Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Face recognition based on subset selection via metric learning on manifold*

Hong SHAO[1], Shuang CHEN[†‡1], Jie-yi ZHAO[2], Wen-cheng CUI[1], Tian-shu YU[3]

([1]*School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China*)

([2]*The University of Texas Health Science Center at Houston, Houston 77030, USA*)

([3]*Schulich School of Engineering, University of Calgary, Calgary T2N 1N4, Canada*)

[†]E-mail: chenshuang19891129@gmail.com

Received Mar. 19, 2015; Revision accepted Aug. 10, 2015; Crosschecked Nov. 11, 2015

Abstract: With the development of face recognition using sparse representation based classification (SRC), many relevant methods have been proposed and investigated. However, when the dictionary is large and the representation is sparse, only a small proportion of the elements contributes to the l^1-minimization. Under this observation, several approaches have been developed to carry out an efficient element selection procedure before SRC. In this paper, we employ a metric learning approach which helps find the active elements correctly by taking into account the interclass/intraclass relationship and manifold structure of face images. After the metric has been learned, a neighborhood graph is constructed in the projected space. A fast marching algorithm is used to rapidly select the subset from the graph, and SRC is implemented for classification. Experimental results show that our method achieves promising performance and significant efficiency enhancement.

Key words: Face recognition, Sparse representation, Manifold structure, Metric learning, Subset selection
doi:10.1631/FITEE.1500085 **Document code:** A **CLC number:** TP391

1 Introduction

Face recognition is one of the most important research topics in the computer vision community. Given a query image of the human face, the task of face recognition is to identify the class of this image from a large human face database. To accomplish this task, several difficulties have to be overcome, including illumination change, pose and facial expression variation, corruption or occlusion of facial features, computational complexity, etc.

Unlike traditional face recognition schemes using feature extraction, Wright *et al.* (2009) proposed a new approach based on sparse representation based classification (SRC). In this method, rather than extracting complex features from the image, the image itself can be directly used as a vectorized form, under the assumption that the vectorized face image lies in a subspace sparsely spanned by other images from the same class. The success of many face recognition methods based on SRC proved that the assumption is reasonable. According to the variable selection and shrinkage property, l^1-norm is robust to occlusion and outlier. However, the optimization of the sparsity-constrained problem is computationally expensive, resulting in difficulty in practical use.

There has been a lot of work on this topic, some of which aims at reducing the size or the dimension of the feature (namely the length of the vectorized feature). Yu *et al.* (2013) proposed to first obtain the illumination-invariant face images called

‡ Corresponding author

* Project supported by the Natural Science Foundation of Liaoning Province, China (No. 201202162)

ORCID: Shuang CHEN, http://orcid.org/0000-0001-7441-4749

the gradientface, then reduce the dimensionality over the whole image set via principal component analysis (PCA), and classify the query image using the dimension-reduced vectorized image. By reducing the length of the dimension of the subspace, l^1-optimization will run faster. Yang and Zhang (2010) combined the Gabor feature extraction and SRC. On the other hand, alternative methods have been proposed to reduce the number of the features via discriminative dictionary learning (Zhang and Li, 2010; Jiang *et al.*, 2011; Deng *et al.*, 2012; Patel *et al.*, 2012), while the dictionary has been learned over the whole image database (Zhang and Li, 2010; Jiang *et al.*, 2011), or been learned intraclass (Deng *et al.*, 2012; Patel *et al.*, 2012). In view of the intrinsic robustness to occlusion and outliers, SRC is also employed to recognize partial faces (Wagner *et al.*, 2012; Liao *et al.*, 2013).

Some effort has been made to explain the SRC in other work. He *et al.* (2011) first modeled the face recognition problem by introducing the maximum correntropy criterion. By using the half-quadratic form, the optimization on each iteration is reduced to a nonnegativity regularized weighted linear least square problem. Yang *et al.* (2011) explored the SRC using a probabilistic interpretation by modeling face recognition as a maximum likelihood estimation (MLE) problem with respect to a sparsity-constrained robust regression, which is yet another form of weighted least square problem with a sparsity constraint. He *et al.* (2014) proposed unifying error detection and error correction in face recognition under a half-quadratic framework. Error correction and error detection are realized by using the additive form and multiplicative form of half-quadratic functions, respectively. Image classification based on the manifold property has been intensively investigated. Because the variation and changes of human faces can be modeled as a manifold, some researchers try to interpret face recognition as a manifold regularized SRC problem (Zhou *et al.*, 2011; Lai *et al.*, 2013; Wang *et al.*, 2015), in which the structures of face images in terms of Reimannian geometry can be fully integrated into the SRC. Aside from regarding one image as a point on the manifold, Lu *et al.* (2013) introduced a method that uses the image patch as a single point on the manifold. Using a manifold matching scheme, images from the same class can be effectively identified. In their method, only one training image is necessary for each class. Zhang *et al.* (2011) pointed out that rather than sparse representation, collaborative representation is more important in the face recognition problem, and they verified this assumption in extensive experiments. In their work, they used l^2-norm to model the energy function, so the problem becomes a regularized least square problem, which can be solved efficiently. Motivated by Zhang *et al.* (2011), methods that use l^2-norm instead of l^1-norm were proposed by Xu *et al.* (2012; 2013) and Ortiz and Becker (2014), achieving significant reduction in computational cost.

Our method is stimulated by Xu *et al.* (2012; 2013), He *et al.* (2013), and Ortiz and Becker (2014). All of the mentioned methods have employed some iterative strategies to reduce the size of the candidate set for coefficient optimization. By doing this, the size of the optimization space can be remarkably reduced. In Xu *et al.* (2013), the first stage was conducted using least square regression over all training images, after selecting candidate classes via reconstruction residuals. In the second stage least square regression was carried out again over the obtained candidate classes. In Xu *et al.* (2012), rather than selecting the candidate classes, the authors aimed at finding the candidate images by observing the absolute values of the coefficients. Very recently, He *et al.* (2013) introduced an algorithm based on the divide-and-conquer strategy, which consists of a metric learning stage and a non-negative sparse representation stage. The above methods concentrate on reducing the optimization space using some coarse method with low computational complexity, and then compute the precise classification via fine optimization. In this paper, we also integrate a metric learning method into the recognition framework. By using the learned metric, we try to interpret the face images as manifolds in the new metric space. In this space it is easier and more natural to select a subset for the SRC.

In this paper, our contribution includes: (1) By exploring the data structure of face images, we employ a metric learning scheme considering both interclass/intraclass relationship and manifold structure of the face images. (2) Rather than k-nearest selection on unstructured data points, we introduce the selection of the subset by using fast marching along the constructed face image graph, which represents the coinciding manifolds. (3) We present a more

accurate subset selection framework for l^1-optimization on face recognition.

