}_i - \mathbf{B}^i\gamma^i)^\top\mathbf{M}(\mathbf{x}_i - \mathbf{B}^i\gamma^i) + \lambda||\gamma^i||_1\,,\end{aligned} \tag{5}$$

where $\beta^i \in \mathbb{R}^{n_1\times 1}$ and $\gamma^i \in \mathbb{R}^{n_2\times 1}$ are the reconstruction coefficients of the intra-class and inter-class samples of $\mathbf{x}_i$, respectively. They can be obtained by the l_1-norm minimization when $\mathbf{M}$ is fixed [33].

In the training procedure, the intra-class reconstruction residual over all training samples is expected to be as small as possible. Moreover, similar to the idea of LMNN, the inter-class reconstruction residual is expected to be larger than the intra-class reconstruction residual with a margin. Hence, the objective function of Reconstruction based Metric Learning (RML) problem is formulated as:

$$\begin{aligned} \min_{\mathbf{M}\in\mathbb{S}_+^d,\beta,\gamma} & \sum_i d^2_{1,\mathbf{M}}(\mathbf{x}_i,\mathbf{A}^i) + \mu\sum_i \xi_i + \nu\sum_i (||\beta^i||_1 + ||\gamma^i||_1) \\ s.t. \quad & d^2_{2,\mathbf{M}}(\mathbf{x}_i,\mathbf{B}^i) - d^2_{1,\mathbf{M}}(\mathbf{x}_i,\mathbf{A}^i) \geq 1-\xi_i, \\ & \xi_i \geq 0, \mathbf{M}\succeq 0\,. \end{aligned} \tag{6}$$

Unlike conventional metric learning methods, RML takes advantage of reconstruction residual information to measure the intra-class and inter-class variations and then learns the distance matrix. Although this method not only considers the label information of training samples but also the relationship between them, it is still not robust to deal with the malicious occlusions or corruptions in real-world data.

2.3 Maximum Correntropy Criterion for Reconstruction-Based LMNN

Images captured in real-world conditions include large variations in shape and occlusions, *e.g.*, human faces may be partially occluded due to the use of accessories such as hats and sunglasses, scenery images may be corrupted because of rain drops and heavy fog. The errors caused by such occlusions could be arbitrarily large in magnitude, and thus cannot be characterized by the simple L_2-norm.

Recently, the maximum correntropy criterion (MCC) has been proposed to process non-Gaussian noise and has achieved good results. It is formulated as:

$$\max_{\theta} \frac{1}{m}\sum_{j=1}^{m} g(e_j)\,, \tag{7}$$

where $g(x) = \exp(-\frac{x^2}{2\sigma^2})$ is a Gaussian kernel function, e_j is the error and θ is the parameter in the criterion to be specified. MCC adaptation is applicable in any noise environment when its distribution has the maximum at the origin [16]. We hope to use it to correlate each sample $\mathbf{x}_i$ to its intra-class sample sets $\mathbf{A}^i$ and inter-class sets $\mathbf{B}^i$.

Because $\mathbf{M} = \mathbf{L}^\top\mathbf{L}, \mathbf{L}\in\mathbb{R}^{m\times m}$, we have:

$$\begin{aligned} (\mathbf{x}_i - \mathbf{A}^i\beta^i)^\top\mathbf{M}(\mathbf{x}_i - \mathbf{A}^i\beta^i) &= ||\mathbf{L}(\mathbf{x}_i - \mathbf{A}^i\beta^i))||_2^2\,, \\ (\mathbf{x}_i - \mathbf{B}^i\gamma^i)^\top\mathbf{M}(\mathbf{x}_i - \mathbf{B}^i\gamma^i) &= ||\mathbf{L}(\mathbf{x}_i - \mathbf{B}^i\gamma^i))||_2^2\,. \end{aligned} \tag{8}$$

For the sake of convenience in writing, let $\mathbf{Y} = \mathbf{L}\mathbf{X} = [y_{ij}] \in \mathbb{R}^{m\times n}$, $\mathbf{P}^i = \mathbf{L}\mathbf{A}^i = [p_{ij}]^i \in \mathbb{R}^{m\times n_1}$, $\mathbf{Q}^i = \mathbf{L}\mathbf{B}^i = [q_{ik}]^i \in \mathbb{R}^{m\times n_2}$. Thus, the intra-class and inter-class sparse reconstruction (squared) Mahalanobis distances Eq. (5) become the following correntropy-based sparse model:

$$\begin{aligned} \max_{\beta^i} \Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) &= \sum_{j=1}^{m} g(y_{ij} - \sum_{k=1}^{n_1} p^i_{kj}\beta^i_k) - \lambda||\beta^i||_1, \\ \max_{\gamma^i} \Im_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i) &= \sum_{j=1}^{m} g(y_{ij} - \sum_{k=1}^{n_2} q^i_{kj}\gamma^i_k) - \lambda||\gamma^i||_1. \end{aligned} \tag{9}$$

Based on correntropy, the individual pixels of the representation are treated differently. The pixels belonging to the same class as testing sample $\mathbf{y}_i$ will be given larger weights. In other words, if there are occlusions and corruptions in a testing sample $\mathbf{y}_i$, the pixels corresponding to outliers will have small contributions to the correntropy. Thus, the noise can be handled uniformly within the correntropy framework [8].

