Neural Comput & Applic (2019) 31:3117–3140
DOI 10.1007/s00521-017-3258-3

ORIGINAL ARTICLE

Non-negative enhanced discriminant matrix factorization method with sparsity regularization

Ming Tong[1] · Haili Bu[1] · Mengao Zhao[1] · Shengnan Xi[1] · Hailong Li[1]

Received: 18 May 2017 / Accepted: 14 October 2017 / Published online: 2 November 2017

Abstract Efficient low-rank representation of data plays a significant role in the field of computer vision and pattern recognition. In order to obtain a more discriminant and sparse low-dimensional representation, a novel non-negative enhanced discriminant matrix factorization method with sparsity regularization is proposed in this paper. Firstly, the local invariance and discriminant information of the low-dimensional representation are incorporated into the objective function to construct a new within-class encouragement constraint term, and the weighted coefficients are introduced to further enhance the compactness between the samples that belong to the same class in the new base space. Secondly, a new between-class penalty term is constructed to maximize the difference between different classes of samples, and meanwhile, the weighted coefficients are introduced to further enhance the discreteness and discriminativeness between classes. Finally, to learn the part-based representation of data better, the sparse constraint term is further introduced, and consequently, the sparseness of data representation, the local invariance, and the discriminativeness are integrated into a unified framework. Moreover, the optimization solution and the convergence proof of objective function are given. The extensive experiments demonstrate the strong robustness of the proposed method to face recognition and image classification under occlusions.

Keywords Non-negative matrix factorization · Face recognition · Feature extraction · Data representation · Discriminant information

1 Introduction

Data representation plays a significant role in many pattern recognition and image processing tasks. A good representation of data should reveal the latent structure of a dataset and therefore is used to improve performance and reduce redundancy. The essence of data representation lies in finding an appropriate low-rank representation, and matrix factorization is just one of the most basic tools for this representation. Most popular matrix factorization methods include principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], and so on, which have been widely used in the fields of dimensionality reduction [3–5] and feature extraction [6, 7]. However, the above methods cannot guarantee the non-negativity of factorization results even though all the input data are positive, but negative values lack explicit physical meaning in some practical applications. Meanwhile, the data obtained by above methods do not possess intelligent characteristic that the whole can be perceived by the parts. Fortunately, non-negative matrix factorization (NMF) [8] provides a novel technical way to solve the above problems.

NMF is a matrix decomposition method under non-negative constraint, which has good data representation

✉ Ming Tong
mtong@xidian.edu.cn

Haili Bu
bettybhl@163.com

Mengao Zhao
mazhao_0@163.com

Shengnan Xi
snxi@stu.xidian.edu.cn

Hailong Li
lufeilong1991@163.com

[1] School of Electronic Engineering, Xidian University, Xi’an 710071, China

ability, and can significantly reduce data dimension. Meanwhile, the factorization characteristic conforms to the intuitive experience of human visual perception. Further, factorization results possess strong interpretability and explicit physical meaning. Based on the above advantages, NMF method has been successfully used in many applications, such as dimensionality reduction [9, 10], feature extraction [11], image processing [12, 13], and face recognition [14, 15]. However, the sparseness of basic NMF method is uncontrollable. Moreover, the basic NMF method fails to take advantage of the latent geometry structure of data and ignores the discriminant information of data. Therefore, some improved NMF methods have been proposed successively.

In order to make the factorization results reflect the local feature information as much as possible, Li et al. [16] proposed a local non-negative matrix factorization (LNMF) method, which can maximize the sparseness of coefficient matrix and simultaneously make the base matrix possess more orthogonal and more localized feature representation in a simple form. To further precisely control the sparseness, Hoyer [17] proposed an extended NMF method by introducing the nonlinear projection operator, namely NMF with Sparseness Constraints (NMFSC), which can make both base matrix and coefficient matrix achieve the desired sparseness. In order to control the sparseness adaptively, Yang et al. [18] proposed an adaptive non-smooth NMF (Ans-NMF) method to learn the factor matrix **S** through an adaptive method, which makes the sparseness of base matrix and coefficient matrix be controlled separately, and consequently, Ans-NMF obtains better-localized feature.

The above methods do not take the discriminant information of data into account, and accordingly, fail to deal with the problems of recognition and classification well. To further introduce discriminant information, Wang et al. [19] presented the Fisher-NMF (FNMF) by imposing fisher constraint. Zafeiriou et al. [20] further extended the FNMF method by adding different divergence terms into the objective function, so as to obtain the discriminant subspace for dealing with the facial expression recognition problem. Nikitidis et al. [21] utilized within-class multimodal distribution of data to divide the data into subclasses, and obtained more discriminative projection representation by introducing the clustering-based discriminant criteria into the objective function. Guan et al. [22] developed the manifold regularized discriminative NMF (MD-NMF) by maximizing the margins between different classes, and successfully improved the performance of NMF in face recognition tasks. Lu et al. [23] took into account the incoherent information of base matrix and coefficient matrix in basic NMF, and constructed objective function to enhance the discriminative ability of the learned base matrix. Meanwhile, the method combined the low-dimensional representation with the subspace of base matrix to regularize NMF for the learning of discriminant subspace. Chen et al. [24] presented a novel supervised and nonlinear approach to improve the classification performance of NMF. By projecting the input data into a reproducing kernel Hilbert space (RKHS), the nonlinear relationships between data are mined. At the same time, the discriminant analysis was utilized to assure the within-class separation to be small and between-class separation to be large in the RKHS.

