IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2019, accepted May 11, 2019, date of publication May 22, 2019, date of current version June 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918149

Intrinsic Metric Learning With Subspace Representation

LIPENG CAI[1], SHIHUI YING[1], (Member, IEEE), YAXIN PENG[1], (Member, IEEE), CHANGZHOU HE[2], AND SHAOYI DU[3], (Member, IEEE)

[1]Department of Mathematics, Shanghai University, Shanghai 200444, China
[2]Qualcomm (Shanghai) Co. Ltd., Shanghai 201210, China
[3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Yaxin Peng (yaxin.peng@shu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700800, in part by the National Natural Science Foundation of China under Grant 11771276, Grant 61573274, and Grant 61731009 and in part by the Capacity Construction Project of Local Universities in Shanghai under Grant 18010500600.

ABSTRACT The accuracy of classification and retrieval significantly depends on the metric used to compute the similarity between samples. For preserving the geometric structure, the symmetric positive definite (SPD) manifold is introduced into the metric learning problem. However, the SPD constraint is too strict to describe the real data distribution. In this paper, we extend the intrinsic metric learning problem to semi-definite case, by which the data distribution is better described for various classification tasks. First, we formulate the metric learning as a minimization problem to the SPD manifold on subspace, which not only considers to balance the information between inner classes and inter classes by an adaptive tradeoff parameter but also improves the robustness by the low-rank subspaces presentation. Thus, it benefits to design a structure-preserving algorithm on subspace by using the geodesic structure of the SPD subspace. To solve this model, we develop an iterative strategy to update the intrinsic metric and the subspace structure, respectively. Finally, we compare our proposed method with ten state-of-the-art methods on four data sets. The numerical results validate that our method can significantly improve the description of the data distribution, and hence, the performance of the image classification task.

INDEX TERMS Metric learning, subspace representation, low-rank optimization, structure preserving, image classification.

I. INTRODUCTION

Metric, as a measure defined on a data set, plays a crucial role in the description of data distribution. Deferent metrics can offer different views of data. How to learn a suitable metric to well describe the data distribution and further improve the separability of the data becomes a fundamental issue in machine learning [1], [2].

Metric learning aims to find a proper metric for a given collection of pairs with similar/dissimilar samples [3]. It is widely used in classification [4]–[6], image and 3D object retrieval [7]–[9], face recognition [10]–[13], and person re-identification tasks [14], [15], etc. From the geometric viewpoint, metric learning can be divided into two categories: linear and nonlinear metric learning. In linear metric learning the metric is formulated by a globally linear mapping, while in nonlinear metric learning it is defined locally. The nonlinear metric learning is often based on the linear metric learning, while the local linearization method [16] and the kernel based method [17], [18] are two mainly used techniques in nonlinear metric learning. Therefore, the linear metric learning is the cornerstone and will improve the performance of the nonlinear metric case.

In past decades, several linear metric learning studies have been developed. These include the Mahalanobis distance metric for clustering [19], the metric matrix parameterization method to learn a weighting diagonal matrix [20], the Neighborhood Component Analysis (NCA) method in a probabilistic framework [21], and the Large Margin Nearest Neighbors (LMNN) method [22]. As in Support Vector Machines (SVMs), LMNN proposes a margin criterion based on hinge loss. Then, Der and Saul adopt the alternating iterative method to solve it more efficiently [23].

The associate editor coordinating the review of this manuscript and approving it for publication was Bilal Alatas.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

On the other hand, Shen et al. propose a dual method to improve the computational efficiency [24]. Kunapuli and Shavlik propose the mirror descent metric learning method for online Mahalanobis distance learning [25]. Recently, Nguyen and Baets solve the metric learning model by using the difference of convex functions programming [28].

Focusing on the energy function, Davis et al. develop the Information-Theoretic Metric Learning (ITML) method by introducing an information measure [26]. ITML is important because it introduces the LogDet divergence regularization used in several other Mahalanobis distance learning methods [27]. Also, Nguyen et al. develop the Distance Metric Learning through Maximization of the Jeffrey divergence (DMLMJ) between two distributions derived from local pairwise constraints [28]. Later, Gu et al. compare the Schatten norm and the vector norm in metric learning [29], and Zhang et al. design an FLS-SVM-ML algorithm by combining the fuzzy least squares SVM and metric learning [30].

From the statistic viewpoint, Li et al. propose a maximum margin criterion, in which an insightful rule for distribution of data and a feature extraction method are established [31]. Liu et al. further discuss this criterion [32]. Later, Li et al. develop a distributed approach to discriminative distance metric learning [33]. Recently, Li and Chen propose a non-parametric metric learning approach based on Gaussian process (GP-Metric) and extend the bilinear similarity into a non-parametric form [34].

Although the above-mentioned metric learning approaches have shown excellent performance, they do not consider the geometric structure of the set of all metrics, thus less improvement in the accuracy and efficiency is achieved. To solve these issue, Zadeh et al. revisit the metric learning from the viewpoint of geometric mean on the manifold of all positive definite matrices [35]. In this work, the best linear metric is learned on the manifold of positive definite matrices, which ensures that the metric in each iteration is positive definite. Reference [36] proposes an intrinsic structure-preserving semi-supervised approach (ISSML) for the linear metric learning, where a parameter is introduced into the objective function to fit the data distribution between inner and inter classes.

Nevertheless, the distribution of real data may not be fully filled in the whole feature space. That is, features may be located on a subspace of the feature space. On the other hand, noise and outliers may influence the learned metric. Therefore, it is important to design a more robust method for linear metric learning with noise and outliers.

Fortunately, many effective approaches, especially the subspace based methods, have been proposed to reduce the dimension of the feature space, as well as to improve the robustness of learning algorithms [37]–[41]. They attempt to find the underlying low dimension structure from the high dimensional data, which realizes the effective description and reduces the computation complexity. One fundamental method is Principal Component Analysis (PCA), which seeks the subspace by maximizing the variance of projected samples [38]. It offers a common technique for dimension reduction and feature representation. To improve the robustness of PCA, Candès et al. propose the Robust PCA (RPCA) method, which decomposes the data into low-rank background and sparse noise parts, and hence greatly promotes the robustness of data recovery [42]. Qiao et al. develop an explicit nonlinear mapping for manifold learning [43]. Further, He et al. propose a Local Preserving Projection (LPP) to find subspaces by preserving local structure of samples [44]. Later, Zhang develops the Linear Discriminant Analysis (LDA) by considering the inner class and inter classes information [39]. Li and Fu well realize the subspace discovery by introducing the low-rank constraints [40], [41]. They also apply this approach to balanced and unbalanced graphs learning [45]. Recently, Ding and Fu improve it to multi-view data analysis by collective low-rank subspace learning [46].

The subspace based methods well represent the features for learning issues, but they do not fully consider the metric factor in data representation. On the other hand, although the LDA considers the inner class and inter classes information with the basic idea to maximize the quotient of inter class matrix and inner class matrix after projecting samples to low dimensional space, the dimension of the projected space is fixed. Recently, Bhutani et al. propose a low-rank variant Low Rank Geometric Mean Metric Learning (LR-GMML) method of the GMML [35] based on the subspace methods [56], which makes GMML scalable in a high dimensional data sets. However, this work does not pay enough attention to the noise distribution. Therefore, this paper will synthesize the advantages of intrinsic metric learning, subspace representation and noise sparsity to form a more robust metric learning method. Our contributions are twofold below.

1) Model: We extend our SPD metric learning model [36] to the subspace SPD metric learning for better fitting the real data distribution, which not only considers to balance the information between inner class and inter classes by an adaptive tradeoff parameter, but also improves the robustness by the low rank subspaces representation and sparsity of noise.

2) Algorithm: We propose an alternate structure preserving algorithm. The algorithm benefits much from the geodesic structure of positive-definite matrix group, which can transfer an SPD constrained optimization problem to an unconstrained problem on an SPD sub-manifold.