2 Background and observations

2.1 Revisiting SRC

Suppose that all images are represented in vector form. Given training images $\boldsymbol{d}_{i,j}$, where $i \in \{1,2,\ldots,q\}$ represents the class of the image and $j \in \{1,2,\ldots,n_i\}$ represents the number of images in the ith class, we denote $\boldsymbol{D} = [\boldsymbol{d}_{1,1},\ldots,\boldsymbol{d}_{1,n_1},\ldots,\boldsymbol{d}_{q,1},\ldots,\boldsymbol{d}_{q,n_q}]$ as the dictionary. As proposed by Wright *et al.* (2009), the conventional sparse coding model seeks to solve the following problem:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{s.t. } \boldsymbol{D}\boldsymbol{x} = \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{y}$ is the query image and $\boldsymbol{x}$ corresponds to the coefficient over each training image from $\boldsymbol{D}$. According to the theory of compressive sensing, only a few entries of $\boldsymbol{x}$ are non-zero corresponding to the same class as the query image, while the rest should all be zero. In practice, small noise is inevitable, so the problem is re-modeled as

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{s.t. } \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|_2 < \epsilon, \tag{2}$$

where ϵ is the value representing noise or occlusion. In Candès (2008), it has been shown that the reconstruction error is upper-bounded with respect to the noise level, so the solution to this problem can be regarded as reliable and robust. There are several variants of this problem such as

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \text{s.t. } \|\boldsymbol{x}\|_1 < \epsilon, \tag{3}$$

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1. \tag{4}$$

To solve the optimization problem, several approaches have been proposed (Tibshirani, 1996; Efron *et al.*, 2004; Candès and Tao, 2007). It should be noted that both the training images and the query image should be well aligned and sheared to obtain an optimal solution since SRC is very sensitive to translation and scale changes.

Once the optimal coefficient $\boldsymbol{x}$ is obtained, the query image is classified using such a strategy. For each class i, let $\delta_i : \mathbb{R}^n \to \mathbb{R}^n$ be the function to select the entries corresponding to the ith class from all training images. Using the selected entries, one can reconstruct the query image in the ith linear subspace spanned by the images from class i. Among all the classes, the one with the smallest reconstruction error is considered to be the class assigned to the query image. This procedure can be modeled as

$$\hat{i} = \arg\min_i \|\boldsymbol{y} - \boldsymbol{D}\delta_i(\hat{\boldsymbol{x}})\|_2. \tag{5}$$

From the algorithm of SRC, we see that all the training images in $\boldsymbol{D}$ participate in the l^1-optimization, while only a few of them contribute to the reconstruction.

2.2 Manifold structure in single face recognition

There has been some work on interpreting the face recognition problem under a manifold framework (Arandjelović *et al.*, 2005; Wang *et al.*, 2008), where the problem of face recognition using image sets is easily modeled as measuring the similarity between manifolds. It is natural to model a set of face images from the same class as manifold since face images from a single person change smoothly along with the pose and expression changes. For the image set, a structure can be established using the positions and relationships among points, which helps depict the local or global geometry of a manifold. This makes measurement over a manifold possible. However, as for single image classification, it is difficult to determine the neighborhood relationship over a manifold for a query image with an unknown class, so nothing about the geometry can be obtained.

Manifold learning techniques (Roweis and Saul, 2000; Tenenbaum *et al.*, 2000; Belkin and Niyogi, 2001) are always used to represent a manifold structure in a low-dimensional space while preserving some global and local features on the original samples. The global method seeks to find a projection that preserves the global relationship such as geodesics, while the local methods focus on finding a projection that preserves the weights assigned to the neighbors of one sample point.

Fig. 1 illustrates a typical single image classification problem in the 3D space, where curves with different colors represent different smooth manifolds, and the query image is abstracted as a single point. The l^2-distance is used as the metric to calculate the distance between any two points in this space.

Note that any metric that satisfies the metric criteria (namely non-negativity, coincidence axiom, symmetry, and triangle inequality) can be used to measure the distance.

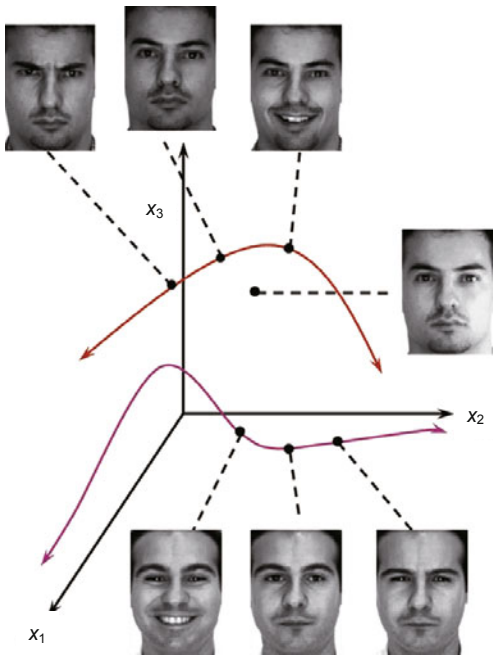

Fig. 1 An illustration on face recognition in terms of manifold structure (modified from Seung and Lee (2000)). The upper three images and the lower three images are the training samples from two classes, respectively. Training images from the same class form a low-dimensional smooth manifold embedded in high-dimensional space. The red and pink curves represent two separate manifolds. The face image in the middle corresponds to the query image. References to color refer to the online version of this figure

Several observations can be concluded from the manifold structure of face images. First, different manifolds may coincide on a local region, which means that there exist analogues near the common boundary of manifolds. Second, if the manifold is smooth and samples are dense enough, an image should lie in the subspace spanned by the ones that are near to this image on the manifold from the same class. Third, though manifolds coincide with each other randomly, the proportion of the common boundary is very small compared to the manifold itself. So, a random sample point on the manifold lies on the common boundary of manifolds with a very low probability. From these three observations, we conclude that, except for common boundaries, a graph is capable of representing most of the manifold structure and the boundaries only slightly influence the structure. This motivates the proposed approach.

3 Methodology

The proposed method seeks to solve the classic face recognition problem, while it has to be noticed that all the images from the training set and the test set are well aligned. Given a query image $\boldsymbol{y}$, we want to classify the query image using a training database $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_q]$, where $\boldsymbol{D}_i$ is constructed by stacking n_i faces from class i, and q is the number of the classes. Thus, matrix $\boldsymbol{D}$ consists of all training sets. According to the assumption in Wright *et al.* (2009), $\boldsymbol{y}$ can be represented using a sparse linear combination of the images from $\boldsymbol{D}$ as $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$, where the element x_j in $\boldsymbol{x} \in \mathbb{R}^{n_1+n_2+\ldots+n_q}$ represents the weight that each training image contributes. In the proposed approach, we first learn a metric that minimizes the interclass distance and maximizes the intraclass distance among all the training images. We next construct a graph that represents the manifold structure over all training images. Then a subset from the training images is rapidly selected taking into account the connectivity of structure. The SRC procedure is finally implemented to classify the query image into the corresponding class. The overall flowchart of our method is depicted in Fig. 2.