In the above improved NMF algorithms with a certain degree of discrimination, the relationship between sample and mean is taken into account, and the within-class and between-class constraint terms are constructed based on it. In fact, the within-class constraint terms are not sufficient to well aggregate all the within-class samples, and the between-class constraint terms are insufficient to well separate the similar samples between classes, which tend to cause confusion. This is especially true when dealing with the problems of face recognition and image classification under occlusion. The above analyses show that all of these methods do not make full use of the similarity between the samples that belong to the same class and the difference between different classes of samples.

In order to reveal and utilize the intrinsic geometry structure of data, GNMF [25], GDNMF [26], LCGNMF [27], and NLMF [28] methods used the low-dimensional manifold characteristics of data as a geometric descriptor to construct the graph Laplacian regularization term, which improve the performance of image clustering or classification. Spatial NMF method [29] automatically learned the structural features of data as much as possible by introducing the divergence constraint term. Feng et al. [30] incorporated a sparse noise term into the objective function of original NMF, and meanwhile, constructed a locally weighted sparse graph regularization term to exploit the local geometric structure information of data.

On the basis of above methods, a novel non-negative enhanced discriminant matrix factorization method with sparsity regularization (NEDMF_SR) is presented in this paper. The main contributions and innovations are summarized as follows: (1) The local invariance and discriminativeness are incorporated into the objective function to construct a new within-class encouragement constraint term, which enhances the within-class compactness. (2) A new between-class penalty term is constructed to maximize the difference between the samples of any two classes in the new base space, which further enhances the discriminativeness. (3) The sparse constraint term is further introduced into the objective function, and consequently, the sparseness, the local invariance, and the discriminativeness are integrated into a unified framework. Meanwhile, the optimization solution and convergence proof are given.

The subsequent chapters of this paper are organized as follows. Section 2 briefly reviews the basic discriminant NMF method. Section 3 presents the NEDMF_SR method in detail. Section 4 shows the experimental results and analysis. Section 5 concludes this paper with direction for future work.

2 Basic NMF method

The basic NMF can be stated in the following manner. Assuming that there are q non-negative p-dimensional sample vectors $\mathbf{b}_i$ $(i = 1, 2, \ldots, q)$, which form an original non-negative matrix $\mathbf{B} \in R_+^{p \times q}$ of size $p \times q$, then the approximate non-negative matrix factorization is applied to matrix $\mathbf{B}$, such that:

$$\mathbf{B} \approx \mathbf{ZH} \tag{1}$$

where $\mathbf{Z} \in R_+^{p \times f}$ is the base matrix, and $\mathbf{H} \in R_+^{f \times q}$ is the coefficient matrix. f denotes the factorization dimension and subjects to the condition of $f < pq/(p+q)$.

When the Generalized Kullback–Leibler Divergence (GKLD) is taken as the objective function, NMF can be transformed into the following constrained optimization problem [8]:

$$\begin{cases} \min\limits_{Z,H} D_{KL}(\mathbf{B} \| \mathbf{ZH}) = \sum_{i,j} \left(B_{i,j} \log \dfrac{B_{i,j}}{\sum_l Z_{i,l} H_{l,j}} - B_{i,j} + (\mathbf{ZH})_{i,j} \right) \\ s.t. \;\; Z_{i,l} \geq 0, H_{l,j} > 0, \;\forall i,j,l \end{cases} \tag{2}$$

This problem can be optimized through the multiplicative update algorithm [8], which is simple and able to achieve good performance, and accordingly, the update rules for the elements of base matrix and coefficient matrix can be given by Eqs. (3) and (4), respectively:

$$Z_{i,l} \leftarrow \frac{Z_{i,l} \sum_j \left[H_{l,j} B_{i,j} \Big/ (\mathbf{ZH})_{i,j} \right]}{\sum_s H_{l,s}} \tag{3}$$

$$H_{l,j} \leftarrow \frac{H_{l,j} \sum_i \left[Z_{i,l} B_{i,j} \Big/ (\mathbf{ZH})_{i,j} \right]}{\sum_v Z_{v,l}} \tag{4}$$

3 NEDMF_SR

In this section, the proposed method, i.e., NEDMF_SR, is introduced. Firstly, the motivation of the proposed method is given, and then, the within-class compact encouragement term and the between-class discrete penalty term are illustrated in detail, both of which are combined with the sparse constraint terms and incorporated into a joint framework to obtain the objective function of the NEDMF_SR method. Subsequently, the update rules and convergence analysis are presented. Finally, a framework based on the NEDMF_SR method for joint feature extraction is given detailedly, as shown in Fig. 1.