To improve the interpretability of the above correntropy-based sparse model (9), we impose nonnegative constraints on the reconstruction coefficients similar as [8]. Thus, we have:

$$\begin{aligned} \max_{\beta^i} \Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) &= \sum_{j=1}^{m} g(y_{ij} - \sum_{k=1}^{n_1} p^i_{kj}\beta^i_k) - \lambda\sum_{k=1}^{n_1}\beta^i_k, \\ & s.t. \quad \beta^i_k \geq 0. \end{aligned} \tag{10}$$

$$\begin{aligned} \max_{\gamma^i} \Im_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i) &= \sum_{j=1}^{m} g(y_{ij} - \sum_{k=1}^{n_2} q^i_{kj}\gamma^i_k) - \lambda\sum_{k=1}^{n_2}\gamma^i_k, \\ & s.t. \quad \gamma^i_k \geq 0. \end{aligned} \tag{11}$$

Based on the above discussions, we propose the following maximum correntropy criterion based metric learning framework:

$$\begin{aligned} \max_{\mathbf{L},\beta,\gamma} \quad & \sum_i \Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) - \mu\sum_i \xi^i - \nu\sum_i (\sum_{k=1}^{n_1}\beta^i_k + \sum_{k=1}^{n_2}\gamma^i_k) \\ s.t. \quad & \Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) - \Im_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i) \geq 1-\xi^i, \xi^i \geq 0, \\ & \beta^i_k \geq 0, \quad \gamma^i_k \geq 0. \end{aligned} \tag{12}$$

With our new model defined in Eq. (12), we can learn a good distance matrix $\mathbf{M}$ even with occlusion data. Our model not only absorbs the benefits of maximum correntropy criterion but also inherits the advantages of traditional metric learning, *e.g.*, the effectiveness in image alignment. In the following section, we will derive an efficient optimization algorithm to solve our proposed new model (12).

3 OPTIMIZATION ALGORITHM

In this section, we first develop an efficient algorithm to optimize problem (12), and then provide its convergence guarantee.

3.1 Algorithm Derivation

We use an alternating optimization approach. We first fix $\mathbf{L}$, update $\beta^i, i = 1, ..., n$ and $\gamma^i, i = 1, ..., n$, and then fix β^i and γ^i, update $\mathbf{L}$, iteratively. Since the solution method of β^i is same as that of γ^i, we only take β^i as an example.

Step 1: We initialize $\mathbf{L}$ as the unit matrix, and obtain β and γ by solving (10) and (11), respectively. Due to the nonlinear property of the objective function in Eq. (10), it is difficult to directly optimize it. Inspired by [8], we use half-quadratic technique [35] and expectation maximization (EM) method [34] to solve this optimization problem. Depending on the property of convex conjugate function [2], we have the follows.

Proposition 1. *There exists a convex conjugate function φ of $g(x)$ such that*

$$g(x) = \max_{h'}\left(h'\frac{||x||^2}{\sigma^2} - \varphi(h')\right), \tag{13}$$

and for a fixed x, the maximum is reached at $h' = -g(x)$ [35].

Substituting (13) into (10), we get:

$$\max_{\beta^i,\mathbf{h}} \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) = \sum_{j=1}^{m}\left(h_j\left(y_{ij}-\sum_{k=1}^{n_1}p_{kj}^i\beta_k^i\right)^2-\varphi(h_j)\right)-\lambda\sum_{k=1}^{n_1}\beta_k^i, \quad s.t. \quad \beta_k^i \geq 0\,, \tag{14}$$

where $\mathbf{h}=[h_1,...,h_m]^\top$ are the auxiliary variables introduced by half-quadratic optimization. Based on Proposition 1, for a fixed β^i, the following equation holds

$$\Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i) = \max_{\mathbf{h}} \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i,\mathbf{h}). \tag{15}$$

It follows that

$$\max_{\beta^i} \Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i) = \max_{\beta^i,\mathbf{h}} \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i,\mathbf{h}). \tag{16}$$

We can draw a conclusion that maximizing $\Im_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i)$ is same as maximizing the augmented function $\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i,\mathbf{h})$. For the sake of convenience in writing, we will abbreviate $\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i,\beta^i,\mathbf{h})$ to $\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i)$ from now on.

Clearly, a local maximizer $(\beta^i,\mathbf{h})$ can be calculated in an alternating way:

$$\begin{aligned}
&\{h_j\}^{t+1} = -g(y_{ij}-\sum_{k=1}^{n_1}p_{kj}^i\{\beta_k^i\}^t),\\
&\{\beta^i\}^{t+1} = arg\max_{\beta^i}(\mathbf{y}_i-\mathbf{P}^i\beta^i)^\top diag(\mathbf{h})(\mathbf{y}_i-\mathbf{P}^i\beta^i)-\lambda\sum_{k=1}^{n_1}\beta_k^i,\\
&\quad s.t. \quad \beta_k^i \geq 0.
\end{aligned} \tag{17}$$

Obviously, if we regard the auxiliary variables $-\mathbf{h}$ as weights in (17), the optimization problem (17) is a weighted linear least squares problem with nonnegative constraint.

The optimal problem in (17) can be reformulated as the following quadratic programming problem:

$$\min_{\beta^i}\left(\frac{\lambda}{2}-\tilde{\mathbf{P}}^{i\top}\tilde{\mathbf{y}}_i\right)^\top\beta^i+\frac{1}{2}\beta^{i\top}\tilde{\mathbf{P}}^{i\top}\tilde{\mathbf{P}}^i\beta^i, \quad s.t. \quad \beta_k^i\geq 0, \tag{18}$$

where $\tilde{\mathbf{P}}^i = diag(\sqrt{-\mathbf{h}^{t+1}})\mathbf{P}^i$ and $\tilde{\mathbf{y}}_i = diag(\sqrt{\mathbf{h}^{t+1}})\mathbf{y}_i$. Since $\tilde{\mathbf{P}}^{i\top}\tilde{\mathbf{P}}^i$ is a positive semidefinite matrix, this quadratic programming problem in (18) is convex. According to the Karush-Kuhn-Tucker optimal conditions, we can derive the following monotone linear complementary problem (LCP) [26]:

$$\alpha^i = \tilde{\mathbf{P}}^{i\top}\tilde{\mathbf{P}}^i\beta^i-\tilde{\mathbf{P}}^{i\top}\tilde{\mathbf{y}}_i+\frac{\lambda}{2}, \alpha^i\geq 0, \beta^i\geq 0, \beta^{i\top}\alpha^i = 0. \tag{19}$$

If the matrix $\tilde{\mathbf{P}}^i$ has full column rank ($rank(\tilde{\mathbf{P}}^i)=n_1$), the convex program (18) and the LCP (19) have unique solutions for each vector $\tilde{\mathbf{y}}_i$ [26].