3.1 Metric learning for face recognition

Unlike Xu *et al.* (2012; 2013), we do not start from calculating an approximate solution to the sparse coding problem. Instead, we employ a similar strategy as in He *et al.* (2013) by first learning a metric that can enhance the discriminative ability. In this step, we employ the metric learning method presented by Weinberger and Saul (2009). Suppose that $(\boldsymbol{d}_i, c_i)$ is a training sample where c_i denotes the class of image $\boldsymbol{d}_i$, and that a binary matrix element $c_{ij} \in \{0, 1\}$ indicates whether or not classes c_i and c_j match. The goal is to learn a linear transformation $\boldsymbol{L} : \mathbb{R}^m \to \mathbb{R}^m$, which is used to compute the distance:

$$\mathrm{Dis}(\boldsymbol{d}_i, \boldsymbol{d}_j) = \|\boldsymbol{L}(\boldsymbol{d}_i - \boldsymbol{d}_j)\|^2. \tag{6}$$

Given an input $\boldsymbol{d}_i$, k' 'target' neighbors are further specified, which means k' other inputs with the

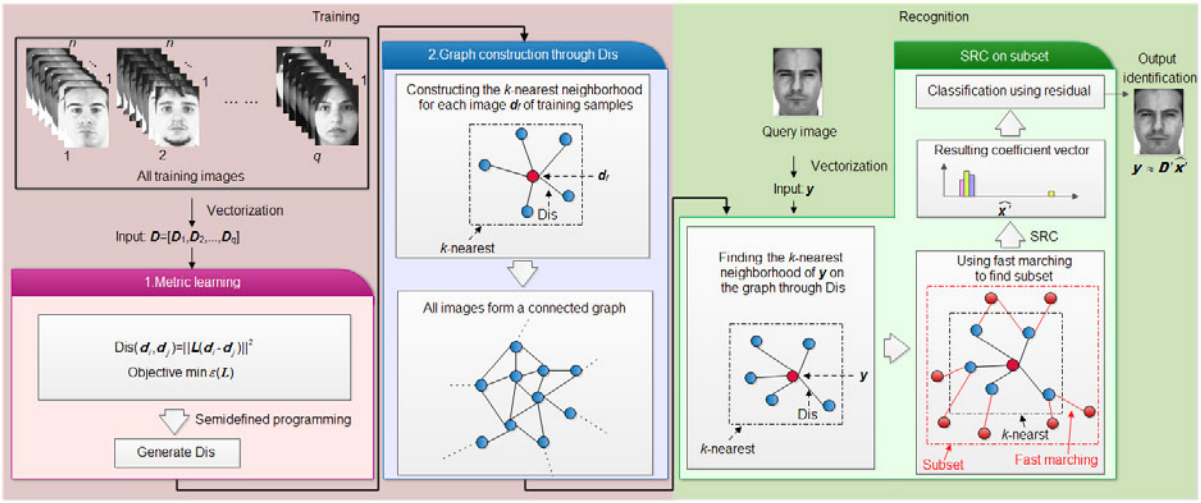

Fig. 2 Flowchart of the proposed framework

same class as c_i and the minimal distance to $\boldsymbol{d}_i$ in terms of Eq. (6). Using matrix element $\eta_{ij} \in \{0,1\}$ to indicate whether input j is a neighbor of i, the cost function over the distance metric is formulated as follows:

$$\varepsilon(\boldsymbol{L}) = \sum_{ij} \eta_{ij} \|\boldsymbol{L}(\boldsymbol{d}_i - \boldsymbol{d}_j)\|^2 + \lambda \sum_{ijl} \eta_{ij}(1 - c_{il}) \cdot [1 + \|\boldsymbol{L}(\boldsymbol{d}_i - \boldsymbol{d}_j)\|^2 - \|\boldsymbol{L}(\boldsymbol{d}_i - \boldsymbol{d}_l)\|^2]_+, \quad (7)$$

where the first term gives the constraint that the distance between a point and its neighbors cannot be too large. The second term gives the constraint that the distance between a point and its neighbors with a different class cannot be too small. Here $[z]_+ = \max(z, 0)$ is the hinge loss function. This loss function can be reformulated into a semidefinite programming problem (Vandenberghe and Boyd, 1996):

$$\min \sum_{ij} \eta_{ij}(\boldsymbol{d}_i - \boldsymbol{d}_j)^{\mathrm{T}} \boldsymbol{M}(\boldsymbol{d}_i - \boldsymbol{d}_j) + \lambda \sum_{ij} \eta_{ij}(1 - c_{il})\xi_{ijl}$$

s.t.

(1) $(\boldsymbol{d}_i - \boldsymbol{d}_l)^{\mathrm{T}} \boldsymbol{M}(\boldsymbol{d}_i - \boldsymbol{d}_l) - (\boldsymbol{d}_i - \boldsymbol{d}_j)^{\mathrm{T}} \boldsymbol{M}(\boldsymbol{d}_i - \boldsymbol{d}_j) \geq 1 - \xi_{ijl}$,

(2) $\xi_{ijl} \geq 0$,

(3) $\boldsymbol{M} \geq 0$ (positive semi-definite), (8)

where $\boldsymbol{M} = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ and ξ_{ijl} is the slack variable. This function is convex, and thus easy to solve. The comprehensive solution for this function has been presented in Vandenberghe and Boyd (1996) and Weinberger and Saul (2009). For face recognition, this stage is to aggregate the face images with the same class as closely as possible, while making images from different classes apart from each other.

In Weinberger and Saul (2009), an experiment on face recognition was implemented using the k-nearest criterion with the learned metric. However, the improvement in performance was not remarkable. We further extend the metric into a subset selection scheme, which helps SRC eliminate the computational influence from the unused samples.

3.2 Graph construction using learned metric

Before calculating the sparse representation of a query image over the training set, one has to obtain a structure that maintains the relationship of the manifolds. As discussed in Section 2.2, a graph is adequate to do this. At this stage we aim at constructing a graph representing the coinciding manifolds.

If we observe the loss function introduced by metric learning, we can find that this is very similar to the Laplacian eigenmap embedding (LEE) (Belkin and Niyogi, 2001). Given data samples $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, LEE seeks to minimize

$$\sum_{i,j} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2 \boldsymbol{W}_{i,j}, \quad (9)$$

where $\boldsymbol{W}_{i,j}$ is a weight matrix that represents the reconstruction contribution and neighborhood relationship. In Eq. (8), η_{ij} plays a similar function. The difference is that one aims to find a linear transformation $\boldsymbol{L}$ rather than a representation in a

low-dimensional space that best preserves the neighborhood relationship. However, since either Eq. (8) or Eq. (9) can preserve the locality, the local structure of the manifold is anticipated to be preserved by the learned transformation $\boldsymbol{L}$. We next consider the problem of constructing a graph from the training samples by taking into account the locality of the manifold.