3.1 Motivation

NMF aims to search for a set of base vectors that are utilized to best approximate the raw data. A natural assumption here could be that if two points $\mathbf{b}_i$, $\mathbf{b}_l$ to be factorized are adjacent in the intrinsic geometry structure of data distribution, then the representations of the two points in regard to the new base space are also adjacent to each other; similarly, if two points $\mathbf{b}_i$, $\mathbf{b}_l$ are discrete in the intrinsic geometry of data distribution, then the representations of the two points in regard to the new base space are also far away from each other. It is universally acknowledged as the local invariance assumption [31, 32], which has a significant influence on the progress of various kinds of algorithms for dimensionality reduction.

Therefore, the within-class compact encouragement term and the between-class discrete penalty term are proposed to introduce the discriminant information and intrinsic geometry structure of data effectively, and then, the two constraints are incorporated into the objective function of NEDMF_SR, which makes the NEDMF_SR method learn a more compact and discriminative low-dimensional representation.

3.2 Within-class compact encouragement term

In order to make the within-class samples more compact in the new base space, low-dimensional cohesion degree between any two within-class samples is fully considered to construct the within-class constraint term. Meanwhile, to further aggregate within-class samples, the different within-class compact weight coefficients are imposed to form within-class compact encouragement term. The specific construction steps are as follows:

1. Construction of within-class constraint term

Let $\boldsymbol{\eta}_j^{(i)}$ denote the representation of the j-th sample for class i in the new base space, that is, the coefficient vector for the j-th sample of class i. The within-class cohesion degree s_i between $\boldsymbol{\eta}_j^{(i)}$ is measured by using Euclidean distance and is calculated as Eq. (5).

$$s_i = \frac{1}{N_i(N_i - 1)} \sum_{j \neq m}^{N_i} \left\| \boldsymbol{\eta}_j^{(i)} - \boldsymbol{\eta}_m^{(i)} \right\|_2^2 \tag{5}$$