The rest of this paper is organized as follows. In Section II, after recalling the traditional metric learning model, we introduce the low rank constraints to the model to represent the optimal subspace for supervised metric learning. Then, to solve the model, we propose an alternately iterative strategy in Section III, where a structure preserving algorithm is designed by using the manifold structure of positive definite matrix group. In Section IV, we demonstrate the effectiveness of our method, as well as compare it with ten state-of-the-art methods for classification on four real data sets, including ORL and Extended YaleB facial data sets, COIL-100 object

data set [54], and USPS digit data set [55]. Finally, this paper is concluded in Section V.

II. THE PROPOSED MODEL

A. METRIC LEARNING MODEL ON MANIFOLD

Given a data set $X = \{x_1, \ldots, x_n\}$ with the label set $Y = \{y_1, \ldots, y_n\}$, where $x_i \in \mathbb{R}^d$, $y_i \in \mathcal{L} = \{c_1, \ldots, c_K\}$, d is the dimension of sample, n is the number of samples, and K is the number of class. Traditional supervised metric learning is to seek the best Mahalanobis distance metric M, such that samples are as close as possible in the same class and as far as possible in different classes.

Denote the Mahalanobis distance between x_i and x_j by

$$D_{ij}^2 = (x_i - x_j)^T M (x_i - x_j), \tag{1}$$

where M is a positive definite matrix (i.e. $M \succ 0$). Then, a triplet set $\Gamma := \{(x_i, x_j, x_k) : D_{ij}^2 < D_{ik}^2\}$ is used to distinguish the relationship between samples, where x_i and x_j are in the same class, i.e. $y_i = y_j$, and x_i, x_k are in different classes, i.e. $y_i \neq y_k$.

Then, the supervised linear metric learning problem [22] is modeled by

$$\min_{M \succ 0} \sum_{\Gamma} (D_{ij}^2 - D_{ik}^2). \tag{2}$$

To better balance the influences of the inner-class and inter-class data, we introduce an adaptive parameter $\gamma = 1/(1 + \bar{D}^{-1})$ into the objective function, where $\bar{D}$ is the mean distance of the data set $\mathcal{D}$ [36]. Then, the model is rewritten by

$$\min_{M \succ 0} \sum_{\Gamma} (D_{ij}^2 - \gamma D_{ik}^2). \tag{3}$$

To reduce the computational complexity, we only calculate the distances of k nearest neighbors in the same class for each sample. Therefore, the energy function is translated into

$$\min_{M \succ 0} \sum_{i,j,k=1}^{n} \eta_{ij}(1 - y_{ik})(D_{ij}^2 - \gamma D_{ik}^2), \tag{4}$$

where η_{ij} and y_{ij} are two indicative functions defined by

$$\eta_{ij} = \begin{cases} 1, & x_i \text{ and } x_j \text{ are neighbors in the same class} \\ 0, & \text{otherwise} \end{cases}$$

$$y_{ij} = \begin{cases} 1, & y_i = y_j, \\ 0, & y_i \neq y_j. \end{cases}$$

Let $Y = (y_{ij})_{n \times n}$, $H = (\eta_{ij})_{n \times n}$, e be a column vector with all elements 1, and E_i be a column vector with all elements 0 except the ith element 1. Then the energy function is further simplified by

$$\min_{M \succ 0} \sum_{i,j=1}^{n} C_{ij} D_{ij}^2, \tag{5}$$

where $C_{ij} = \left(E_i^T (ee^T - Y)e\right) \eta_{ij} - \gamma E_i^T H e (1 - y_{ij})$.

Further, we rewrite the objective function by matrix form and the model is translated to

$$\min_{M \succ 0} 2\mathrm{tr}(X(D_c - C)X^T M), \tag{6}$$

where $D_c = \mathrm{diag}(Ce)$ and $C = (C_{ij})_{n \times n}$. For more detail deduction, we refer to [36].

To enhance the generalization ability, inspired by [47] we introduce a locally topological structure of data to (6) in [36]. Although it improves the robustness, there are still some shortages: 1) The metric matrix M is always assumed to be positive definite, but the distribution of data may not be fully filled in the whole feature space in real cases, and 2) the computation of the distances of the nearest neighbors is time-consuming. Therefore, we should find a more efficient way to improve the robustness. Fortunately, subspace based methods have been proposed from another way to reduce the dimension of the feature space, as well as to improve the robustness of learning algorithms. They always assume that the samples may be located on a subspace of the feature space and offer a way to improve the robustness for feature representation. Therefore, we can improve the robustness of metric learning by a more robust feature representation.

B. METRIC LEARNING MODEL ON SUB-MANIFOLD

In subspace learning, one efficient way is the sparse low rank self-representation [40]. That is, all samples can almost be linearly represented by few samples in the data set $X = XZ + E$, where Z is an $n \times n$ coefficient matrix and E is the noise. From another viewpoint, samples used to represent the whole data set X span a subspace. Therefore, we reformulate the metric learning problem on subspace by

$$\begin{aligned} \min_{M \succ 0, Z, E} \ & \mathrm{tr}(XZ(D_c - C)Z^T X^T M) + \mu \cdot \mathrm{r}(Z) + \lambda \|E\|_{2,1} \\ \text{s.t.} \ & X = XZ + E \end{aligned}$$

where $\mathrm{r}(Z)$ is the rank of Z, μ and λ are two balance parameters, and $\|\cdot\|_{2,1}$ is the $l_{2,1}$ norm, which can model the group-wise regularity of noise [37], [48]. In fact, the second and third terms provide a way to settle the above issues by considering the metric learning on such subspace. That is, we can introduce the low rank self-representation for the data set X, and then consider the metric learning on the projected samples.

The rank of Z can be approximated by its nuclear norm. Then, the objective function can be rewritten as

$$\begin{aligned} \min_{M \succ 0, Z, E} \ & \mathrm{tr}(XZ(D_c - C)Z^T X^T M) + \mu \|Z\|_* + \lambda \|E\|_{2,1} \\ \text{s.t.} \ & X = XZ + E \end{aligned} \tag{7}$$

where $\|\cdot\|_*$ is the nuclear norm of the matrix. The first term represents the weighted difference of inner-class distances and inter-class distances on the subspace determined by Z when Z is fixed. Minimizing the second term assumes that the dimension of subspace is as low as possible, while the third term means that XZ well represents the data set X. Therefore, this model well balances the robustness and the description of data distribution.

III. ITERATIVE STRATEGY AND ALGORITHMS

In this section, we propose an iterative strategy to solve the model (7). The detail is described as follows.

There are two kinds of independent variables in (7): the metric $M \succ 0$ and the variables of low rank self representations Z and E. Hence a common solving strategy is the alternate iteration. Concretely, the model can be solved by iteratively minimizing two subproblems below.

S1) For current fixed Z and E, we update M by solving the following minimization problem.

$$\min_{M\succ 0} \operatorname{tr}(XZ(D_c - C)Z^T X^T M). \tag{8}$$

S2) For current fixed M, we update Z and E by solving the following minimization problem.

$$\begin{aligned} &\min_{Z,E} \operatorname{tr}(XZ(D_c - C)Z^T X^T M) + \mu\|Z\|_* + \lambda\|E\|_{2,1} \\ &\text{s.t.} \quad X = XZ + E \end{aligned} \tag{9}$$

For the subproblem S1), we can adopt the same intrinsic steep descent method on the positive definite matrix group $\mathcal{P}(d)$ in [36]. Only we need to modify is to replace the data set X in [36] by XZ^k. That is, the iterative format for solving M by the geodesic structure of $\mathcal{P}(d)$ can be defined by

$$M^{k+1} = [M^k]^{\frac{1}{2}} \exp(-\alpha G(M^k))[M^k]^{\frac{1}{2}}, \tag{10}$$

where α is the optimal step, and $G(M^k)$ is the gradient of the objective function (8) at M^k. It is calculated by

$$G(M^k) = [M^k]^{-\frac{1}{2}} XZ^k(D_c - C)Z^{k^T} X^T [M^k]^{-\frac{1}{2}}. \tag{11}$$

The detailed deduction can be found in [36] and we use the symmetry of $XZ^k(D_c - C)Z^{k^T} X^T$.