A graph construction consists of finding the neighboring relationship for each point and assigning a weight or distance to each neighboring point pair. Conventional graph construction methods on the Euclidean space naturally employ l^2-distance as the metric. However, this is not a good choice for face recognition since the intraclass and interclass information is not fully considered. An example is shown in the left of Fig. 3 using l^2-distance. Though from the same class, two images depart farther away from each other than two face images from different classes. Thus, the neighboring relationship can be imprecise. Therefore, the learned metric Dis is adopted to calculate the distance between two images. An example is illustrated in the right of Fig. 3. One can observe that the neighboring relationship agrees with the class by using the metric Dis. In the experiments, we will further show that instead of l^2-distance, this metric is more powerful for intraclass and interclass variations.

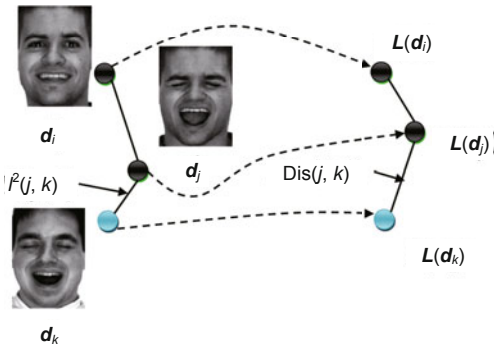

Fig. 3 Relationship between face images under l^2-distance and the learned metric Dis. In the left, though d_i and d_j are from the same class, d_j is closer to d_k which is from another person according to severe expression changes. However, as shown in the right, after metric learning, the projected point $L(d_j)$ is closer to $L(d_i)$ than to $L(d_k)$ under the metric Dis

Using the metric Dis, we implement the k-nearest method to construct a weighted graph $G = (V, E)$ on the whole training set $\boldsymbol{D}$, where V_i corresponds to the ith image $\boldsymbol{d}_i$ from $\boldsymbol{D}$, $E_{i,j} = \mathrm{Dis}(\boldsymbol{d}_i, \boldsymbol{d}_j)$, and k is larger than the size of any training sample from the same class. The second constraint is to guarantee that the graph over all training samples is connected.

3.3 SRC on subset

In this part, we present a way to select a subset from all the training images to effectively reduce the computational cost of optimization. Given the graph G, we seek to find the subset with size e along the coinciding manifolds, where e is the number of the images that are selected as the candidates to form the dictionary for the SRC. We first compute the k-nearest images from the training set with respect to the query image in terms of metric Dis. Normally $k < e$. After obtaining the k starting images, the remaining $e - k$ images are selected using the fast marching propagation algorithm (Sethian, 1999). Along with the starting k images, the $e - k$ images with the shortest geodesic distances to the query image on G form the new dictionary $\boldsymbol{D}'$. Note that the subset selection is in the new metric space with Dis, while the classification is in the original image space. Assuming that the query image lies in the subspace spanned by the images from the new dictionary $\boldsymbol{D}'$, we rewrite the constraint part as

$$\boldsymbol{y} = \boldsymbol{D}'\boldsymbol{x}', \tag{10}$$

where $\boldsymbol{x}'$ represents the coefficients that correspond only to the selected images. With this new dictionary $\boldsymbol{D}'$ and the coefficient vector $\boldsymbol{x}'$, the l^1-minimization is reformulated as

$$\hat{\boldsymbol{x}}' = \arg\min_{\boldsymbol{x}'} \|\boldsymbol{y} - \boldsymbol{D}'\boldsymbol{x}'\|_2^2 + \lambda\|\boldsymbol{x}'\|_1. \tag{11}$$

We use the same classification scheme as in Wright *et al.* (2009) by measuring the reconstruction error in each class. Since most of the images in the training set are discarded, the optimization and classification can run rapidly. This can achieve more speed-up than Xu *et al.* (2012) and Ortiz and Becker (2014).

It can also be remarked that another speed-up on the recognition phase is achieved compared with Xu *et al.* (2012) and Ortiz and Becker (2014). In their methods, a least square solution is employed to calculate a potential subset, and further optimization is implemented on this subset. This means when a new query image comes, the least square should run

again to obtain a coefficient over all training images. Though the least square solution is much faster than l^1-optimization, it still incurs extra computational cost especially when the training set is very large. On the other hand, the pre-calculated linear transformation $\boldsymbol{L}$ and the establishment of the graph G make reuse possible. In our method, we need only to compute a limited fast marching to obtain the subset, while the matrix $\boldsymbol{L}$ and the structure of the graph always remain the same. This mechanism introduces a local computation rather than a global one, and hence can save more computational cost.

4 Experiments

We conduct several experiments in this section to evaluate the performance by comparing our method with the original SRC and other subset selection methods (Wright *et al.*, 2009; He *et al.*, 2013; Xu *et al.*, 2013; Ortiz and Becker, 2014), which are denoted as SRC, TSRC, CSRC, and WSRC, respectively. Specifically, we use the down-sampled image as the single feature as in Wright *et al.* (2009), and choose the MSTR setting in He *et al.* (2013) for comparison, since the MSTR setting is reported with the best recognition performance. We further test the performance of our approach under occlusion. Finally, to evaluate the efficiency of the algorithms, we summarize the computational speed in terms of different subset sizes. All the experiments are performed on a computer with i7 CPU and 12 GB memory. We realize the counterparts using Matlab. In all of the following experiments, the face images are well aligned.

4.1 Recognition evaluation without occlusion

Experiments on four datasets are implemented in this section. We detail the datasets and the experimental results below.

AR database (Martínez and Benavente, 1998): this database consists of over 4000 images of 126 persons, with 70 males and 56 females. In this experiment, we randomly select 1400 images without occlusion from 50 males and 50 females, and 14 for each person. The image size is 165×120; we also down-sample the images into the size 80×60 in all the tests below. Fig. 4 shows some of the sample images from a single person. In the test, we use the AR database to evaluate the recognition performance in various expression and illumination conditions. Seven images from each person in Session 1 are selected as the training image, and we test the performance on the remaining seven images selected from Session 2. We test different numbers of subset size e, and set the value e to the one that reaches a good recognition accuracy. Then in the following comparison test on this dataset, we set e constant. The performance under varying e values is shown in Fig. 5a. In the AR test we set $e = 20$.