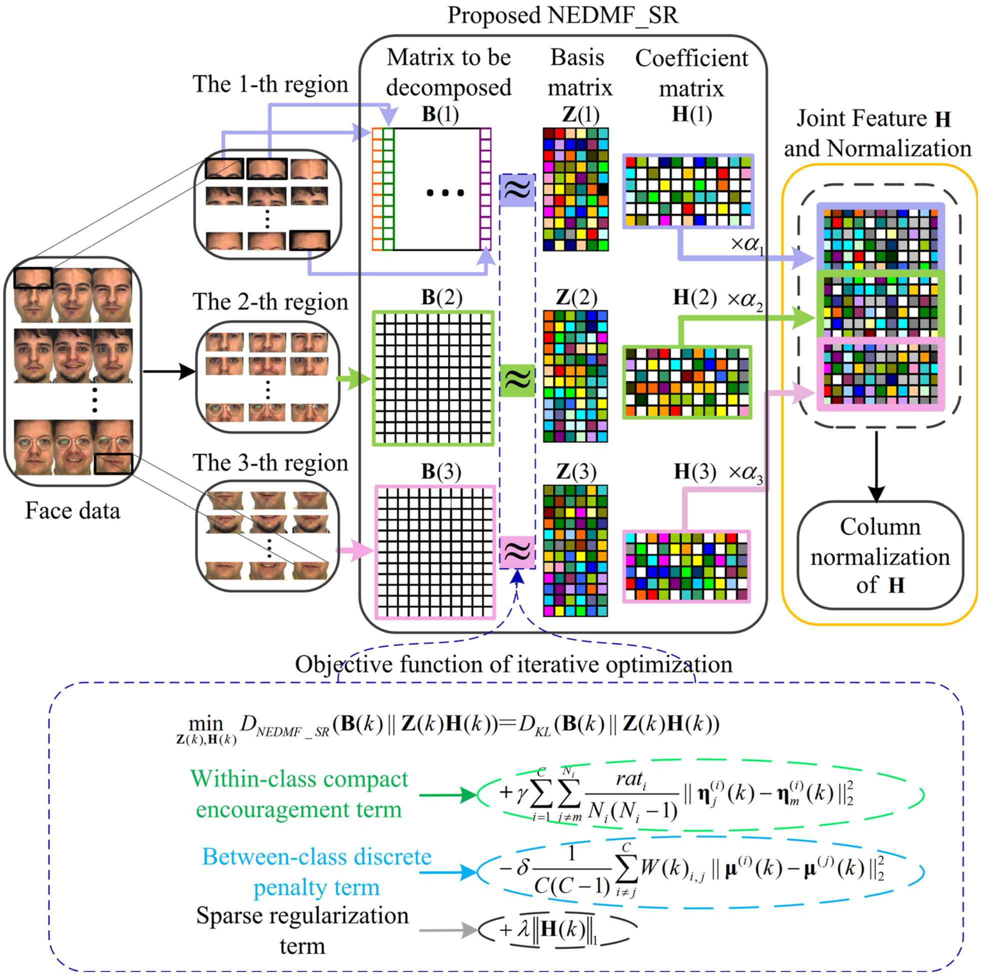

Proposed NEDMF_SR
Matrix to be decomposed B(1)
Basis matrix Z(1)
Coefficient matrix H(1)
The 1-th region
Joint Feature H and Normalization
×α1
The 2-th region
B(2)
Z(2)
H(2) ×α2
Face data
The 3-th region
B(3)
Z(3)
H(3) ×α3
Column normalization of H
Objective function of iterative optimization
$\min_{\mathbf{Z}(k),\mathbf{H}(k)} D_{NEDMF\_SR}(\mathbf{B}(k)\|\mathbf{Z}(k)\mathbf{H}(k)) = D_{KL}(\mathbf{B}(k)\|\mathbf{Z}(k)\mathbf{H}(k))$
Within-class compact encouragement term
$+\gamma\sum_{i=1}^{C}\sum_{j\neq m}^{N_i}\frac{rat_i}{N_i(N_i-1)}\|\boldsymbol{\eta}_j^{(i)}(k)-\boldsymbol{\eta}_m^{(i)}(k)\|_2^2$
Between-class discrete penalty term
$-\delta\frac{1}{C(C-1)}\sum_{i\neq j}^{C}W(k)_{i,j}\|\boldsymbol{\mu}^{(i)}(k)-\boldsymbol{\mu}^{(j)}(k)\|_2^2$
Sparse regularization term
$+\lambda\|\mathbf{H}(k)\|_1$

Fig. 1 New framework for joint feature extraction

where N_i stands for the number of samples belonging to class i.

Furthermore, the total within-class cohesion degree S_w of C classes is defined as the within-class constraint term, which can be obtained by Eq. (6):

$$S_w = \sum_{i=1}^{C} s_i = \sum_{i=1}^{C}\sum_{j\neq m}^{N_i}\frac{1}{N_i(N_i-1)}\left\|\boldsymbol{\eta}_j^{(i)} - \boldsymbol{\eta}_m^{(i)}\right\|_2^2 \tag{6}$$

2. Construction of within-class compact encouragement term

Due to the large within-class variation in the same class of images, the mean of within-class cohesion degree for C classes, i.e., $mean = \sum_{i=1}^{C} s_i \big/ C$, is used as the criterion in order to aggregate within-class samples better. The larger the value of s_i is, the smaller the cohesion degree is. Therefore, the within-class compact weight coefficient is imposed on the samples with small cohesion degree, while the samples with great cohesion degree have no constraint. Accordingly, the within-class compact weight coefficient rat_i of class i is calculated as Eq. (7).

$$rat_i = \begin{cases} \dfrac{mean}{s_i}, & s_i > mean \\ 1, & s_i \leq mean \end{cases} \tag{7}$$

Then, the proposed within-class compact encouragement term S_w' can be given by Eq. (8).

$$S_w' = \sum_{i=1}^{C}\sum_{j\neq m}^{N_i}\frac{rat_i}{N_i(N_i-1)}\left\|\boldsymbol{\eta}_j^{(i)} - \boldsymbol{\eta}_m^{(i)}\right\|_2^2 \tag{8}$$

3.3 Between-class discrete penalty term

In order to make the between-class samples more discrete in the new base space, the low-dimensional cohesion degree between the samples of any two classes is fully considered to construct the between-class constraint term. Meanwhile, to further enhance the separability of between-class samples, the distinct between-class discrete weight coefficients are imposed on the samples of different classes to form the between-class discrete penalty term. The specific construction steps are as follows:

1. Construction of between-class constraint term

Let $\boldsymbol{\mu}^{(i)}$ denote the sample mean for class i in the new base space, $\boldsymbol{\mu}^{(i)} = \sum_{j=1}^{N_i} \boldsymbol{\eta}_j^{(i)} \big/ N_i$, that is, the mean vector of coefficient vectors $\boldsymbol{\eta}_j^{(i)}$ for class i. The between-class separation degree S_b between $\boldsymbol{\mu}^{(i)}$, namely the between-class constraint term, is measured by Euclidean distance, and calculated as Eq. (9).

$$S_b = \frac{1}{C(C-1)} \sum_{i \neq j}^{C} \left\| \boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)} \right\|_2^2 \tag{9}$$

2. Construction of between-class discrete penalty term

Firstly, the similarity between sample means of any two classes is measured by Eq. (10).

$$W_{i,j} = e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2}{\sigma}} \tag{10}$$

where $W_{i,j}$ represents the similarity between class i and class j, and σ is the heat kernel parameter. $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the mean vectors of class i and class j, respectively.