For solving the subproblem S2) easier, we introduce a slack variable V with respect to Z, and then the model (9) equals to

$$\begin{aligned} &\min_{V,Z,E} \operatorname{tr}(Z(D_c - C)Z^T X^T M^{k+1} X) + \mu\|V\|_* + \lambda\|E\|_{2,1} \\ &\text{s.t.} \quad X = XZ + E, Z = V \end{aligned} \tag{12}$$

Then, by using the augmented Lagrange method, the model is translated into

$$\begin{aligned} &\min_{V,Z,E} \operatorname{tr}(Z(D_c - C)Z^T X^T M^{k+1} X) + \mu\|V\|_* + \lambda\|E\|_{2,1} \\ &\quad + \operatorname{tr}(L_1^T(X - XZ - E)) + \operatorname{tr}(L_2^T(Z - V)) \\ &\quad + \frac{\sigma}{2}(\|X - XZ - E\|_F^2 + \|Z - V\|_F^2), \end{aligned} \tag{13}$$

where L_1 and L_2 are two Lagrange multipliers, $\sigma > 0$ is a penalty parameter, and $\|\cdot\|_F$ is the Frobenius norm.

Using the same strategy, the minimization problem (13) can be solved by iterating the following three steps.

S2-1) For current fixed Z^k and E^k, we update the V^{k+1} by solving

$$\min_V \mu\|V\|_* + \operatorname{tr}(L_2^T(Z^k - V)) + \frac{\sigma}{2}\|Z^k - V\|_F^2. \tag{14}$$

To solve (14), we adopt the singular value thresholding (SVT) technique [49]. That is,

$$V^{k+1} = S_{\frac{\mu}{\sigma}}(Z^k + \frac{L_2}{\sigma}), \tag{15}$$

where $S_\tau(\cdot)$ is the shrinkage thresholding operator.

S2-2) For current fixed V^{k+1} and E^k, we update Z^{k+1} by solving

$$\begin{aligned} \min_Z \ &\operatorname{tr}(Z(D_c - C)Z^T X^T M^{k+1} X) \\ &+ \operatorname{tr}(L_1^T(X - XZ - E^k)) + \operatorname{tr}(L_2^T(Z - V^{k+1})) \\ &+ \frac{\sigma}{2}(\|X - XZ - E^k\|_F^2 + \|Z - V^{k+1}\|_F^2). \end{aligned} \tag{16}$$

From the KKT condition, we have

$$\begin{aligned} 0 = \ &2X^T M^{k+1} XZ(D_c - C) - X^T L_1 + L_2 \\ &+ \sigma(-X^T X + X^T XZ + X^T E^k + Z - V^{k+1}). \end{aligned} \tag{17}$$

Then, we can solve Z^{k+1} by a fast algorithm in [50].

S2-3) For current fixed Z^{k+1}, we update E^{k+1} by solving

$$\begin{aligned} \min_E \ &\lambda\|E\|_{2,1} + \operatorname{tr}(L_1^T(X - XZ^{k+1} - E)) \\ &+ \frac{\sigma}{2}\|X - XZ^{k+1} - E\|_F^2. \end{aligned} \tag{18}$$

The minimization problem (18) can be solved by the method in [51]. Concretely,

$$E(:, i) = \begin{cases} \dfrac{\|Q_i\| - \alpha}{\|Q_i\|} Q_i, & \|Q_i\| > \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $Q = X - XZ^{k+1} + \frac{L_1}{\sigma}$, Q_i is the i-th column of Q, and $\alpha = \frac{\lambda}{\sigma}$.

Finally, it should be pointed out that the Lagrange multipliers L_1, L_2 and σ are updated by

$$L_1^{k+1} = L_1^k + \sigma^k(X - XZ^{k+1} - E^{k+1}), \tag{20}$$

$$L_2^{k+1} = L_2^k + \sigma^k(Z^{k+1} - V^{k+1}), \tag{21}$$

and

$$\sigma^{k+1} = \min(\rho\sigma^k, \sigma_{\max}). \tag{22}$$

Then, we obtain the algorithm for metric learning via the subspace representation and summarize it as Algorithm 1.

IV. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed robust metric learning algorithm, in this section, we compare it with ten state-of-the-art methods for classification on four real data sets, including ORL [52] and Extended YaleB [53] facial data sets, COIL-100 object data set [54], and USPS digit data set [55]. All programs are written in Matlab 2013a and run by PC with Intel(R) Core(TM) i7-7500U CPU and 32GB RAM.

Algorithm 1 Metric Learning Algorithm

Require: Data matrix X, parameters μ, λ, $\rho = 1.3$ and $\sigma_{\max} = 10^{10}$.

Ensure: M^*, Z^*, and E^*.

Initialize $M^0 = \mathrm{Id}$, $Z^0 = 0$, $E^0 = 0$, $\sigma^0 = 0.1 > 0$, $\varepsilon = 10^{-8}$.
while Error $> \varepsilon$ or $k <$ maxIteration **do**
  Calculate the gradient $G(M^k)$ by (11);
  Calculate the optimal stepsize α by inexact line search.
  $M^{k+1} = [M^k]^{\frac{1}{2}} \exp(-\alpha G(M^k))[M^k]^{\frac{1}{2}}$.
  Update V^{k+1} by (15) with Z^k fixed.
  Update Z^{k+1} by solving (17) with V^{k+1} and E^k fixed.
  Update E^{k+1} by (19) with Z^{k+1} fixed.
  Update the multipliers L_1^{k+1} and L_2^{k+1} by
    $L_1^{k+1} = L_1^k + \sigma^k(X - XZ^{k+1} - E^{k+1})$,
    $L_2^{k+1} = L_2^k + \sigma^k(Z^{k+1} - V^{k+1})$,
  Update $\sigma^{k+1} = \min(\rho\sigma^k, \sigma_{\max})$
  Calculate the Error^{k+1} by Eq. (13) at M^{k+1}, V^{k+1}, Z^{k+1} and E^{k+1}.
  Set $k \Leftarrow k + 1$.
end while
Output: $M^* = M^k$, $Z^* = Z^k$ and $E^* = E^k$.

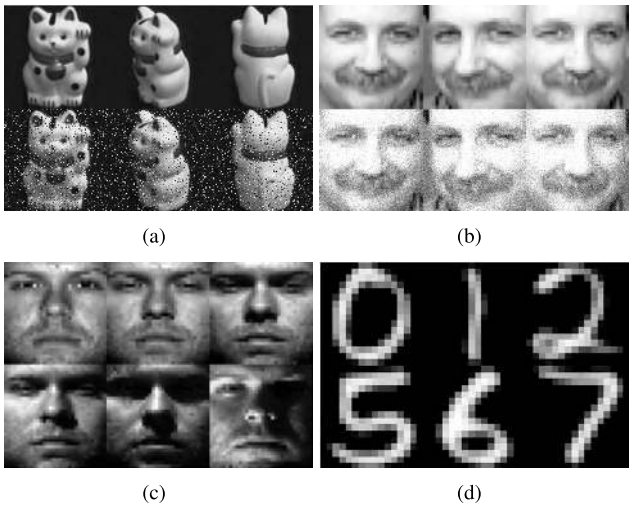


FIGURE 1. Representative samples from (a) COIL-100, (b) ORL, (c) extended YaleB, and (d) USPS.

A. DATASETS AND COMPARED METHODS

We compare our method with ten state-of-the-art methods on four widely used data sets. Fig. 1 shows some representative images in four data sets while Table 1 shows the attributes of the data sets.

Fig. 1(a) shows the representative samples of the COIL-100 object data set. This data set contains 100 objects. All images of one object were taken 5 degrees apart because the object is imaged on a rotated turntable, and hence each object has 72 images. The size of each image is 32×32 pixels with 256 grey levels per pixel. Thus, the original dimension of each image is 1024. The bottom low shows the image with 10% salt & pepper noise.

TABLE 1. Attributes of four data sets.

Data set	Type	Classes	Samples	Dimension
ORL	Face	40	400	1024
Extended YaleB	Face	38	2414	1024
COIL-100	Object	100	7200	1024
USPS	Digit	10	1000	256

Fig. 1(b) shows the representative samples of the ORL facial data set. This data set contains 400 images with 40 classes acquired under different lighting conditions. All images in data sets are cropped and resized to the size of 32×32. Thus, the original dimension of each image is 1024. The bottom row shows the image with 10% Gaussian noise.