We define F and G as two subsets of $\{1,...,n_1\}$ such that $F\cup G=\{1,...,n_1\}$ and $F\cap G=\varnothing$. And let F and G be the working set and inactive set in the active set algorithm, respectively. Consider the following column partition of the matrix $\tilde{\mathbf{P}}^i$

$$\tilde{\mathbf{P}}^i = [\tilde{\mathbf{P}}_F^i,\tilde{\mathbf{P}}_G^i], \tag{20}$$

where $\tilde{\mathbf{P}}_F^i\in\mathbb{R}^{m\times|F|}$, $\tilde{\mathbf{P}}_G^i\in\mathbb{R}^{m\times|G|}$, and $|F|$, $|G|$ are the numbers of F and G, respectively. We can rewrite (19) as:

$$\begin{bmatrix}\alpha_F^i\\ \alpha_G^i\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{P}}_F^{i\top}\tilde{\mathbf{P}}_F^i & \tilde{\mathbf{P}}_F^{i\top}\tilde{\mathbf{P}}_G^i\\ \tilde{\mathbf{P}}_G^{i\top}\tilde{\mathbf{P}}_F^i & \tilde{\mathbf{P}}_G^{i\top}\tilde{\mathbf{P}}_G^i\end{bmatrix}\begin{bmatrix}\beta_F^i\\ \beta_G^i\end{bmatrix}-\begin{bmatrix}\tilde{\mathbf{P}}_F^{i\top}\tilde{\mathbf{y}}_i\\ \tilde{\mathbf{P}}_G^{i\top}\tilde{\mathbf{y}}_i\end{bmatrix}+\frac{\lambda}{2}, \tag{21}$$

where $\beta_F^i,\alpha_F^i\in\mathbb{R}^{|F|}$, $\beta_G^i,\alpha_G^i\in\mathbb{R}^{|G|}$, $\beta^i=(\beta_F^i,\beta_G^i)$, and $\alpha^i=(\alpha_F^i,\alpha_G^i)$. After that, we can compute the values of variables β_F^i and β_G^i by the following iterative procedure [26]:

$$\begin{aligned}
&\min_{\beta_F^i\in\mathbb{R}^{|F|}} ||\tilde{\mathbf{P}}_F^i\beta_F^i-\tilde{\mathbf{y}}_i||_2^2+\lambda\sum_{k\in F}\beta_k^i,\\
&\alpha_G^i = \tilde{\mathbf{P}}_G^{i\top}(\tilde{\mathbf{P}}_F^i\beta_F^i-\tilde{\mathbf{y}}_i)+\frac{\lambda}{2}.
\end{aligned} \tag{22}$$

The optimal solution is given by $\beta^i=(\beta_F^i,0)$ and $\alpha^i=(0,\alpha_G^i)$. Solving β^i and γ^i are two independent problems, and we use same strategy to get $\gamma^i, i=1,...,n$.

Step 2: Having obtained $\beta^i, i=1,2,...,n$ and $\gamma^i, i=1,2,...,n$, problem (12) can be rewritten as:

$$\begin{aligned}
&\max_{\mathbf{L}} \quad \sum_i\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i)-\mu\sum_i\xi^i\\
&s.t. \quad \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i)-\hat{\Im}_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i)\geq 1-\xi^i, \xi^i\geq 0.
\end{aligned} \tag{23}$$

For the sake of convenience in writing, Eq. (23) can be further reformulated as:

$$\min_{\mathbf{L}} -\sum_i\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i)+\mu[1+\hat{\Im}_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i)-\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i)]_+. \tag{24}$$

Notably, we can re-write the intra-class and inter-class sparse reconstruction (squared) Mahalanobis distances as:

$$\begin{aligned}
&\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i,\mathbf{A}^i) = (\mathbf{L}(\mathbf{x}_i-\mathbf{A}^i\beta^i))^\top diag(\mathbf{h}_A)\mathbf{L}(\mathbf{x}_i-\mathbf{A}^i\beta^i)-\lambda\sum_{k=1}^{n_1}\beta_k^i,\\
&\hat{\Im}_{2,\mathbf{L}}(\mathbf{x}_i,\mathbf{B}^i) = (\mathbf{L}(\mathbf{x}_i-\mathbf{B}^i\gamma^i))^\top diag(\mathbf{h}_B)\mathbf{L}(\mathbf{x}_i-\mathbf{B}^i\gamma^i)-\lambda\sum_{k=1}^{n_2}\gamma_k^i,
\end{aligned} \tag{25}$$

where $\mathbf{h}_A$ and $\mathbf{h}_B$ are the auxiliary variables corresponding to the calculation of $\beta^i, i=1,...,n$ and $\gamma^i, i=1,...,n$, respectively. In each iteration, $\mathbf{L}$ is updated by performing a subgradient descent, *i.e.*,

$$\mathbf{L}\Leftarrow\mathbf{L}-\eta\nabla\mathbf{L}, \tag{26}$$

where the subgradient of problem (24) with respect to $\mathbf{L}$ is:

$$\begin{aligned}
\nabla\mathbf{L} = 2\sum_i &diag(-\mathbf{h}_A)\mathbf{L}(\mathbf{x}_i-\mathbf{A}^i\beta^i)(\mathbf{x}_i-\mathbf{A}^i\beta^i)^\top\\
&+\mu\sum_i(diag(-\mathbf{h}_B)\mathbf{L}(\mathbf{x}_i-\mathbf{B}^i\gamma^i)(\mathbf{x}_i-\mathbf{B}^i\gamma^i)^\top\\
&-diag(-\mathbf{h}_A)\mathbf{L}(\mathbf{x}_i-\mathbf{A}^i\beta^i)(\mathbf{x}_i-\mathbf{A}^i\beta^i)^\top).
\end{aligned} \tag{27}$$

The specific procedures are summarized in Alg. 1.