Then, the different between-class discrete weight coefficients $W'_{i,j}$ are imposed on between-class samples to further separate the similar samples among different classes. The mean W_{mean} of upper triangular elements in the matrix $\mathbf{W}$ ($\mathbf{W} \in R^{C \times C}$ is a symmetric matrix) is used as the criterion, $W'_{i,j}$ can be constructed by Eq. (11).

$$W'_{i,j} = \begin{cases} W_{i,j}, & W_{i,j} \geq W_{mean} \\ 0, & W_{i,j} < W_{mean} \end{cases} \tag{11}$$

It can be shown that the larger the $W_{i,j}$ is, that is, the greater the similarity between classes i and j is, and thus, a large discrete constraint needs to be imposed on the pairwise classes, for simplicity, the value of $W'_{i,j}$ is assigned to $W_{i,j}$; otherwise, no constraint is imposed, and the value of $W'_{i,j}$ is assigned to 0. Finally, the between-class discrete penalty term S'_b can be given by Eq. (12).

$$S'_b = \frac{W'_{i,j}}{C(C-1)} \sum_{i \neq j}^{C} \left\| \boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)} \right\|_2^2 \tag{12}$$

3.4 Sparse regularization term

According to the compressed sensing and sparse coding theory [33], the underlying assumption of sparse models is that the input vectors can be reconstructed accurately as a linear combination of the dictionary atoms only with a small number of nonzero coefficients. And consequently, the sparsity enhanced regularizer $\psi(\mathbf{H})$ is used as the sparse constraint term of coefficient matrix to improve its ability of sparse representation, and this term can be defined by Eq. (13).

$$\psi(\mathbf{H}) = \lambda \|\mathbf{H}\|_1 \tag{13}$$

where λ is the sparse parameter, and $\lambda > 0$.

3.5 Objective function

According to the motivations presented in the previous section, the within-class compact encouragement term and the between-class discrete penalty term are defined, and meanwhile, the coefficient matrix is required to possess the ability of sparse representation. And consequently, the objective function of the proposed NEDMF_SR method can be given as follows:

$$\begin{cases} \min\limits_{\mathbf{Z}(k),\mathbf{H}(k)} D_{NEDMF\_SR} = \sum\limits_{j}^{n} \sum\limits_{i}^{m_k} \left(B_{i,j}(k) \left(\log \frac{B_{i,j}(k)}{(\mathbf{Z}(k)\mathbf{H}(k))_{i,j}} - 1 \right) + (\mathbf{Z}(k)\mathbf{H}(k))_{i,j} \right) \\ \qquad + \gamma \sum\limits_{i=1}^{C} \sum\limits_{j \neq m}^{N_i} \frac{rat_i}{N_i(N_i-1)} \left\| \boldsymbol{\eta}_j^{(i)}(k) - \boldsymbol{\eta}_m^{(i)}(k) \right\|_2^2 \\ \qquad - \delta \frac{1}{C(C-1)} \sum\limits_{i \neq j}^{C} W'_{i,j}(k) \left\| \boldsymbol{\mu}^{(i)}(k) - \boldsymbol{\mu}^{(j)}(k) \right\|_2^2 + \lambda \|\mathbf{H}(k)\|_1 \\ \qquad = D_{KL} + \gamma S'_w(k) - \delta S'_b(k) + \lambda \|\mathbf{H}(k)\|_1 \\ s.t.\ \mathbf{B}(k) \in R_+^{m_k \times n}, \mathbf{Z}(k) \in R_+^{m_k \times f_k}, \\ \quad \mathbf{H}(k) \in R_+^{f_k \times n}, f_k < < \min(m_k, n), k \in [1, K] \end{cases} \tag{14}$$

In practice, when dealing with image recognition problems, it is usually necessary to be blocked. Where k denotes the k-th module, $k \in [1, K]$; K represents the total number of blocks; $\mathbf{B}(k)$ is the matrix consisting of all the k-th modules; $\boldsymbol{\eta}_j^{(i)}(k) = \mathbf{H}_s(k)$, $s = N_i(i-1) + j$, $j \in [1, N_i]$, $\mathbf{H}_s(k)$ denotes the s-th column of coefficient matrix $\mathbf{H}(k)$, γ is the regulation parameter of within-class compact encouragement term, δ represents the regulation parameter of between-class discrete penalty term, λ denotes the sparseness control parameter for coefficient matrix.

3.6 Update rules

The objective function of the proposed NEDMF_SR method is non-convex with respect to both of the variables $\mathbf{Z}(k)$ and $\mathbf{H}(k)$ together, and thus, it seems to be impossible to search the global minimum of D_{NEDMF_SR}, and consequently, the iteration rule is utilized to obtain the local minimum. Let G be the auxiliary function for D_{NEDMF_SR} and is defined by Eq. (15).

$$
\begin{aligned}
G(\mathbf{H}(k),\mathbf{H}^{(t)}(k)) &= \sum_{i,j}\left(B_{i,j}(k)\log B_{i,j}(k) - B_{i,j}(k)\right)\\
&-\sum_{i,j}B_{i,j}(k)\sum_{m}a_m\left(\log(Z_{i,m}(k)H_{m,j}(k)) - \log a_m\right)\\
&+\sum_{i,j,m}Z_{i,m}(k)H_{m,j}(k)+\gamma\sum_{i=1}^{C}\sum_{j\neq m}^{N_i}\frac{rat_i}{N_i(N_i-1)}\left\|\boldsymbol{\eta}_j^{(i)}(k)-\boldsymbol{\eta}_m^{(i)}(k)\right\|_2^2\\
&-\frac{\delta}{C(C-1)}\sum_{i\neq j}^{C}W'_{i,j}(k)\left\|\boldsymbol{\mu}^{(i)}(k)-\boldsymbol{\mu}^{(j)}(k)\right\|_2^2+\lambda\sum_{i,j}H_{i,j}(k)
\end{aligned}
\tag{15}
$$

where $a_m = Z_{i,m}(k)H^{(t)}_{m,j}(k)\Big/\sum_m Z_{i,m}(k)H^{(t)}_{m,j}(k)$.