Fig. 1(c) shows the representative samples of the Extended YaleB facial data set. This data set contains a total of 2414 images, including 64 frontal pose images of 38 different subjects. The variability between images of the same person is mainly due to different lighting conditions. The images are automatically centered by using optical flow and then converted to vectors. All images in data sets are cropped and resized to the size of 32×32. Thus, the original dimension of each image is 1024.

Fig. 1(d) shows the representative samples of the USPS data set. This data set is one of the standard data sets for handwritten digit recognition. It contains 9298 images with the digits from 0 to 9, which have been normalized to size of 16×16. Thus, the original dimension of each image is 256.

To validate the abilities of classification and dimension reduction, we compare our method with ten state-of-the-art metric learning methods (ITML [26], LMNN [22], DMLMJ [28], LR-GMML [56], LDA [39], LPP [44], LSDA [57] and SRRS [41]) and two baseline methods (Euclidean metric (EU) and PCA). For fair comparison, we only adopt the k-NN classifier under the learned metrics, because the ability of classification depends on the classifiers. Also, we use the recognition error rate to evaluate the performance of all compared methods. The recognition error rate is defined as

$$\text{The recognition error rate} = \frac{\sum_{i=1}^{n} f(x_i) \neq y_i}{n}.$$

where n is the number of samples, $f(x_i)$ is the prediction label of the sample, and y_i is the real label of the sample.

B. NUMERICAL RESULTS AND DISCUSSION

1) COIL-100 OBJECT DATA SET

For COIL-100 object data set, the first 20 classes are selected as the subset of samples. Then, we randomly select 10 images in each class as the training samples and the rest as the testing samples. We repeat this process 20 times. Further, to validate the robustness of algorithms, we randomly add the salt & pepper noise from 10% to 50% on the data, respectively. The bottom of Fig. 1(a) shows the images with 10% salt & pepper noise. Then, the numerical results are displayed in Table 2 and Figs. 2–5.

Also, we adopt the area under ROC curve (AUC) and $F1$-measure as the evaluation metric, we report AUCs and $F1$s

TABLE 2. The recognition error rates (Average ± Std) on COIL-100 data set with different percentages of salt & pepper noise.

Method	0%	10%	20%	30%	40%	50%	Average
EU	10.18±0.86(1024)	50.62±2.19(1024)	73.13±1.05(1024)	80.51±1.19(1024)	83.89±1.45(1024)	87.13±0.94(1024)	64.24
ITML	7.91±1.24(1024)	30.88±3.82(1024)	52.55±2.76(1024)	65.86±2.11(1024)	75.00±2.51(1024)	82.51±2.40(1024)	52.45
LMNN	8.58±0.98(1024)	30.56±2.42(1024)	40.53±2.90(1024)	51.66±2.78(1024)	63.69±2.84(1024)	75.25±2.84(1024)	45.05
DMLMJ	8.02±1.35(43)	25.12±1.84(1024)	35.70±0.96(1024)	47.25±1.80(1024)	58.44±1.22(1024)	69.86±1.77(1024)	40.73
LR-GMML	10.17±1.10(156)	24.08±1.62(18)	34.92±1.78(18)	46.60±2.23(18)	57.36±1.80(18)	69.77±2.77(18)	40.48
PCA	9.16±0.90(35)	20.23±2.20(11)	33.89±2.31(8)	53.66±2.64(7)	70.25±2.71(5)	82.91±1.98(4)	45.02
LDA	13.57±1.19(19)	28.82±1.93(19)	40.14±2.58(19)	51.52±2.02(19)	62.67±1.65(19)	72.42±1.74(19)	44.86
LPP	14.15±1.61(22)	28.77±1.19(21)	41.00±1.51(21)	52.68±1.53(22)	62.65±1.23(22)	72.76±1.60(20)	45.34
LSDA	14.19±1.43(30)	27.15±1.48(21)	38.46±1.91(21)	50.16±1.56(19)	60.67±1.39(21)	71.73±1.68(19)	43.73
SRRS	8.82±1.09(26)	12.31±1.09(31)	20.16±1.10(52)	35.13±2.02(10)	52.65±2.38(5)	66.17±2.59(4)	32.54
Ours	**7.55±1.10(25)**	**11.79±1.03(31)**	**18.05±1.05(37)**	**29.44±1.59(44)**	**45.35±1.18(54)**	**61.99±1.15(56)**	**29.03**

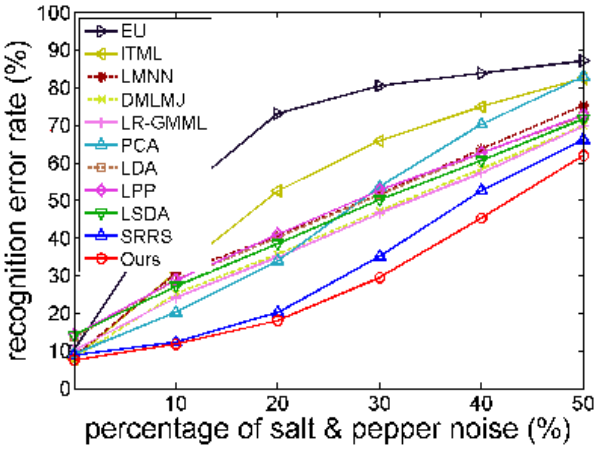


FIGURE 2. Averaged error rates of classification on COIL-100 data set with different percentages of salt & pepper noise.

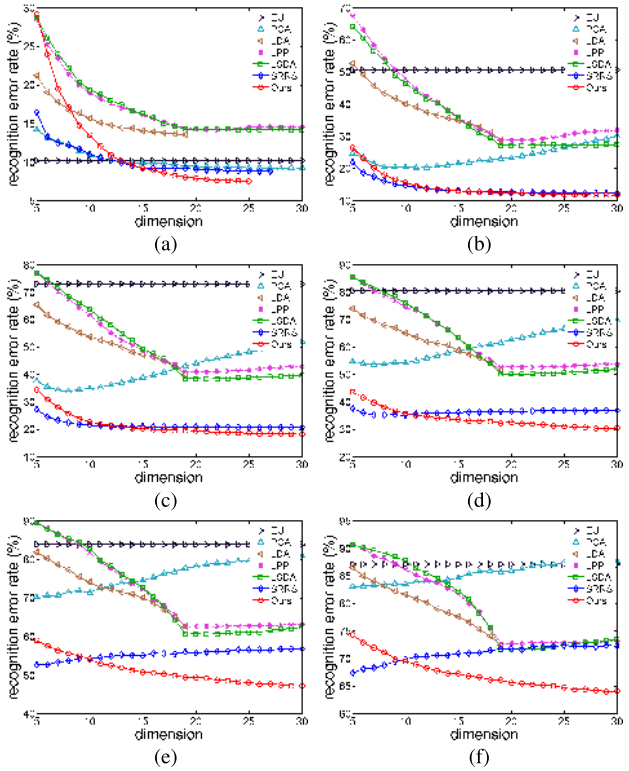


FIGURE 3. Averaged error rates of classification of seven methods on COIL-100 data set with different percentages of salt & pepper noise and dimensions of feature space. (a) 0%. (b) 10%. (c) 20%. (d) 30%. (e) 40%. (f) 50%.

in Table 3. For the multi-classification task, we divide it into C_K^2 (K is the number of class) binary-classification tasks to calculate the mean of AUCs and $F1$s. $F1$ is defined as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \times P \times R}{P + R}.$$

where P, TP, FN, R, FP and TN represent precision, the true positive, the false negative, recall, the false positive and the true negative, respectively.

Table 2 shows the recognition error ratios of classification task for all methods in the different noise levels. The number in the bracket is the dimension of reduced feature space. We have the following three observations. 1) All learned metrics improve the performance of classification. It is validated from the results of ITML and LMNN. Their performance is better than the EU-based method under the learned metric in the original feature space with dimension of 1024. 2) The dimension reduction also contributes to improve the classification performance. It is demonstrated from the comparison results between EU method and all dimension reduction based methods. 3) Our proposed method outperforms listed state-of-the-art methods under different salt & pepper noise in aspects of accuracy and robustness, as shown in Fig. 2.