Just like any other kernel methods, the choice of kernel size will influence the performance of the proposed model, and kernel size is often determined empirically. In this paper, we compute the kernel size (bandwidth) by:

$$\sigma^2 = \frac{\theta}{2m}(\mathbf{P}_F^i\beta_F^i-\mathbf{y}_i)^T(\mathbf{P}_F^i\beta_F^i-\mathbf{y}_i)\,, \tag{28}$$

Algorithm 1 Algorithm to solve Eq. (12)

Input: $\mathbf{X} \in \mathbb{R}^{m \times n}$
Output: $\mathbf{L} \in \mathbb{R}^{m \times m}$
Initialization: $\mathbf{L} = \mathbf{I}_m$
repeat
  for all $\mathbf{x}_i \in \mathbf{X}$ **do**
    Initialization: $\mathbf{h}_A = -\mathbf{e}$, $\mathbf{h}_B = -\mathbf{e}$
    repeat
      Calculate β^i and γ^i using Eqs. (18)-(22);
      Update $\mathbf{h}_A$ and $\mathbf{h}_B$ using Eq. (17);
      Update kernel size σ according to Eq. (28);
    until Converge
  repeat
    Calculate $\nabla_{\mathbf{L}}$ using Eq. (27);
    Update $\mathbf{L} \Leftarrow (\mathbf{L} - \eta \nabla_{\mathbf{L}})$;
  until Converge
until Converge

where θ is a constant to control the noise. We set θ to 1 throughout the paper.

3.2 Convergence Analysis

Theorem 1. *The Alg. 1 will monotonically increase the objective of the problem in Eq. (12) in each iteration and converge to the local optimum solution to the problem.*

Proof. According to Eq. (17) and Proposition 1, we have

$$\begin{aligned}\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \{\beta^i\}^t, \{\mathbf{h}\}^t) \leq & \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \{\beta^i\}^t, \{\mathbf{h}\}^{t+1}) \\ \leq & \hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \{\beta^i\}^{t+1}, \{\mathbf{h}\}^{t+1}).\end{aligned} \tag{29}$$

The cost function increases at each alternating maximization step. Therefore, the sequence $\{\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \{\beta^i\}^t, \{\mathbf{h}\}^t), t = 1, 2, ...\}$ is nondecreasing. We can verify that $\Im_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \beta^i)$ is bounded (property of correntropy [16]) and, by Eq. (16), $\hat{\Im}_{1,\mathbf{L}}(\mathbf{x}_i, \mathbf{A}^i, \{\beta^i\}^t, \{\mathbf{h}\}^t)$ is also bounded. Consequently, **step 1** converges.

After obtained the optimal $\beta^i, i = 1, 2, ..., n$ and $\gamma^i, i = 1, 2, ..., n$, we can use gradient descent method to optimizing function (24). By choosing the suitable step size η, we know that the convergence of **step 2** can be guaranteed by [1].

In Alg. 1, we optimize problem (12) by alternatively updating $\{\beta^i, i = 1, 2, ..., n, \gamma^i, i = 1, 2, ..., n\}$ and $\mathbf{L}$. Thus, the objective function (12) is monotonically increasing for each iteration. Considering the boundedness of correntropy, the objective function (12) is also bounded. Thus, the proposed Alg. 1 finally converges to a local optimum point. □

4 EXPERIMENTAL RESULTS

In this section, we propose to evaluate our metric learning method on different databases, including real-world malicious occlusion datasets, contiguous occlusion and corruption datasets, and kinship verification dataset. There are two main goals in our experiment: first, we will show that our model is more robust to be applied to solve real-world occlusion problems; second, our model is able to outperform the state-of-the-art metric learning methods.

4.1 Real-World Malicious Occlusion

We applied our model to real face recognition scenarios against malicious occlusion, *i.e.*, NUST Robust Face database (NUST-RF) [4]. In our experiment, we focus on face recognition task, namely identifying a person from a digital image.

- **NUST Robust Face database** - NUST Robust Face database (NUST-RF) is mainly designed for robust face recognition under various occlusions [4]. Except occlusion, it also includes variations of illumination, expression and pose. We use a subset face images of NUST-RF database, and there are 50 subjects captured in two environments (indoor and outdoor). We manually cropped the face portion of the image and then normalized it to 80×60 pixels. Fig. 1 shows an example of several selected images of one subject.

We extracted LOMO features for each image [15], which not only achieve some invariance to viewpoint changes, but also capture local region characteristics of a person. PCA is further applied to reduce the feature dimension to 30 dimension.

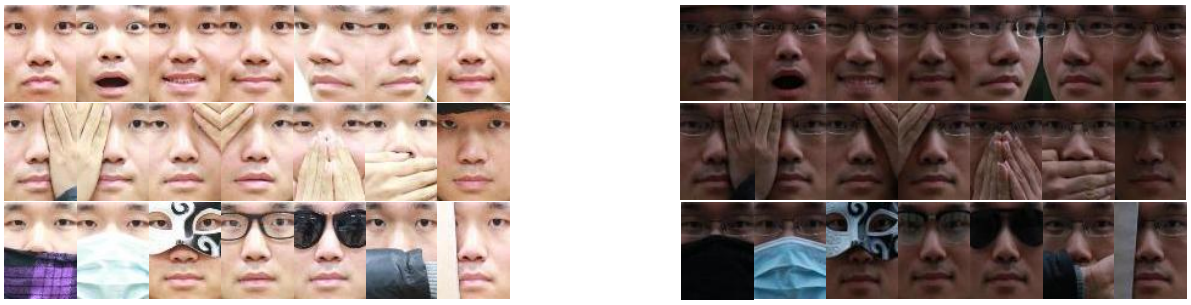
Setting: In the experiment, we use 5-fold cross validation and compute average accuracy and standard deviation for each method as final performance. All the regularization parameters are tuned from range $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. For CAP and FANTOPE methods, parameter rank k of matrix $\mathbf{M}$ is tuned from $[10 : 5 : 30]$.