In order to search for the local minimum of D_{NEDMF_SR}, the partial derivative of each term in auxiliary function G with respect to $\mathbf{H}(k)$ is calculated, respectively, and let $\partial G(H(k),H^{(t)}(k))/\partial H_{m,l}(k)=0$. Let $H_{m,l}(k)$ denote the m-th element of the ρ-th sample vector of class r, and thus $H_{m,l}(k)=\eta^{(r)}_{\rho,m}(k)$. To simplify the calculation, the partial derivative of the fourth term in Eq. (15) is firstly calculated by Eq. (16).

$$
\begin{aligned}
\frac{\partial S'_w(k)}{\partial\eta^{(r)}_{\rho,m}(k)} &= \frac{\partial\sum_{i=1}^{C}\sum_{j\neq l}^{N_i}\frac{rat_i}{N_i(N_i-1)}\left\|\boldsymbol{\eta}_j^{(i)}(k)-\boldsymbol{\eta}_l^{(i)}(k)\right\|_2^2}{\partial\eta^{(r)}_{\rho,m}(k)}\\
&=\frac{\partial\sum_n\sum_{i=1}^{C}\sum_{j\neq l}^{N_i}\frac{rat_i}{N_i(N_i-1)}\left(\eta^{(i)}_{j,n}(k)-\eta^{(i)}_{l,n}(k)\right)^2}{\partial\eta^{(r)}_{\rho,m}(k)}\\
&=-\sum_{j=1}^{N_r}\frac{rat_r}{N_r(N_r-1)}2\left(\eta^{(r)}_{j,m}(k)-\eta^{(r)}_{\rho,m}(k)\right)\\
&\quad+\sum_{l=1}^{N_r}\frac{rat_r}{N_r(N_r-1)}2\left(\eta^{(r)}_{\rho,m}(k)-\eta^{(r)}_{l,m}(k)\right)\\
&=4\sum_{l=1}^{N_r}\frac{rat_r}{N_r(N_r-1)}\eta^{(r)}_{\rho,m}(k)-4\sum_{j=1}^{N_r}\frac{rat_r}{N_r(N_r-1)}\eta^{(r)}_{j,m}(k)\\
&=\frac{4rat_r}{N_r-1}\left(\eta^{(r)}_{\rho,m}(k)-\mu^{(r)}_m(k)\right)
\end{aligned}
\tag{16}
$$

Then, the partial derivative of the fifth term in Eq. (15) is calculated by Eq. (17).

$$
\begin{aligned}
\frac{\partial S'_b(k)}{\partial\eta^{(r)}_{\rho,m}(k)} &= \frac{\delta}{C(C-1)}\frac{\partial\sum_{i\neq j}^{C}W'_{i,j}(k)\left\|\boldsymbol{\mu}^{(i)}(k)-\boldsymbol{\mu}^{(j)}(k)\right\|_2^2}{\partial\eta^{(r)}_{\rho,m}(k)}\\
&=\frac{\delta}{C(C-1)}\frac{\partial\sum_n\sum_{i\neq j}^{C}\left(\mu_n^{(i)}(k)-\mu_n^{(j)}(k)\right)^2W'_{i,j}(k)}{\partial\eta^{(r)}_{\rho,m}(k)}\\
&=\frac{\delta}{C(C-1)}\left(\sum_{i\neq r}^{C}2\left(\mu_m^{(i)}(k)-\mu_m^{(r)}(k)\right)\frac{-1}{N_r}W'_{i,r}(k)\right.\\
&\quad\left.+\sum_{j\neq r}^{C}2\left(\mu_m^{(r)}(k)-\mu_m^{(j)}(k)\right)\frac{1}{N_r}W'_{r,j}(k)\right)\\
&=\frac{4\delta}{N_rC(C-1)}\sum_{i\neq r}^{C}\left(\mu_m^{(r)}(k)-\mu_m^{(i)}(k)\right)W'_{i,r}(k)
\end{aligned}
\tag{17}
$$

Integrating the above two terms, Eq. (18) is derived by setting $\partial G(\mathbf{H}(k),\mathbf{H}^{(t)}(k))/\partial H_{m,l}(k)=0$.