Table 3 shows AUCs and $F1$s for all methods on COIL-100 data set. We can find that our method outperforms the other state-of-the-art methods under different evaluation metrics.

Fig. 3 shows the relationship between the recognition error ratios and dimension of the learned feature space. We have selected six methods to compare with our proposed method, since some methods are less efficient without dimension reduction. At the same time, we only display the EU curves to be baselines. Our proposed method almost obtains the best subspace representation for the samples.

To show the distributions under the corresponding metrics, we project all learned features to two dimensional spaces by PCA and display them in Fig. 4. Visually, the separability of data are all improved in different degrees under different learned metrics, where different colors represent

TABLE 3. AUCs and $F1$s for all methods on COIL-100 data set with different percentages of salt & pepper noise.

Method	0%		10%		20%		30%		40%		50%		Average	
	AUC	$F1$	AUC	$F1$	AUC	$F1$	AUC	$F1$	AUC	$F1$	AUC	$F1$	AUC	$F1$
EU	0.898	0.896	0.494	0.462	0.269	0.223	0.195	0.138	0.161	0.098	0.129	0.069	0.358	0.314
ITML	0.921	0.919	0.689	0.661	0.487	0.448	0.348	0.305	0.224	0.169	0.182	0.128	0.475	0.438
LMNN	0.914	0.913	0.692	0.677	0.598	0.585	0.488	0.471	0.364	0.341	0.239	0.207	0.549	0.532
DMLMJ	0.920	0.919	0.749	0.738	0.643	0.628	0.528	0.506	0.416	0.392	0.301	0.279	0.593	0.577
LR-GMML	0.899	0.897	0.760	0.752	0.644	0.634	0.534	0.521	0.424	0.408	0.306	0.290	0.595	0.584
PCA	0.908	0.907	0.798	0.790	0.661	0.648	0.463	0.439	0.298	0.284	0.171	0.155	0.550	0.537
LDA	0.864	0.860	0.712	0.695	0.599	0.578	0.485	0.460	0.373	0.349	0.276	0.258	0.551	0.533
LPP	0.859	0.854	0.712	0.690	0.594	0.565	0.474	0.440	0.372	0.336	0.274	0.249	0.548	0.522
LSDA	0.858	0.854	0.729	0.713	0.615	0.595	0.498	0.473	0.393	0.369	0.283	0.269	0.563	0.546
SRRS	0.912	0.910	0.877	0.875	0.798	0.793	0.649	0.640	0.474	0.462	0.338	0.327	0.645	0.668
Ours	**0.927**	**0.926**	**0.882**	**0.880**	**0.820**	**0.817**	**0.705**	**0.700**	**0.545**	**0.537**	**0.380**	**0.372**	**0.710**	**0.705**

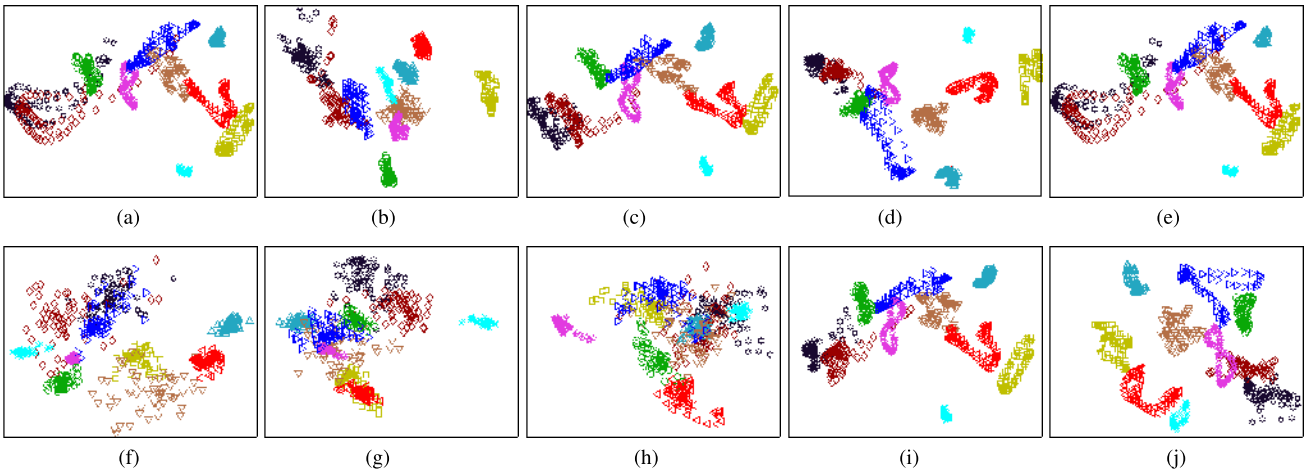


FIGURE 4. Data set visualization of ten methods on two dimensional feature spaces without noise on COIL-100 data set. Different colors represent different classes with ten classes. (a) EU. (b) ITML. (c) LMNN. (d) DMLMJ. (e) PCA. (f) LDA. (g) LPP. (h) LSDA. (i) SRRS. (j) Ours.

TABLE 4. The recognition error rates (Average ± Std) on ORL data set with different training samples.

Training samples	3	4	5	6	Average
EU	23.20±2.49(1024)	17.75±2.14(1024)	13.88±2.32(1024)	10.69±2.67(1024)	16.38
ITML	17.05±4.11(1024)	12.27±4.16(1024)	8.45±3.74(1024)	6.44±3.42(1024)	11.05
LMNN	14.98±1.68(1024)	12.00±3.71(1024)	8.95±2.75(1024)	6.91±2.12(1024)	10.71
DMLMJ	13.70±2.74(36)	9.31±2.87(44)	6.08±1.81(46)	4.50±2.11(50)	8.40
LR-GMML	12.20±1.71(36)	8.27±2.17(36)	**5.30±1.76(37)**	4.00±1.83(38)	7.44
PCA	23.20±2.56(119)	17.75±2.20(159)	13.82±2.37(198)	10.69±2.74(239)	16.37
LDA	15.64±2.27(39)	10.54±2.37(39)	7.35±1.87(39)	5.63±1.87(39)	9.79
LPP	13.39±1.82(39)	9.29±1.58(39)	6.75±1.69(39)	5.34±1.78(39)	8.69
LSDA	13.50±1.95(39)	8.96±2.05(49)	6.55±1.78(41)	5.47±1.75(42)	8.62
SRRS	12.69±1.77(119)	8.02±1.68(159)	6.38±1.36(60)	4.58±2.01(68)	7.92
Ours	**10.57±1.85(119)**	**6.58±2.24(159)**	5.33±1.69(199)	**3.47±1.27(239)**	**6.49**

different classes. The results of EU and PCA are the same because the final projections are the same. Further, the separability of data by our projection is the best.

Finally, to test the sensitivity of the model parameters, we validate the parameters in a large range. Concretely, we select μ from 10^2 to 10^5 and λ from 10^3 to 10^6. The results are shown in Fig. 5. The recognition error ratio is not sensitive to the model parameters and hence the model is robustness.

2) ORL AND EXTENDED YALEB FACIAL DATA SETS

To further validate the efficiency, we conduct the classification tasks on the ORL and Extended YaleB facial data sets. In experiments, we further consider the change of classification performance with the number of training samples on these two data sets.

First, we randomly select 3, 4, 5 and 6 images in each class as the training samples and the rest as testing samples. We repeat this process 20 times, and record the recognition error ratios with their standard variations by all methods, respectively. The results are displayed in Table 4 and Fig. 6.

From Table 4, we obtain the following facts. 1) All learned metrics improve the classification performance. It is validated from the results of ITML and LMNN. Their performance is better than the EU-based method under the learned metric in the original feature space. 2) The dimension reduction also contributes to improve the performance of classification. It is demonstrated from the comparison results between EU method and all dimension reduction based methods. Especially, compared with PCA, our proposed method has a large improvement even the dimensions of feature spaces are close.