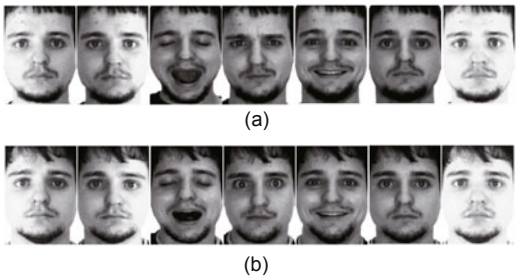

Fig. 4 Fourteen face images with a variety of expressions and illuminations from a single person in the AR dataset: (a) the seven images from Session 1 acting as the training images; (b) the seven images from Session 2 acting as the test images

The comparison of the selected methods is shown in Fig. 5b. We can observe that our method performs better than SRC. An improvement can also be seen compared to CSRC and WSRC. This is because by using the learned metric, we can accurately find the training samples from the same class as the query image that really contributes to the reconstruction. Even compared to TSRC, our method is competitive. The accuracy of our method reaches around 98.4%. This result indicates that our method is very robust to illumination and expression changes.

Extended Yale B (Georghiades *et al.*, 2001): there are 2414 front-face images from 38 different persons. The images are captured with various illumination conditions. Each image is of the size 192×168. In the test, we down-sample the image to 96×84. For each person, 32 images are randomly selected as the training samples, and the remaining ones are regarded as the test samples. Fig. 6 shows some of the example images from a single person. We choose $e = 30$ for the comparison test. The performance on varying e is shown in Fig. 7.

Because the sample images from each class are much better than the AR dataset, the accuracies of

SRC, CSRC, and WSRC are remarkably enhanced. Namely, the query image is more likely to lie in the subspace sparsely spanned by some training images from the same class. In this test, the accuracy of the proposed method reaches 98.7%, outperforming the selected counterparts.

ORL database: this database contains 40 distinct persons with 10 images per person. The images are taken at different time instances, with varying pose, expression, and detail (glasses or no glasses) conditions. Some displacement is contained in this dataset, so we have all images pre-aligned. The image size is 112×92, and we down-sample the images into size 56×46. Some example images from two persons are shown in Fig. 8. We randomly select five images from each person as the training samples, and tests are conducted on the remaining images. Since

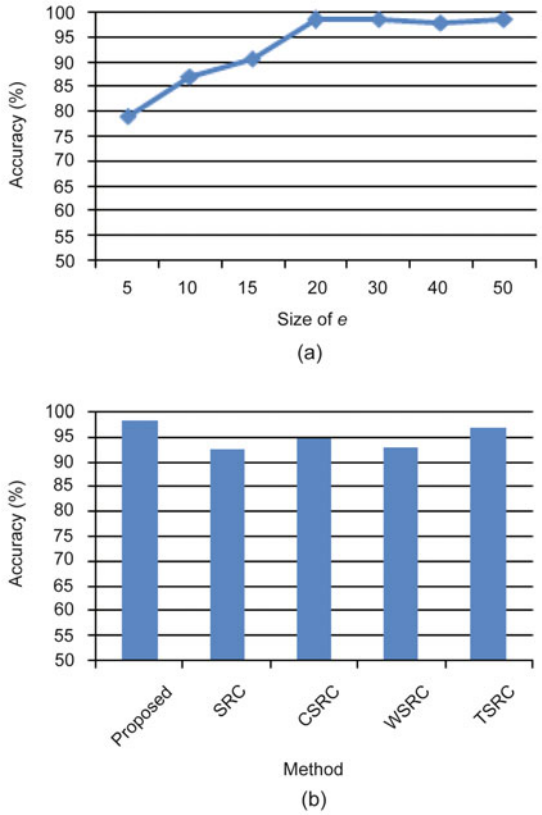

Fig. 5 Analysis of recognition performance on the AR database: (a) performance of our method under varying e values; (b) comparison of the selected methods in accuracy

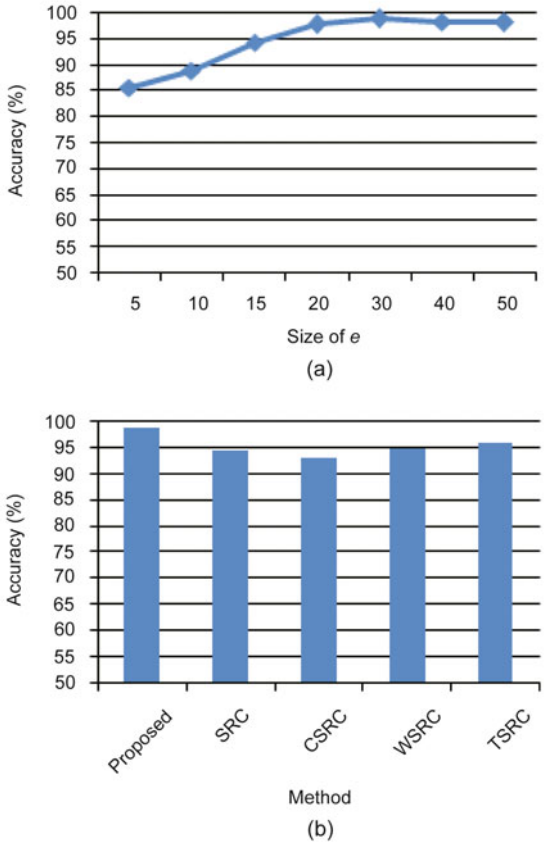

Fig. 7 Analysis of recognition performance on the extended Yale B database: (a) performance of our method under varying e values; (b) comparison of the selected methods in accuracy

Fig. 6 Face images from a single person in the extended Yale B database

the number of the training images from each person is sufficient, we follow the flow in AR and Extended Yale B tests by first choosing the best value of e through experiment. The performance using varying e is shown in Fig. 9a. In this test, we set $e = 20$. This size achieves a good trade-off between recognition accuracy and computational efficiency.

The result of the comparison test is shown in Fig. 9b. This is very similar to the result of the AR test. Our method classifies the face images more accurately than the subset selection method (Xu *et al.*, 2013; Ortiz and Becker, 2014). Moreover,

Fig. 8 Several face images of two classes from the ORL database

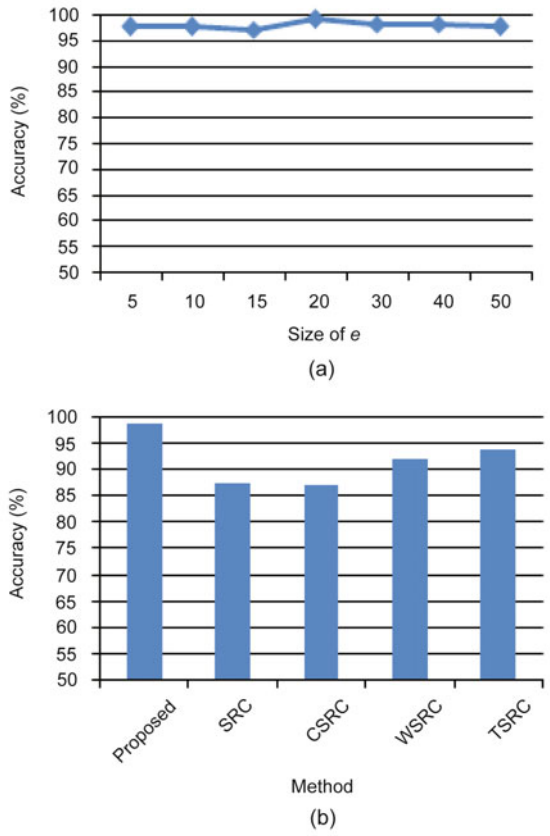

Fig. 9 Analysis of recognition performance on the ORL database: (a) performance of our method under varying e values; (b) comparison of the selected methods in accuracy

the recognition accuracy of the proposed method is even higher than that of He *et al.* (2013). The result demonstrates that, by combining metric learning and subset selection, the proposed method performs well on moderate category size against pose and expression changes.