$$
\begin{aligned}
\frac{\partial G(\mathbf{H}(k),\mathbf{H}^{(t)}(k))}{\partial H_{m,l}(k)} &= -\sum\nolimits_i B_{i,l}(k)\frac{Z_{i,m}(k)H^{(t)}_{m,l}(k)}{\sum_n Z_{i,n}(k)H^{(t)}_{n,l}(k)}\frac{1}{H_{m,l}(k)}\\
&+\sum\nolimits_i Z_{i,m}(k)+\frac{4rat_r\gamma}{N_r-1}\left(H_{m,l}(k)-\mu_m^{(r)}(k)\right)\\
&-\frac{4\delta W'_{i,r}(k)}{N_rC(C-1)}\sum_{i\neq r}^{C}\left(\mu_m^{(r)}(k)-\mu_m^{(i)}(k)\right)+\lambda=0
\end{aligned}
\tag{18}
$$

It can be seen that Eq. (18) is a quadratic equation of $H_{m,l}(k)$, and then, both sides of the equation are simultaneously multiplied by $H_{m,l}(k)$ to obtain the following expanded equation.

$$
\begin{aligned}
&-\sum\nolimits_i B_{i,l}(k)\frac{Z_{i,m}(k)H^{(t)}_{m,l}(k)}{\sum_n Z_{i,n}(k)H^{(t)}_{n,l}(k)}\\
&+\left((\lambda+1)-\left(\frac{4rat_r\gamma}{N_r-1}+\frac{4\delta}{N_rC(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\right)\right.\\
&\times\frac{1}{N_r}\sum\nolimits_{\alpha\neq l}H_{m,\alpha}(k)\\
&\left.+\frac{4\delta}{N_rC(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\mu_m^{(i)}(k)\right)H_{m,l}(k)\\
&+\left(\frac{4rat_r\gamma}{N_r-1}-\left(\frac{4rat_r\gamma}{N_r-1}+\frac{4\delta}{N_rC(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\right)\frac{1}{N_r}\right)\\
&H^2_{m,l}(k)=0
\end{aligned}
\tag{19}
$$

By solving the above quadratic equation, the following iterative rule can be derived as shown in Eq. (20).

$$
\begin{aligned}
&H^{(t+1)}_{m,l}(k)\\
&=\frac{T+\sqrt{T^2+16\left(\frac{rat_r\gamma}{N_r}-\frac{\delta}{N_r^2C(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\right)\sum_i B_{i,l}(k)\frac{Z^{(t)}_{i,m}(k)H^{(t)}_{m,l}(k)}{\sum_n Z^{(t)}_{i,n}(k)H^{(t)}_{n,l}(k)}}}{8\left(\frac{rat_r\gamma}{N_r}-\frac{\delta}{N_r^2C(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\right)}
\end{aligned}
\tag{20}
$$

where $T=\left(\frac{4rat_r\gamma}{N_r-1}+\frac{4\delta}{N_rC(C-1)}\sum_{i\neq r}^{C}W'_{i,r}(k)\right)\frac{1}{N_r}\sum_{\alpha\neq l}H_{m,\alpha}(k)$
$-\frac{4\delta\sum_{i\neq r}^{C}W'_{i,r}(k)\mu_m^{(i)}(k)}{N_rC(C-1)}-(\lambda+1).$

The iteration rule for base matrix is the same as in NMF [8] and is shown as in Eq. (21).

$$Z_{i,m}^{(t+1)}(k) = Z_{i,m}^{(t)}(k)\frac{\sum_j H_{m,j}^{(t+1)}(k)B_{i,j}(k)\Big/\sum_n Z_{i,n}^{(t)}(k)H_{n,j}^{(t+1)}(k)}{\sum_j H_{m,j}^{(t+1)}(k)} \tag{21}$$

Regarding the iteration rules of Eqs. (20) and (21), the following Theorem 1 holds. Theorem 1 shows that the update rules of $\mathbf{Z}(k)$ and $\mathbf{H}(k)$ will converge to the local optimum eventually. Note: A detailed proof of Theorem 1 is given in the "Appendix".

Theorem 1 *The objective function* $D_{NEDMF\_SR}(\mathrm{B}(k)||\mathbf{Z}(k)\mathbf{H}(k))$ *in Eq.* (14) *is nonincreasing under the update rules in Eqs.* (20) and (21).

3.7 NEDMF_SR for image classification and recognition

In this section, the proposed NEDMF_SR method is utilized to deal with the practical problems of occluded image classification and recognition. In fact, these problems could be effectively handled by applying the modular representation approach [34]. In the proposed modular schemes, the occluded image is equally divided into many modules, each of which is processed independently, and the information from all the modules is further fused to make the final determination. On this basis, a quite simple and efficient fusion strategy is introduced, which significantly weaken the influence of occluded modules and improve the image recognition accuracy. This strategy constructs two kinds of classifiers, including the local classifier and the global classifier, and the specific construction steps are as follows.

1. Construction of local classifier

   (a) Modular processing of images. For the training dataset, all the images of size $\varsigma \times \tau$ are equally divided into K non-overlapping modules and are expressed as K corresponding matrices. Then, the k-th matrix of the j-th image is transposed and arranged sequentially column by column to form the m_k-dimensional column vector $\mathbf{b}_j(k)$, where $m_k = (\varsigma \times \tau)/K$ and $k = 1, 2, \ldots, K$.