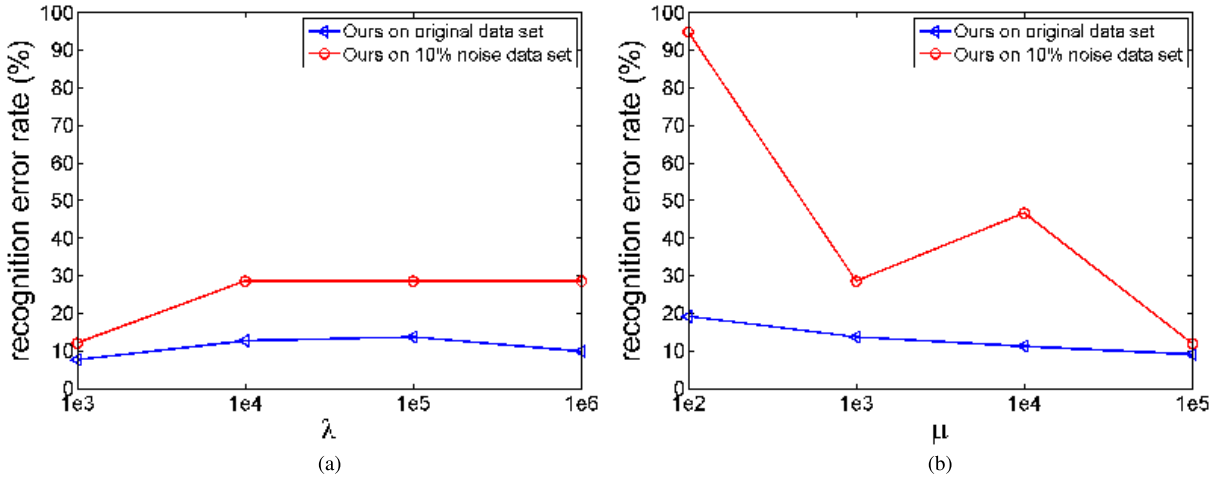


FIGURE 5. Averaged error rates for different model parameters on COIL-100 data set. (a) Sensitivity of parameter λ. (b) Sensitivity of parameter μ.

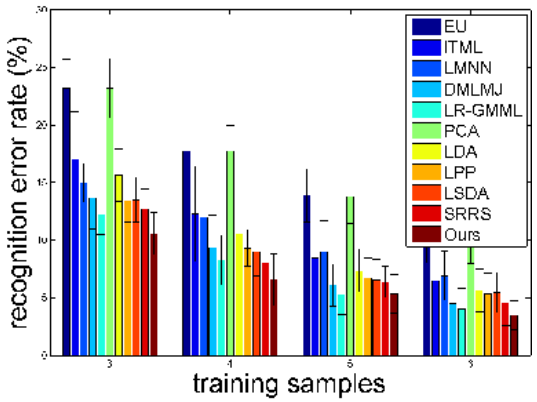


FIGURE 6. Averaged error rates of classification on ORL data set with different training samples.

3) The recognition error ratio of each method decreases with the increment of the training samples. On the other hand, at the same number of training samples, our proposed method significantly outperforms all other methods. Further, our proposed method obtains more promotion in the case of small training samples.

To further validate the robustness of the proposed method, we randomly add the Gaussian noise from 10%–50% on the data. In each noisy case, we randomly select 5 images in each class as training samples and the rest as testing samples, and repeat this process 20 times. Then, we obtain the results shown in Table 5 and Fig. 7.

Table 5 shows that the above three facts still hold in the noisy cases. Further, Fig. 7(a) shows the relationships between recognition error ratios and dimension of the learned feature space in the case of 50% Gaussian noise. In this figure, we also exclude the ITML and LMNN methods because they do not reduce the dimension of feature space. At the same time, we only display the EU curves to be baselines. It is seen that our proposed method almost obtains the best subspace representation for the samples. On the other hand,

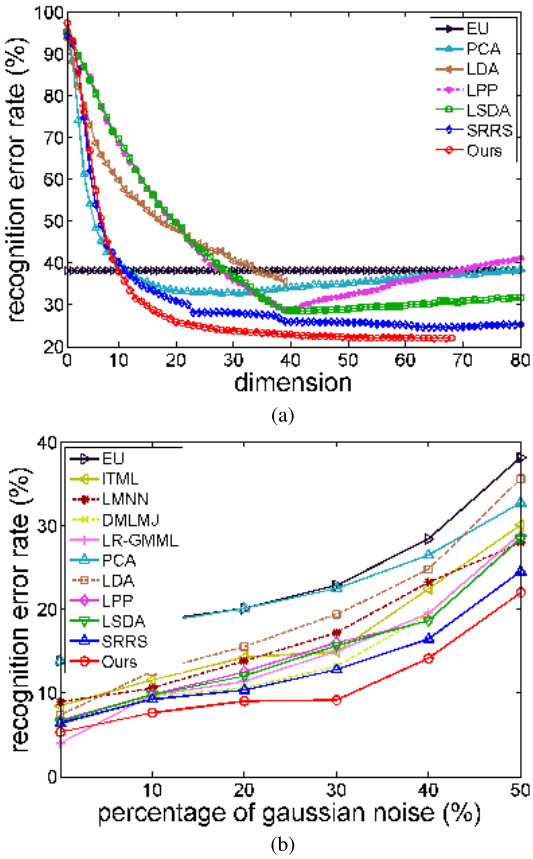


FIGURE 7. Results on ORL data set. (a) Averaged error rates on ORL data set with 50% Gaussian noise. (b) Averaged error rates of on ORL data set with different percentages of Gaussian noise.

from Fig. 7(b), we conclude that recognition error ratios increase with increment of the noise, while our proposed method obtain the best performance in all noisy cases.

Further, the above process is repeated on the Extended YaleB facial data set. We randomly select 10, 20, 30 and 40 images in each class as training samples and the rest as testing samples. We repeat this process 20 times again, and

TABLE 5. The recognition error rates (Average ± Std) on ORL data set with different percentages of Gaussian noise.

Noise	10%	20%	30%	40%	50%
EU	18.63±2.42(1024)	20.13±2.36(1024)	22.88±3.02(1024)	28.45±2.61(1024)	38.15±3.53(1024)
ITML	11.55±4.98(1024)	14.28±4.95(1024)	14.93±5.16(1024)	22.45±4.77(1024)	30.10±4.12(1024)
LMNN	10.60±1.85(1024)	13.85±2.35(1024)	17.23±3.15(1024)	23.23±3.08(1024)	28.00±5.82(1024)
DMLMJ	9.60±2.75(56)	10.63±2.67(38)	13.20±3.16(39)	19.40±3.05(36)	29.23±4.13(48)
LR-GMML	9.78±1.99(36)	11.40±2.51(37)	14.98±2.93(37)	19.52±3.30(37)	29.05±4.10(38)
PCA	18.38±2.63(36)	20.13±2.42(199)	22.50±3.30(39)	26.50±3.00(35)	32.75±3.10(31)
LDA	12.60±2.05(39)	15.55±2.96(39)	19.40±3.08(39)	24.80±3.98(39)	35.63±3.48(39)
LPP	9.82±1.66(39)	12.53±1.78(40)	16.02±2.02(39)	18.65±2.08(39)	28.57±2.76(39)
LSDA	9.78±1.87(40)	12.05±2.31(41)	15.63±1.89(41)	18.68±1.96(40)	28.45±3.10(42)
SRRS	9.24±2.56(117)	10.35±2.79(119)	12.79±3.01(88)	16.48±2.59(69)	24.52±2.35(63)
Ours	**7.63±1.88(199)**	**9.03±1.76(125)**	**9.18±1.62(120)**	**14.15±2.16(97)**	**22.02±1.82(65)**

TABLE 6. The recognition error rates (Average ± Std) on extended YaleB data set with different training samples.