Compared methods:

- KNN: We use k-nearest neighbor method ($k = 1$) as classifier and compute Euclidean distance to measure the similarity between any two images. This method works as a baseline.
- ITML: Metric learning method proposed in [5]. They use LogDet divergence as regularization so that they do not need do explicit positive semi-definite projection.
- LMNN: It is one of the most widely-used Mahalanobis distance metric learning methods [30]. In this method, they use labeled information to generate triplet constraints.
- FANTOPE: Based on LMNN, this method utilizes a fantope regularization which minimizes sum of k smallest singular values of distance matrix $\mathbf{M}$ [14].
- CAP: Based on LMNN, it uses a capped trace norm to penalize the singular values of distance matrix $\mathbf{M}$ that are less than a threshold adaptively learned in the optimization [9].
- RML: It uses a reconstruction criterion to learn the discriminative distance metric [18].
- Proposed: The maximum correntropy criterion is introduced into the metric learning model to deal with occlusion data.

Results: Table 1 shows the recognition performance of different methods on NUST-RF database of two environments. Obviously our method consistently outperforms other competing methods in both indoor and outdoor cases. As shown in Fig. 1, if occlusions exist, it is unlikely that the test image will be very close to any single training image of the same class, so that the KNN classifier performs poorly. Although LMNN, CAP and FANTOPE can improve the recognition rates compared to KNN, their improvements are limited. RML also performs poorly because it is based on the MSE criterion which is sensitive to outliers.

Fig. 2 shows the recognition accuracy using different regularization parameter μ on NUST-RF database in two environments. From

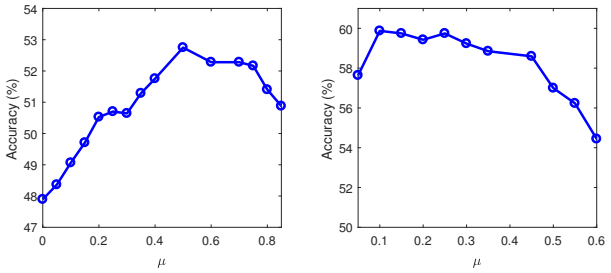


(a) Indoor (b) Outdoor

Figure 1: Cropped images of one subject captured in two environments in NUST-RF database, *i.e.*, (a) indoor, and (b) outdoor.

Table 1: Recognition accuracy (%) and standard deviation of different methods on NUST-RF database in two environments.

	KNN	ITML	LMNN	FANTOPE	CAP	RML	Proposed
Indoor	36.14 ± 2.70	48.88 ± 0.48	36.20 ± 3.30	41.87 ± 2.50	41.70 ± 2.86	35.56 ± 3.04	**52.75 ± 1.79**
Outdoor	45.24 ± 1.51	59.04 ± 1.14	46.01 ± 2.06	58.72 ± 1.33	58.34 ± 1.33	42.81 ± 2.16	**59.74 ± 2.86**



(a) indoor (b) outdoor

Figure 2: The recognition accuracy using different regularization parameter μ on NUST-RF database in two environments, *i.e.*, (a) indoor, and (b) outdoor.

the figure we can see that our method outperforms other compared methods except ITML even with inappropriate regularization parameter μ. This illustrates the superiority of the proposed model on real-world malicious occlusion data.

A well-tuned kernel size σ is much more effective in eliminating the effect of outliers and noise because it controls all robust properties of correntropy. In this paper, we set the Gaussian kernel size as a single function defined in Eq. (28). We plot the Gaussian function with respect to θ in Fig. 3a. When the value of θ is small, the outliers will receive smaller weights. When the value of θ is large, the outliers will receive larger weights. Fig. 3b and Fig. 3c demonstrate how the kernel size affects the accuracy in two environments. They both seem to reach the maximum around $\theta = 1$. That's mainly because they have similar corrupted pixels.

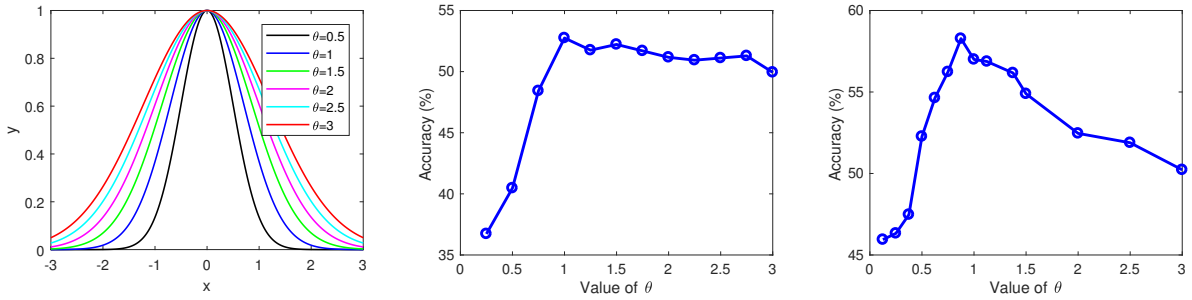
4.2 Sparse Noise and Contiguous Occlusion

In this section, we did three groups of occlusion experiments associated with three datasets to validate the robustness of the proposed algorithm.

- **Traffic video database** - Traffic video database consists of 254 video sequences of highway traffic in Seattle [3], which contains a variety of traffic patterns and weather conditions, like rain drops, overcast and *et al* [3]. 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 [3]. For each video clip, we subtract the mean image and then use mean gray values of all the frames as feature representation.
- **OSR dataset** - Outdoor Scene Recognition (OSR) dataset is from [25], and there are 2688 images from 8 scene categories. We extracted gist features as representation [23].
- **PubFig database** - We use a subset face images of PubFig database, and there are 771 images from 8 face categories [12]. Similarly with NUST-RF database, we extracted LOMO features as representation.