FERET (Phillips *et al.*, 1998): the test set of FERET consists of 1400 face images from 200 persons (seven for each). The images are labeled with 'ba', 'bd', 'be', 'bf', 'bg', 'bj', and 'bk'. 'ba' refers to the front-image of the human face. 'bd' and 'bg' are with the rotation +25 degrees and −25 degrees, respectively. 'be' and 'bf' are with the rotation +15 degrees and −15 degrees, respectively. 'bj' consists of expression changes and 'bk' consists of illumination changes. The size of the FERET image is 80×80. We use this dataset to test the robustness of our method against pose change. Some sample images from a single person are shown in Fig. 10. We first take all images with labels 'ba' and 'bj' as the training samples, and images with 'bk' as the test samples. In the following four tests, 'ba', 'bj', and 'bk' are regarded as the training samples and images with the rest of the labels as the test samples. Since the training images from each person are severely insufficient, we set $e = 8$ in the experiment.

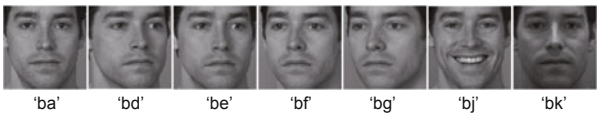

Fig. 10 Seven face images of one class from the FERET database

This test is much more challenging than AR and Extended Yale B, since the number of the training images from each person seems insufficient. The experimental results of FERET are presented in Fig. 11, which supports our assumption. From the curves, we see that the number of training images from each person heavily influences the recognition performance of the methods using SRC. In the first test, because the training samples and testing samples are both front-face images, the proposed method reaches an accuracy of 89%, performing similarly to the remaining methods. However, in the following tests, only three training samples are selected with the same pose, which cannot span a subspace that fully represents face changes upon poses. We can see that the recognition accuracy drops drastically when

the rotation angle gets larger. Specifically, for the method in SRC, the recognition accuracy is lower than 40% when there is severe rotation. In the case of rotating +25 degrees or −25 degrees, though the accuracy of the proposed method drops to 83.5% and 79.0% in this test respectively, the metric learning procedure guarantees that the accuracy remains relatively robust. The method of WSRC works well because they use robust combined features. The performance of the proposed method shows that the metric learning procedure also works similarly to the function of robust feature extraction.

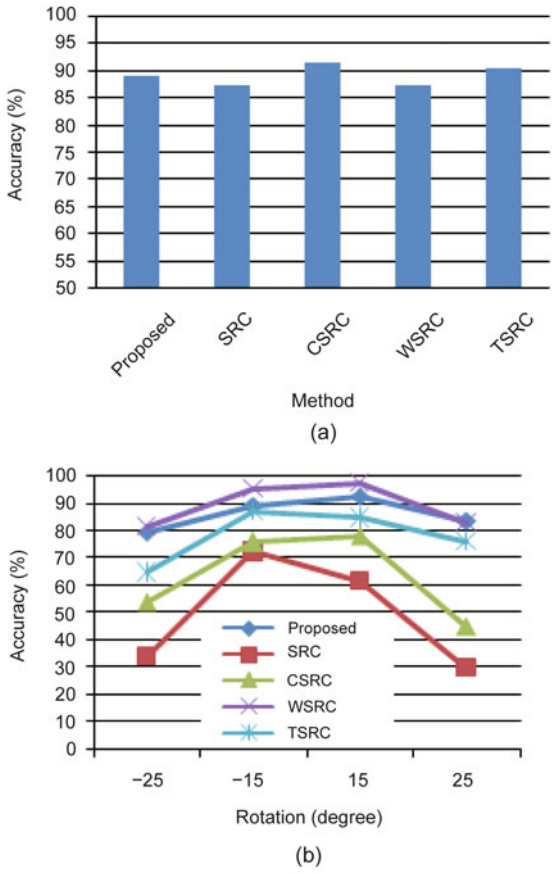

Fig. 11 Experimental results of five tests on the FERET database: (a) comparison of the selected methods in test 1; (b) performance of each selected method under various degrees of rotation (tests 2–5)

4.2 Recognition evaluation with occlusion

In this subsection, we evaluate the robustness of the proposed and the selected methods under different levels of occlusion, including sunglasses, scarf, and random block occlusion. In the experiments on sunglasses and scarf, we follow the training set setting as in the AR test without occlusion: 700 images collected from 50 males and 50 females, and 7 images for each person. We consider two separate test sets of 200 images for sunglasses occlusion and scarf occlusion, respectively. For images with sunglasses, the proportion of the occlusion is around 20% of each test image. The second test set includes persons wearing scarves, which roughly occludes 40% of the image. Four sample images for sunglasses and scarf occlusion are shown in Fig. 12a.

In Fig. 12b, the comparison results on both cases are presented. In the case of sunglasses occlusion, the proposed method reaches an accuracy of 92.0%, higher than those of SRC, CSRC, and WSRC, and close to that of TSRC. For scarf occlusion, the recognition accuracy obviously drops for each method, yet the proposed method reaches 75.5%, better than those of SRC, CSRC, and WSRC. This provides evidence to show that the metric learning introduced in the proposed method works well on corruption and outlier.

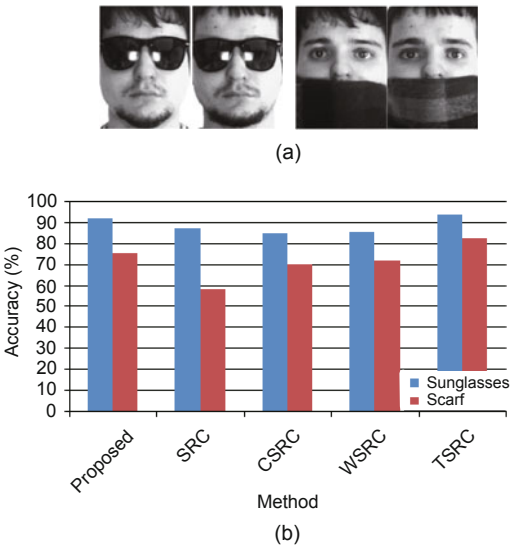

Fig. 12 Recognition under the sunglasses and scarf occlusion: (a) images of sunglasses and scarf occlusion; (b) comparison results for both cases

Since sunglasses and scarf appear on a specific region of a human face image, to avoid the bias introduced, we further conduct a test by casting block occlusion onto the test images with varying sizes and positions. The ratios of block side length to the face image side length are selected as $10\%, 20\%, 30\%, 40\%, 50\%$. We manually

generate the occlusion by filling the gray value of the corresponding region to 0. Some sample images are shown in Fig. 13a, and the results are presented in Fig. 13b. Our method outperforms SRC, CSRC, and WSRC in most cases. Compared to TSRC, our method works better when the occlusion percentage is smaller than 30%. At 40% occlusion, the accuracy is 72.5%, while at 50% occlusion, the recognition rate is 54.0%.