Training samples	10	20	30	40	Average
EU	46.25±0.31(1024)	30.77±0.69(1024)	21.38±0.56(1024)	17.38±0.34(1024)	28.95
ITML	24.53±0.79(1024)	13.54±1.46(1024)	9.14±1.39(1024)	7.36±0.78(1024)	13.64
LMNN	20.75±1.78(1024)	10.83±0.86(1024)	7.04±0.59(1024)	5.55±0.56(1024)	11.04
DMLMJ	12.01±1.26(99)	6.29±0.88(130)	4.15±0.70(159)	3.17±0.49(1024)	6.41
LR-GMML	11.12±0.56(484)	14.55±0.86(766)	13.62±0.67(883)	10.69±1.10(903)	12.50
PCA	46.24±0.31(376)	30.77±0.71(710)	21.37±0.58(797)	17.38±0.37(888)	28.94
LDA	12.03±1.16(37)	8.87±0.83(37)	13.64±0.96(37)	4.08±0.66(37)	9.66
LPP	11.25±1.26(76)	7.23±0.65(162)	7.43±0.76(231)	2.02±0.41(226)	6.98
LSDA	11.22±1.05(299)	10.29±0.86(93)	10.44±0.70(593)	3.99±0.57(151)	8.99
SRRS	9.02±0.89(340)	5.72±0.68(759)	3.32±0.35(340)	2.37±0.45(1024)	5.11
Ours	**7.81±0.95(379)**	**4.27±0.49(391)**	**2.19±0.40(452)**	**1.31±0.33(619)**	**3.90**

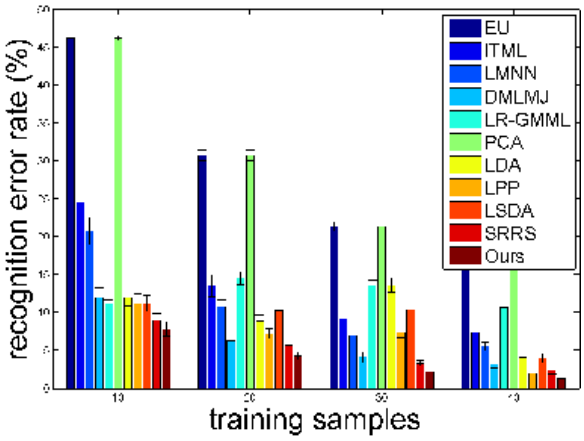


FIGURE 8. Averaged error rates of classification on extended yaleB data set with different training samples.

record the recognition error ratios with their standard variations by all methods, respectively. The results are displayed in Table 6 and Fig. 8.

Fig. 8 reveals the same phenomenon with the results on ORL data set, but the performance of PCA on this data set is not well like it on ORL data set. On the other hand, the performances of SRRS and our method are close, but our method is still better than all other methods in the same size of training samples.

3) USPS DIGITAL DATA SET

Finally, we compare all methods for the image retrieval task on USPS digital data set. Here, we only use the first 100 images for each number to be subset. Then, we randomly select 30 images in each class as training samples and the rest as testing samples. We also repeat this process 20 times. By applying nine methods on them, we obtain the results and display them in Table 7 and Figs. 9 and 10.

TABLE 7. The recognition error rates (Average ± Std) on USPS data set.

Method	Recognition Error Rate
EU	10.17±0.86(1024)
ITML	9.12±1.21(1024)
LMNN	8.66±1.13(1024)
DMLMJ	14.65±1.30(93)
LR-GMML	8.94±0.99(49)
PCA	9.91±0.97(29)
LDA	52.12±2.64(9)
LPP	50.11±3.47(14)
LSDA	46.39±2.97(143)
SRRS	9.54±0.84(50)
OUR	**8.41±0.89(50)**

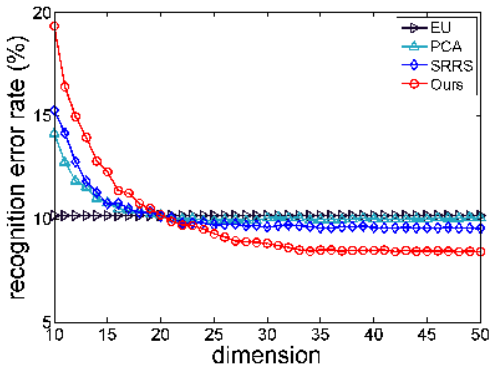


FIGURE 9. Averaged error rates on USPS data set.

From Table 7, we see that 1) All learned metrics improve the performance of retrieval. It is validated from the results of ITML and LMNN. The performances are better than the

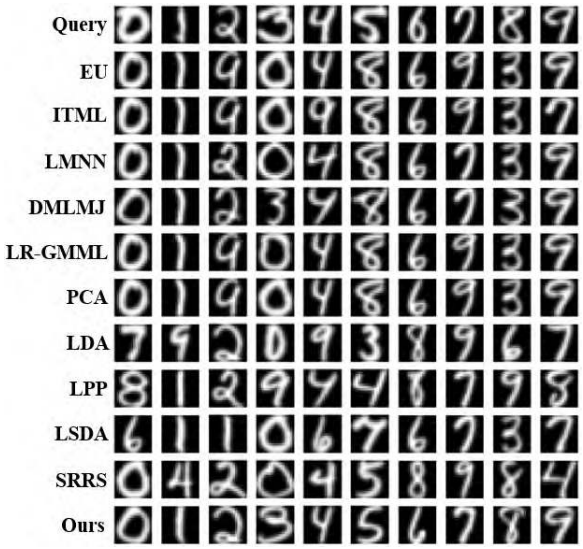


FIGURE 10. Nearest neighbor samples from USPS data set. The first row shows the queries. The rest rows correspond to the nearest neighbors of the queries obtained under eleven methods.

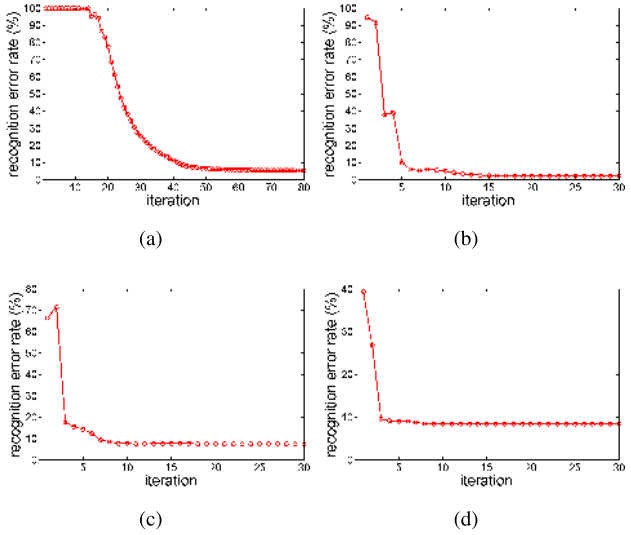


FIGURE 11. Convergence curves of our proposed method on four data sets. (a) ORL. (b) Extended YaleB. (c) COIL-100. (d) USPS.

EU-based method under the learned metric in the original feature space with dimension of 256; 2) The dimension reduction also contributes to improve the performance of retrieval. It is demonstrated from the comparison results between EU method and all dimension reduction based methods; and 3) Our proposed method still outperforms the rest state-of-the-art methods.

Fig. 9 shows the relationship between the recognition error ratios and the dimension of the learned feature space. Here, we only display the PCA, SRRS and our method with the EU baselines. It is seen that our proposed method almost obtains the best subspace representation for the samples.

To intuitively display the results, we portrait the image retrieval results in Fig. 10. Our proposed method finds the most similar images by using the learned metric.

Finally, we give the convergence analysis by the numerical experiments. Fig. 11 shows the convergence curves of our proposed method on four data sets without noise. It is seen that our proposed method has a fast convergence rate.

V. CONCLUSION

In this paper, we have proposed a robust intrinsic metric learning method based on the subspace representation for samples. Concretely, we formulate the metric learning problem as a minimization problem on the SPD manifold. To extend this model to the semi-definite cases, we introduce the robust subspace representation to our geodesic preserving metric learning method by applying the low-rank and sparse representations. To solve this model, we develop an iterative strategy to update the metric and the subspace structure, respectively. In the step of updating the metric, we construct a structure-preserving algorithm. Finally, the numerical results validate that our method can significantly improve the performance of image classification, even under high noise.

REFERENCES

[1] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2013.

[2] D. Li and Y. Tian, "Survey and experimental study on metric learning methods," *Neural Netw.*, vol. 105, pp. 447–462, Sep. 2018. doi: 10.1016/j.neunet.2018.06.003.

[3] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Michigan State Univ., East Lansing, MI, USA, Tech. Rep., 2006, pp. 1–51. [Online]. Available: https://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf

[4] Z. Ding and Y. Fu, "Robust transfer metric learning for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 660–670, Feb. 2017.