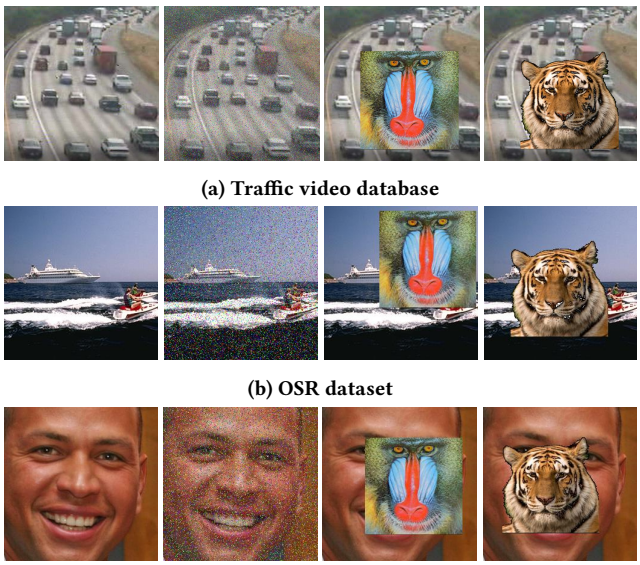
Setting: We simulated various types of contiguous occlusion by adding sparse noise to both training and testing data or by replacing a randomly selected local region in each image with an unrelated square block of the "baboon" image for regular occlusion and a randomly located "tiger" image for irregular occlusion. The size of the added image is 60% of the size of the original image. Fig. 4 shows a clean image and its noisy versions from three datasets. Since the differences between the pixels of the unrelated "baboon" image or "tiger" image and the pixels of the images from three datasets are relatively small, the contiguous occlusion caused by these unrelated images is much more challenging than by random black or white dots.

As a benchmark for comparison, for traffic video database, we use the pre-specified training/testing split, which is generated for



(a) Gaussian function $y = \exp(-x^2/\theta)$ (b) Indoor (c) Outdoor

Figure 3: Recognition performance versus θ in Gaussian kernel size σ. (a) Gaussian function $y = \exp(-x^2/\theta)$ with respect to θ. (b) In indoor environment, the average accuracy under various values of θ on NUST-RF database. (c) In outdoor environment, the average accuracy under various values of θ on NUST-RF database.



(a) Traffic video database

(b) OSR dataset

(c) PubFig dataset

Figure 4: Example pairs of images from three datasets, *i.e.*, (a)traffic video database, (b) OSR dataset, (c) PubFig dataset.

4-fold cross validation [3]. For PubFig datasets, we use the same experiment setup as [14]. Each person contributes 30 images as training data to learn Mahalanobis metric matrix $\mathbf{M}$ and builds classifier, other images are used as testing data to evaluate classification performance. Each time, we select training data randomly and repeat this procedure 5 times. For OSR dataset, we use 5-fold cross validation. For all datasets, PCA is further applied to reduce the feature dimension to 30 dimension. We compute average accuracy and standard deviation for each method as the final performance. All the regularization parameters are tuned from range $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. For CAP and FANTOPE methods, parameter rank k of matrix $\mathbf{M}$ is tuned from $[10 : 5 : 30]$.

Compared methods: We use the same compared methods as we used in last subsection, namely KNN, ITML [5], LMNN [30], FANTOPE [14], CAP [9], RML [18].

Results: Table 2, Table 3 and Table 4 show the classification accuracy and the standard derivation of different methods on three datasets, *i.e.*, traffic database, OSR dataset and PubFig dataset. It is obviously our method consistently outperforms other competing methods in all cases. It is interesting that ITML achieves second best results on NUST-RF database and PubFig dataset as shown in Table 1 and Table 4, but gets poor results on traffic video database and OSR dataset as shown in Table 2 and Table 3. The main reason is that we extract LOMO features as representation for NUST-RF database and PubFig dataset, which exactly works for ITML method. For this classification task, both FANTOPE and CAP methods are based on LMNN method. Since they all have similar results, which indicates the low-rank regularization for Mahalanobis distance metric learning is not particularly effective in this case. Especially for regular occlusion that replace a randomly selected local region with "baboon" image and irregular occlusion that replace local region with "tiger" image, LMNN, FANTOPE and CAP achieve almost the same result.

4.3 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 [19]. We use KinFaceW-I dataset without adding extra sparse noise or contiguous occlusion. Some example pairs from KinFaceW-I dataset are shown in Fig. 5.

- **KinFaceW-I dataset** - KinFaceW-I dataset consists of four representative types of kin relations: Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-D) and Mother-Son (M-S), respectively. In the KinFaceW-I dataset, there are 134, 156, 127 and 116 pairs of kinship images for these four relations.

Setting: As a benchmark for comparison, we use the pre-specified training/testing split, which is generated for 5-fold cross validation [19]. We use the given Histogram of Oriented Gradients (HOG) from image blocks as feature representation. PCA is further employed to reduce dimensionality of each vector to 100 dimension.

Table 2: Recognition accuracy (%) and standard deviation of different methods on traffic video database, where Sparse noise, regular and irregular occlusion are added.

	KNN	ITML	LMNN	FANTOPE	CAP	RML	Proposed
Original	83.48 ± 3.87	88.09 ± 2.19	89.37 ± 2.38	90.16 ± 2.69	90.16 ± 2.69	85.05 ± 5.15	**91.34 ± 1.52**
Sparse	68.47 ± 10.22	72.54 ± 2.24	74.08 ± 1.68	77.51 ± 3.80	75.59 ± 4.93	71.66 ± 3.28	**81.49 ± 3.28**
Regular	54.75 ± 4.52	57.09 ± 1.54	61.83 ± 4.49	64.91 ± 1.99	63.01 ± 4.56	55.13 ± 3.68	**67.32 ± 3.49**
Irregular	58.67 ± 5.07	66.44 ± 3.22	69.32 ± 5.15	69.32 ± 5.15	69.32 ± 5.15	60.25 ± 2.70	**73.28 ± 8.01**

Table 3: Recognition accuracy (%) and standard deviation of different methods on OSR dataset, where Sparse noise, regular and irregular occlusion are added.