4.3 Efficiency comparison

We evaluate the computational efficiency of the proposed algorithm and the selected counterparts in this section. All the algorithms are realized in Matlab on a computer with i7 CPU and 12 GB memory. There are two phases in our algorithm, as presented in Fig. 2, namely training and recognition. In realistic applications, since the training phase can be implemented beforehand, the bottleneck of the efficiency is mainly in the recognition phase. In this section, we focus our attention on the comparison of recognition speed.

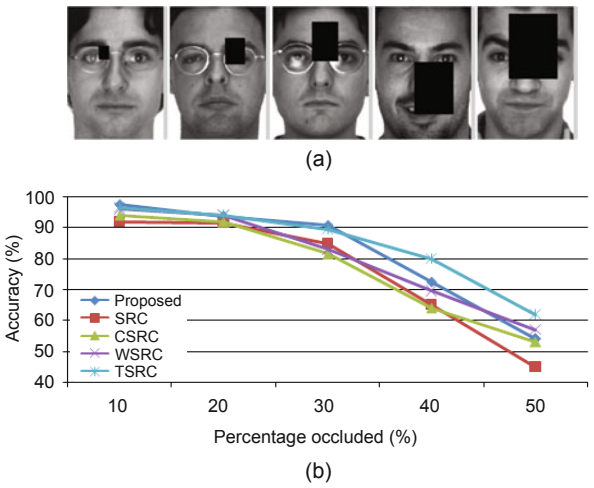

Fig. 13 Recognition under block occlusion: (a) images of block occlusion; (b) performance of the proposed and the selected methods under various block occlusions

We follow the experimental settings in Section 4.1 with AR, Yale B, ORL, and FERET databases, which reflect different image dimensions and sizes of training sets. Therefore, we test the computational efficiency of our algorithm under various subset sizes, along with the selected counterparts. The results are given in Table 1. The number on each entry represents the average recognition time, followed with number in brackets, which is the extra time for maximum recognition cost. For example, 0.207(+0.211) in the second column and the second row means that the average recognition time of the proposed method with subset size 5 is 0.207 s, and the maximum recognition time is 0.207+0.211 s, hence 0.418 s.

From the results, firstly, it can be observed that the recognition speed of the proposed method rises slightly along with the increasing subset size. When the size of the subset is small (from 5 to 50), the difference of the recognition speed is negligible, and the recognition speed for a single image is below 0.25 s for all the four databases, which makes our algorithm possible for realistic applications. When the

Table 1 Comparison of recognition speed

Method	Recognition time (s)			
	AR	Yale B	ORL	FERET
Proposed				
Subset size: 5	0.207(+0.211)	0.134(+0.157)	0.073(+0.086)	0.110(+0.145)
Subset size: 10	0.219(+0.208)	0.138(+0.150)	0.082(+0.106)	0.117(+0.151)
Subset size: 20	0.220(+0.203)	0.145(+0.127)	0.102(+0.098)	0.118(+0.172)
Subset size: 30	0.227(+0.232)	0.141(+0.139)	0.108(+0.110)	0.129(+0.161)
Subset size: 50	0.239(+0.276)	0.158(+0.145)	0.110(+0.142)	0.134(+0.173)
Subset size: 100	0.349(+0.321)	0.192(+0.207)	0.171(+0.161)	0.187(+0.199)
Subset size: 150	0.401(+0.410)	0.219(+0.304)	0.203(+0.172)	0.220(+0.275)
SRC	78.405(+94.223)	41.089(+33.460)	11.900(+17.112)	28.142(+34.447)
CSRC	4.419(+7.889)	3.744(+3.412)	2.590(+3.077)	3.198(+4.106)
WSRC	4.018(+4.980)	3.146(+3.484)	2.183(+3.421)	2.871(+3.897)
TSRC	0.632(+0.703)	0.561(+0.475)	0.440(+0.344)	0.479(+0.543)

The number on each entry represents the average recognition time and the number in brackets is the extra time for maximum recognition cost

size of the subset becomes relatively large (larger than 100), the recognition efficiency of the proposed algorithm slows down accordingly. This is because the calculation of l^1-regularized optimization becomes dominant when the subset size is large. However, it remains an acceptable speed for real-time requirement. According to Section 4.1, the recognition rate of our algorithm does not increase significantly when the size of the subset reaches a proper value. Therefore, it can be concluded that with the applicable recognition rate, our algorithm can also enhance the recognition speed-up to a run-time level.

If we compare the recognition speeds of the proposed method and the selected counterparts, we find a great improvement of recognition speed with our algorithm. As in the lower part of Table 1, SRC shows its drawback of speed in calculating the sparse solution when the dictionary (database) is large. This is because l^1-optimization is an extremely computation-intensive process, especially when the sizes of dimension and dictionary are large. CSRC and WSRC both introduce the least square method to estimate an initial distribution of the sparse solution, so they show similar performance on either database. However, least square involves calculation of an inverse matrix, which can be computationally expensive when the database is large. Though CSRC and WSRC show great improvement compared with standard SRC, they still cost several seconds to identify a face image. On the other hand, TSRC introduces a recognition scheme similar to the proposed method, by first learning a diagonal metric matrix, then identifying the face label using two times of non-negative sparse coding. The speed performance of TSRC is quite similar to that of our method. However, the two times of sparse coding doubles the optimization procedure, which leads to more computational time.

Note that the time cost of our algorithm is subject to the training phase, where a metric is learned according to the training samples. This procedure typically costs about 20 min for a database with 2000 images. However, since the metric can be pretrained, this will not influence the recognition speed. TSRC is also in this situation.

In general, the proposed method outperforms the selected counterparts on time performance, without sacrificing recognition accuracy.

5 Conclusions

In this paper, we present a novel face recognition approach integrating metric learning and subset selection taking into account the manifold structure of human face images. The first stage of the proposed method is to learn a robust metric in terms of a linear projection so that the interclass and intraclass relationship, as well as the manifold structure, can be preserved. The second stage involves graph construction using the learned metric as distance measurement. Finally, we use a fast marching algorithm to select a subset from all training images to perform SRC, which is much faster than the conventional one. This method is intuitive and easy to implement. Experimental results show that the proposed approach achieved promising recognition performance with remarkable efficiency.