[5] C. Geng and S. Chen, "Metric learning-guided least squares classifier learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6409–6414, Dec. 2018. doi: 10.1109/TNNLS.2018.2830802.

[6] Z. Zhang, H. Lin, X. Zhao, R. Ji, and Y. Gao, "Inductive multi-hypergraph learning and its application on view-based 3D object classification," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5957–5968, Dec. 2018.

[7] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[8] Y. Gao, R. Ji, P. Cui, Q. Dai, and G. Hua, "Hyperspectral image classification through bilayer graph-based learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2769–2778, Jul. 2014.

[9] Y. Gao, M. Wang, R. Ji, X. Wu, and Q. Dai, "3-D object retrieval with Hausdorff distance learning," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2088–2098, Apr. 2014.

[10] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu, "Discriminative low-rank metric learning for face recognition," in *Proc. IEEE 11th Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.

[11] Z. Huang, R. Wang, S. Shan, L. Van Gool, and X. Chen, "Cross Euclidean-to-Riemannian metric learning with application to face recognition from video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2827–2840, Dec. 2018. doi: 10.1109/TPAMI.2017.2776154.

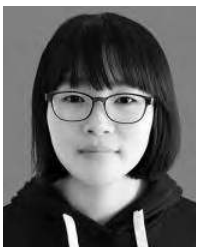
[12] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, Jun. 2018.

[13] X. Zhao, N. Wang, Y. Zhang, S. Du, Y. Gao, and J. Sun, "Beyond pairwise matching: Person reidentification via high-order relevance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3701–3714, Aug. 2018.

[14] C. Sun, D. Wang, and H. Lu, "Person re-identification via distance metric learning with latent variables," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 23–34, Jan. 2017.

[15] X. Xu, W. Li, and D. Xu, "Distance metric learning using privileged information for face verification and person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3150–3162, Dec. 2018. doi: 10.1109/TNNLS.2015.2405574.

[16] Y. Peng, L. Hu, S. Ying, and C. Shen, "Global nonlinear metric learning by gluing local linear metrics," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 423–431.
[17] X. Li, Y. Bai, Y. Peng, S. Du, and S. Ying, "Nonlinear semi-supervised metric learning via multiple kernels and local topology," *Int. J. Neural Syst.*, vol. 28, no. 2, 2018, Art. no. 1750040.
[18] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1950–1962, Sep. 2015.
[19] E. P. Xing, M. I. Jordan, S. J. Russell, A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. NIPS*, 2003, pp. 521–528.
[20] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. Adv. NIPS*, 2004, pp. 41–48.
[21] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. NIPS*, 2005, pp. 513–520.
[22] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[23] M. Der and L. K. Saul, "Latent coincidence analysis: A hidden variable model for distance metric learning," in *Proc. Adv. NIPS*, 2012, pp. 3230–3238.
[24] C. Shen, J. Kim, F. Liu, L. Wang, and A. Van den Hengel, "Efficient dual approach to distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 394–406, Feb. 2014.
[25] G. Kunapuli and J. Shavlik, "Mirror descent for metric learning: A unified approach," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2012, pp. 859–874.
[26] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 209–216.
[27] J. Mei, M. Liu, H. R. Karimi, and H. Gao, "Logdet divergence-based metric learning with triplet constraints and its applications," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4920–4931, Nov. 2014.
[28] B. Nguyen, C. Morell, and B. De Baets, "Supervised distance metric learning through maximization of the Jeffrey divergence," *Pattern Recognit.*, vol. 64, pp. 215–225, Apr. 2017.
[29] Z. Gu, M. Shao, L. Li, and Y. Fu, "Discriminative metric: Schatten norm vs. Vector norm," in *Proc. ICPR*, Nov. 2012, pp. 1213–1216.
[30] S. Zhang, W. Lu, W. Xing, and L. Zhang, "Using fuzzy least squares support vector machine with metric learning for object tracking," *Pattern Recognit.*, vol. 84, pp. 112–125, Dec. 2018. doi: 10.1016/j.patcog.2018.07.012.
[31] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Feb. 2006.
[32] J. Liu, S. Chen, X. Tan, and D. Zhang, "Comments on 'Efficient and robust feature extraction by maximum margin criterion,'" *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1862–1864, Nov. 2007.
[33] J. Li, X. Lin, X. Rui, Y. Rui, and D. Tao, "A distributed approach toward discriminative distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2111–2122, Sep. 2015.
[34] P. Li and S. Chen, "Gaussian process approach for metric learning," *Pattern Recognit.*, vol. 87, pp. 17–28, Mar. 2019. doi: 10.1016/j.patcog.2018.10.010.
[35] P. H. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, Jun. 2016, pp. 2464–2471.
[36] S. Ying, Z. Wen, J. Shi, Y. Peng, J. Peng, and H. Qiao, "Manifold preserving: An intrinsic approach for semisupervised distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2731–2742, Jul. 2018.
[37] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[38] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[39] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.
[40] S. Li and Y. Fu, "Robust subspace discovery through supervised low-rank constraints," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 162–171.
[41] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.
[42] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, May 2011.
[43] H. Qiao, P. Zhang, D. Wang, and B. Zhang, "An explicit nonlinear mapping for manifold learning," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 51–63, Feb. 2013.
[44] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[45] S. Li and Y. Fu, "Learning balanced and unbalanced graphs via low-rank coding," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1274–1287, May 2015.
[46] Z. Ding and F. Yun, "Robust multiview data analysis through collective low-rank subspace," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1986–1997, May 2018.
[47] Q. Wang, P. C. Yuen, and G. Feng, "Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions," *Pattern Recognit.*, vol. 46, no. 9, pp. 2576–2587, 2013.
[48] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2586–2593.
[49] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[50] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation AX + XB = C [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.
[51] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 569–592, 2009.
[52] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[53] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a spatially smooth subspace for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007. doi: 10.1109/CVPR.2007.383054.
[54] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.
[55] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.
[56] M. Bhutani, P. Jawanpuria, H. Kasai, and B. Mishra, "Low-rank geometric mean metric learning," in *Proc. Geometry Mach. Learn. (GiMLi) Workshop Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1–4.
[57] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. IJCAI*, Jan. 2007, pp. 708–713.



LIPENG CAI received the B.Sc. degree in applied mathematics from the Jiangxi University of Science and Technology, Jiangxi, China, in 2016. She is currently pursuing the master's degree with the Department of Mathematics, School of Science, Shanghai University, Shanghai, China. Her current research interests include metric learning and subspace methods.

SHIHUI YING (M'11) received the B.Eng. degree in mechanical engineering and the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2008, respectively. He held a postdoctoral position with the Biomedical Re-search Imaging Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, from 2012 to 2013. He is currently a Professor with the Department of Mathematics, School of Science, Shanghai University, Shanghai, China. His current research interests include geometric theory and methods for medical image processing, and machine learning.

YAXIN PENG (M'15) received the B.Sc. degree in mathematics from Anhui Normal University, Wuhu, China, in 2002, the M.Sc. degree in mathematics from East China Normal University (ECNU), Shanghai, China, in 2005, and the Ph.D. degree in mathematics from the Ecole Normale Suprieure de Lyon, Lyon, France, and ECNU, in 2008. She is currently an Associate Professor with the Department of Mathematics, School of Science, Shanghai University, Shanghai. Her research interests include geometric variation, metric learning, point cloud, and image processing.

CHANGZHOU HE received the B.S. and M.S degrees in computational mathematics from Peking University and the Chinese Academy of Sciences, Beijing, China, in 2001 and 2004, respectively. He is currently a Senior Staff Engineer with Qualcomm (Shanghai) Co. Ltd., Shanghai, China. His current research interests include machine learning and software engineering.

SHAOYI DU (M'11) received the B.S. degree in computational mathematics and in computer science, the M.S. degree in applied mathematics, and the Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China, in 2002, 2005, and 2009, respectively. He was a Postdoctoral Fellow with Xi'an Jiaotong University, from 2009 to 2011, and visited The University of North Carolina at Chapel Hill, from 2013 to 2014. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, machine learning, and pattern recognition.