	KNN	ITML	LMNN	FANTOPE	CAP	RML	Proposed
Original	69.01 ± 1.96	69.09 ± 1.18	74.41 ± 1.20	74.97 ± 0.88	74.45 ± 1.19	61.34 ± 1.62	**75.46 ± 2.15**
Sparse	61.83 ± 1.75	60.93 ± 0.83	66.67 ± 1.70	66.70 ± 1.68	66.67 ± 1.70	56.57 ± 2.60	**68.67 ± 1.98**
Regular	55.34 ± 2.72	56.02 ± 1.20	58.66 ±1.31	58.73 ± 1.43	58.70 ± 1.27	54.38 ± 3.26	**60.48 ± 3.42**
Irregular	52.25 ± 1.74	53.45 ± 2.75	57.02 ± 1.80	57.10 ± 1.74	57.13 ± 1.68	50.45 ± 2.17	**63.99 ± 2.38**

Table 4: Recognition accuracy (%) and standard deviation of different methods on Pubfig dataset, where Sparse noise, regular and irregular occlusion are added.

	KNN	ITML	LMNN	FANTOPE	CAP	RML	Proposed
Original	56.73 ± 1.12	62.04 ± 2.19	61.65 ± 1.63	61.69 ± 1.60	61.80 ± 1.72	55.86 ± 1.54	**63.99 ± 1.60**
Sparse	48.46 ± 1.35	52.52 ± 0.80	51.35 ± 1.55	51.39 ± 1.69	51.39 ± 1.99	48.23 ± 1.26	**54.59 ± 1.45**
Regular	35.30 ± 1.14	40.60 ± 1.42	37.48 ± 1.64	37.71 ± 1.56	38.05 ± 1.57	35.80 ± 1.59	**44.77 ± 1.09**
Irregular	40.94 ± 2.30	44.27 ± 1.73	41.73 ± 3.39	41.88 ± 2.78	42.33 ± 2.35	40.23 ± 2.48	**46.09 ± 2.43**

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

Figure 5: Example pairs of images from the KinFaceW-I database [19].

To measure the facial kinship verification accuracy for all these compared methods, we report a Receiver Operator Characteristic (ROC) curve. Since it is difficult to compare the performance just by the curve, we compute Equal Error Rate (*EER*) [11] of the respective method, and use $1 - EER$ as evaluation criterion, and the method with the lowest *EER*, or the highest $1 - EER$ is the most accurate one.

All the regularization parameters are tuned from range $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$. For CAP and FANTOPE methods, parameter rank k of matrix $\mathbf{M}$ is tuned from $[10 : 5 : 30]$.

Compared methods:

- IDENTITY: We compute Euclidean distance directly as a baseline.
- MAHAL: Traditional Mahalanobis distance between images in a pair is computed, where the metric matrix is inverse of covariance between two vectors.
- KISSME: A metric learning methods based on a statistical inference perspective. It learns a distance metric from equivalence constraints and can be used in large scale dataset [11].
- LDML: It uses logistic discriminant to learn a metric from a set of labeled image pairs [7].
- FANTOPE: It introduces a fantope regularization which minimizes sum of k smallest singular values of distance matrix [14].
- CAP: It introduces a capped trace norm to penalize the singular values of distance matrix $\mathbf{M}$ that are less than a threshold adaptively learned in the optimization [9].

Results: We plot ROC curve for each method in Fig. 6. As shown in the figure, the curves are not very smooth because the number of sample pairs is very limited. Since the curves are mixed together, it is difficult to distinguish which method is better. The proposed model performs relatively well at the beginning of the curve. After

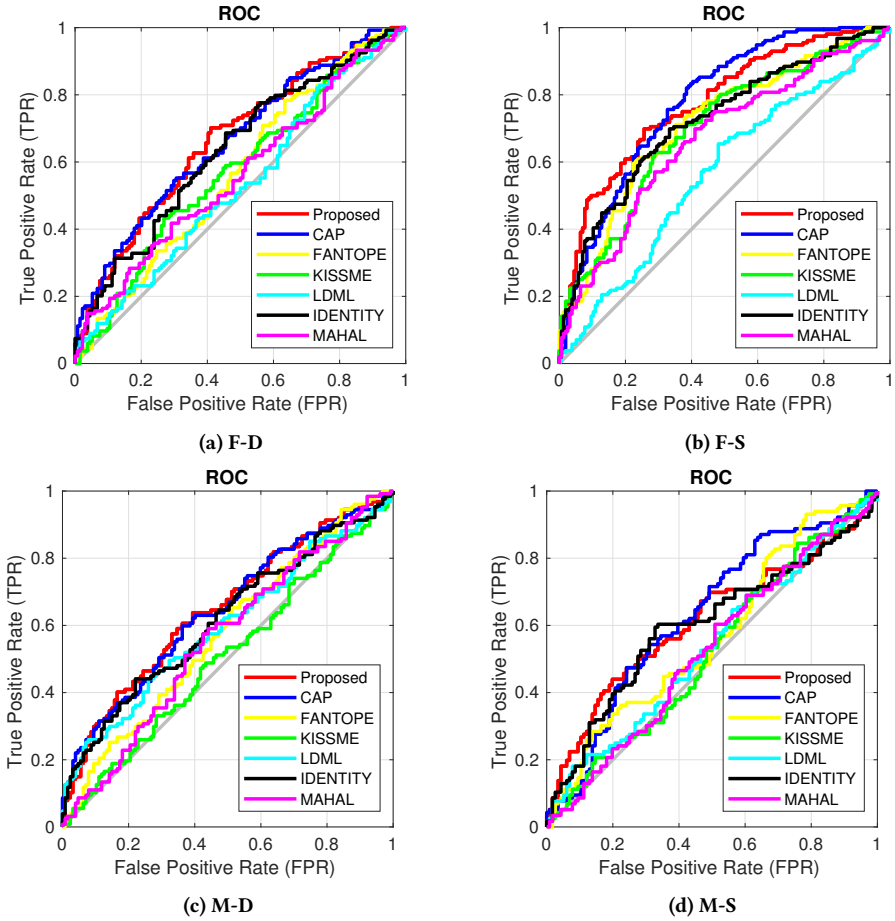


Figure 6: Facial kinship verification results on all kinship relations of KinFaceW-I dataset: (a) Father-Daughter kinship relation; (b) Father-Son kinship relation; (c) Mother-Daughter kinship relation, (d) Mother-Son kinship relation.