References

Arandjelović, O., Shakhnarovich, G., Fisher, J., *et al.*, 2005. Face recognition with image sets using manifold density divergence. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.581-588. [doi:10.1109/CVPR.2005.151]

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS*, **14**:585-591.

Candès, E.J., 2008. The restricted isometry property and its implications for compressed sensing. *Compt. Rend. Math.*, **346**(9-10):589-592. [doi:10.1016/j.crma.2008.03.014]

Candès, E.J., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.*, **35**(6):2313-2351. [doi:10.1214/009053606000001523]

Deng, W.H., Hu, J.N., Guo, J., 2012. Extended SRC: undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Patt. Anal. Mach. Intell.*, **34**(9):1864-1870. [doi:10.1109/TPAMI.2012.30]

Efron, B., Hastie, T., Johnstone, I., *et al.*, 2004. Least angle regression. *Ann. Stat.*, **32**(2):407-499. [doi:10.1214/009053604000000067]

Georghiades, A.S., Belhumeur, P.N., Kriegman, D., 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Patt. Anal. Mach. Intell.*, **23**(6):643-660. [doi:10.1109/34.927464]

He, R., Zheng, W.S., Hu, B.G., 2011. Maximum correntropy criterion for robust face recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, **33**(8):1561-1576. [doi:10.1109/TPAMI.2010.220]

He, R., Zheng, W.S., Hu, B.G., *et al.*, 2013. Two-stage nonnegative sparse representation for large-scale face recognition. *IEEE Trans. Neur. Netw. Learn. Syst.*, **24**(1):35-46. [doi:10.1109/TNNLS.2012.2226471]

He, R., Zheng, W.S., Tan, T.N., *et al.*, 2014. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **36**(2):261-275. [doi:10.1109/TPAMI.2013.102]

Jiang, Z.L., Lin, Z., Davis, L.S., 2011. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1697-1704. [doi:10.1109/CVPR.2011.5995354]

Lai, Z.H., Li, Y.J., Wan, M.H., *et al.*, 2013. Local sparse representation projections for face recognition. *Neur. Comput. Appl.*, **23**(7):2231-2239. [doi:10.1007/s00521-012-1174-0]

Liao, S.C., Jain, A.K., Li, S.Z., 2013. Partial face recognition: alignment-free approach. *IEEE Trans. Patt. Anal. Mach. Intell.*, **35**(5):1193-1205. [doi:10.1109/TPAMI.2012.191]

Lu, J.W., Tan, Y.P., Wang, G., 2013. Discriminative multimanifold analysis for face recognition from a single training sample per person. *IEEE Trans. Patt. Anal. Mach. Intell.*, **35**(1):39-51. [doi:10.1109/TPAMI.2012.70]

Martínez, A., Benavente, B., 1998. The AR Face Database. CVC Technical Report 24.

Ortiz, E.G., Becker, B.C., 2014. Face recognition for web-scale datasets. *Comput. Vis. Image Understand.*, **118**:153-170. [doi:10.1016/j.cviu.2013.09.004]

Patel, V.M., Wu, T., Biswas, S., *et al.*, 2012. Dictionary-based face recognition under variable lighting and pose. *IEEE Trans. Inform. Forens. Secur.*, **7**(3):954-965. [doi:10.1109/TIFS.2012.2189205]

Phillips, P.J., Wechsler, H., Huang, J., *et al.*, 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.*, **16**(5):295-306. [doi:10.1016/S0262-8856(97)00070-X]

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500):2323-2326. [doi:10.1126/science.290.5500.2323]

Sethian, J.A., 1999. Fast marching methods. *SIAM Rev.*, **41**(2):199-235. [doi:10.1137/S0036144598347059]

Seung, H.S., Lee, D.D., 2000. The manifold ways of perception. *Science*, **290**(5500):2268-2269. [doi:10.1126/science.290.5500.2268]

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500):2319-2323. [doi:10.1126/science.290.5500.2319]

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**(1):267-288.

Vandenberghe, L., Boyd, S., 1996. Semidefinite programming. *SIAM Rev.*, **38**(1):49-95. [doi:10.1137/1038003]

Wagner, A., Wright, J., Ganesh, A., *et al.*, 2012. Toward a practical face recognition system: robust alignment and illumination by sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **34**(2):372-386. [doi:10.1109/TPAMI.2011.112]

Wang, L.F., Wu, H.Y., Pan, C.H., 2015. Manifold regularized local sparse representation for face recognition. *IEEE Trans. Circ. Syst. Video Technol.*, **25**(4): 651-659. [doi:10.1109/TCSVT.2014.2335851]

Wang, R.P., Shan, S.G., Chen, X.L., *et al.*, 2008. Manifold-manifold distance with application to face recognition based on image set. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2008.4587719]

Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, **10**:207-244. [doi:10.1145/1577069.1577078]

Wright, J., Yang, A.Y., Ganesh, A., *et al.*, 2009. Robust face recognition via sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(2):210-227. [doi:10.1109/TPAMI.2008.79]

Xu, Y., Zuo, W.M., Fan, Z.Z., 2012. Supervised sparse representation method with a heuristic strategy and face recognition experiments. *Neurocomputing*, **79**:125-131. [doi:10.1016/j.neucom.2011.10.013]

Xu, Y., Zhu, Q., Fan, Z.Z., *et al.*, 2013. Using the idea of the sparse representation to perform coarse-to-fine face recognition. *Inform. Sci.*, **238**:138-148. [doi:10.1016/j.ins.2013.02.051]

Yang, M., Zhang, L., 2010. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. Proc. 11th European Conf. on Computer Vision, p.448-461. [doi:10.1007/978-3-642-15567-3_33]

Yang, M., Zhang, D., Yang, J., 2011. Robust sparse coding for face recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.625-632. [doi:10.1109/CVPR.2011.5995393]

Yu, Z.P., Wu, Z.D., Zhang, J.W., 2013. An illumination robust algorithm for face recognition via SRC and Gradientfaces. Proc. 2nd Int. Conf. on Innovative Computing and Cloud Computing, p.36-40. [doi:10.1145/2556871.2556880]

Zhang, D., Yang, M., Feng, X.C., 2011. Sparse representation or collaborative representation: which helps face recognition? Proc. IEEE Int. Conf. on Computer Vision, p.471-478. [doi:10.1109/ICCV.2011.6126277]

Zhang, Q., Li, B.X., 2010. Discriminative K-SVD for dictionary learning in face recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2691-2698. [doi:10.1109/CVPR.2010.5539989]

Zhou, T.Y., Tao, D.C., Wu, X.D., 2011. Manifold elastic net: a unified framework for sparse dimension reduction. *Data Min. Knowl. Discov.*, **22**(3):340-371. [doi:10.1007/s10618-010-0182-x]