Table 5: The $1 - EER$ of different methods on the KinFaceW-I dataset.

Method	F-D	F-S	M-D	M-S
Euclidean	0.604	0.673	0.575	0.603
Mahalanobis	0.522	0.635	0.575	0.526
KISSME [11]	0.560	0.654	0.512	0.517
LDML [7]	0.507	0.571	0.559	0.526
Cap [9]	0.612	0.699	0.606	**0.603**
Fantope [14]	0.537	0.660	0.551	0.509
Proposed	**0.627**	**0.705**	**0.606**	0.586

that, CAP method seems to have gotten better results. Because our method prefers to solve occlusion or corruption data, there is no obvious advantage in this task. Even so, the proposed model also achieves promising results. To further check the performance difference between these methods, we compute $1 - EER$ as evaluation criterion, as shown in Table 5. Our method gets more promising performance than the other related methods in most cases.

5 CONCLUSION

To deal with the large variations in shape and occlusions presented by real-world data, in this paper, we propose a new robust metric learning model by utilizing the maximum correntropy criterion. We consider the metric learning model as a reconstruction problem, and enforce the intra-class reconstruction residual of each sample to be smaller than the inter-class reconstruction residual by a large margin. To improve the robustness of model, we use correntropy induced metric to characterize this projected reconstruction residual. We derive an efficient optimization algorithm to solve the proposed new model. Experimental results demonstrate the superior performance of our proposed method.

REFERENCES

[1] Stephen Boyd and Almir Mutapcic. 2006. Subgradient methods. *Lecture notes of EE364b, Stanford University, Winter Quarter* 2007 (2006).
[2] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization.* Cambridge university press.
[3] Antoni B Chan and Nuno Vasconcelos. 2005. Probabilistic kernels for the classification of auto-regressive visual processes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 846–851.

[4] Shuo Chen, Jian Yang, Lei Luo, Yang Wei, Kaihua Zhang, and Ying Tai. 2017. Low-Rank Latent Pattern Approximation With Applications to Robust Image Classification. *IEEE Transactions on Image Processing* 26, 11 (2017), 5519–5530.

[5] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*. ACM, 209–216.

[6] Zhengming Ding and Yun Fu. 2017. Robust Transfer Metric Learning for Image Classification. *IEEE Transactions on Image Processing* 26, 2 (2017), 660–670.

[7] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2009. Is that you? Metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 498–505.

[8] Ran He, Wei-Shi Zheng, and Bao-Gang Hu. 2011. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 8 (2011), 1561–1576.

[9] Zhouyuan Huo, Feiping Nie, and Heng Huang. 2016. Robust and Effective Metric Learning Using Capped Trace Norm: Metric Learning via Capped Trace Norm. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1605–1614.

[10] Hongjun Jia and Aleix M Martinez. 2009. Support vector machines in face recognition with occlusions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 136–141.

[11] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2288–2295.

[12] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 365–372.

[13] Marc T Law, Nicolas Thome, and Matthieu Cord. 2013. Quadruplet-wise image similarity learning. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 249–256.

[14] Marc T Law, Nicolas Thome, and Matthieu Cord. 2014. Fantope regularization in metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1051–1058.

[15] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.

[16] Weifeng Liu, Puskal P Pokharel, and José C Príncipe. 2007. Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Transactions on Signal Processing* 55, 11 (2007), 5286–5298.

[17] Jiwen Lu, Junlin Hu, and Yap-Peng Tan. 2017. Discriminative Deep Metric Learning for Face and Kinship Verification. *IEEE Transactions on Image Processing* 26, 9 (2017), 4269–4282.

[18] Jiwen Lu, Gang Wang, Weihong Deng, and Kui Jia. 2015. Reconstruction-based metric learning for unconstrained face verification. *IEEE Transactions on Information Forensics and Security* 10, 1 (2015), 79–89.

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

[20] Lei Luo and Heng Huang. 2018. Matrix Variate Gaussian Mixture Distribution Steered Robust Metric Learning. *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)* (2018), in press.

[21] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. 2012. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. *Computer Vision–ECCV 2012* (2012), 488–501.

[22] Alexis Mignon and Frédéric Jurie. 2012. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2666–2672.

[23] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42, 3 (2001), 145–175.

[24] Shibin Parameswaran and Kilian Q Weinberger. 2010. Large margin multi-task metric learning. In *Advances in neural information processing systems*. 1867–1875.

[25] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 503–510.

[26] Luís F Portugal, Joaquim J Judice, and Luís N Vicente. 1994. A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables. *Math. Comp.* 63, 208 (1994), 625–643.

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

[28] Matthew Schultz and Thorsten Joachims. 2004. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*. 41–48.

[29] Hua Wang, Feiping Nie, and Heng Huang. 2014. Robust Distance Metric Learning via Simultaneous L1-Norm Minimization and Maximization. *The 31st International Conference on Machine Learning (ICML 2014)* (2014), 1836–1844.

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

[31] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2009. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 31, 2 (2009), 210–227.

[32] Jie Xu, Lei Luo, and Heng Huang. 2018. Multi-Level Metric Learning via Smoothed Wasserstein Distance. *27th International Joint Conference on Artificial Intelligence (IJCAI 2018)* (2018), in press.

[33] Allen Y Yang, Zihan Zhou, Arvind Ganesh Balasubramanian, S Shankar Sastry, and Yi Ma. 2013. Fast l_1-Minimization Algorithms for Robust Face Recognition. *IEEE Transactions on Image Processing* 22, 8 (2013), 3234–3246.

[34] Shuang-Hong Yang, Hongyuan Zha, S Kevin Zhou, and Bao-Gang Hu. 2009. Variational graph embedding for globally and locally consistent feature extraction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 538–553.

[35] Xiao-Tong Yuan and Bao-Gang Hu. 2009. Robust feature extraction via information theoretic learning. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 1193–1200.