   (b) Construction of modular matrix. All the column vectors of the k-th module corresponding to n images are arranged by column to form the training matrix $\mathbf{B}(k) = [\mathbf{b}_1(k), \mathbf{b}_2(k), \ldots, \mathbf{b}_n(k)]$ of the k-th module. The same procedure is conducted on the images in testing set to obtain matrix $\mathbf{V}(k) = [\mathbf{v}_1(k), \mathbf{v}_2(k), \ldots, \mathbf{v}_g(k)]$, where g is the number of testing samples.

   (c) Projection processing. The projection coefficient vector $\hat{\mathbf{h}}_j(k) \in R_+^{f_k}$ of $\mathbf{v}_j(k)$ can be obtained by projecting the testing data $\mathbf{v}_j(k) \in R_+^{m_k}$, $j = 1, 2, \ldots, g$ on the base matrix $\mathbf{Z}(k)$, and the projection way is defined as $\hat{\mathbf{h}}_j(k) = (\mathbf{Z}(k))^{\dagger}\mathbf{v}_j(k) = \left((\mathbf{Z}(k))^T\mathbf{Z}(k)\right)^{-1}(\mathbf{Z}(k))^T\mathbf{v}_j(k)$, where † denotes the pseudo inversion of matrix.

   (d) Design of measure criteria. K nearest neighbor (NN) classifiers are constructed as the local classifiers, and the measure criterion of each classifier is defined as follows:

$$\begin{aligned}
\text{Classifier 1}: d_{j\xi}^1 &= \left\|\hat{\mathbf{h}}_j(1) - \mathbf{h}_\xi(1)\right\|_2 \\
\text{Classifier 2}: d_{j\xi}^2 &= \left\|\hat{\mathbf{h}}_j(2) - \mathbf{h}_\xi(2)\right\|_2 \\
&\vdots \\
\text{Classifier } K: d_{j\xi}^K &= \left\|\hat{\mathbf{h}}_j(K) - \mathbf{h}_\xi(K)\right\|_2
\end{aligned} \tag{22}$$

   where $d_{j\xi}^i$ denotes the low-dimensional Euclidean distance between the i-th module of testing sample $\mathbf{v}_j$ and of the training sample $\mathbf{v}_\xi$.

2. Construction of global classifier

In fact, the contribution degree of each local feature to global recognition is different, and the complementary information among all local features is neglected in the training process of each local classifier. It is inevitable to get low recognition accuracy if the local classifiers are directly used for image classification. Therefore, it is indispensable to further construct the global classifier. The different contribution degree of each module feature to recognition is fully considered, and then, the local classifiers are combined in the linearly weighted way and assigned to the different weight coefficients based on discriminability, which weakens the role of occluded modules effectively, and consequently, the classification accuracy is improved. The linear weight coefficient α_k of the k - th local classifier can be calculated by Eq. (23).

$$\alpha_k = \frac{n_k}{\sum_{j=1}^K n_j} \tag{23}$$

where n_k denotes the recognition accuracy of the k - th local classifier. The discriminability of classifier can be measured by α_k. The greater the α_k is, the better the discriminability is.

The measure criterion of global classifier is defined as Eq. (24).

$$
\begin{aligned}
d_{j\xi} &= \alpha_1 \left\|\hat{\mathbf{h}}_j(1) - \mathbf{h}_\xi(1)\right\|_2 + \alpha_2 \left\|\hat{\mathbf{h}}_j(2) - \mathbf{h}_\xi(2)\right\|_2 \\
&\quad + \cdots + \alpha_K \left\|\hat{\mathbf{h}}_j(K) - \mathbf{h}_\xi(K)\right\|_2 \\
&= \left\| \begin{pmatrix} \alpha_1 \hat{\mathbf{h}}_j(1) \\ \alpha_2 \hat{\mathbf{h}}_j(2) \\ \vdots \\ \alpha_K \hat{\mathbf{h}}_j(K) \end{pmatrix} - \begin{pmatrix} \alpha_1 \mathbf{h}_\xi(1) \\ \alpha_2 \mathbf{h}_\xi(2) \\ \vdots \\ \alpha_K \mathbf{h}_\xi(K) \end{pmatrix} \right\|_2 = \left\|\hat{\mathbf{h}}_j - \mathbf{h}_\xi\right\|_2 \qquad (24) \\
&s.t. \quad \alpha_1, \alpha_2, \ldots, \alpha_K \in [0,1], \ \sum\nolimits_{i=1}^{K} \alpha_i = 1
\end{aligned}
$$

where $d_{j\xi}$ denotes the low-dimensional Euclidean distance between the testing sample $\mathbf{v}_j$ and the training sample $\mathbf{v}_\xi$. It is observed that the global classifier takes full advantage of the correlation and information complementarity among all modules, which weakens the adverse effects of continuous occlusions on fusion recognition, and makes the NEDMF_SR possess better adaptability and strong robustness to the occluded image classification and recognition.

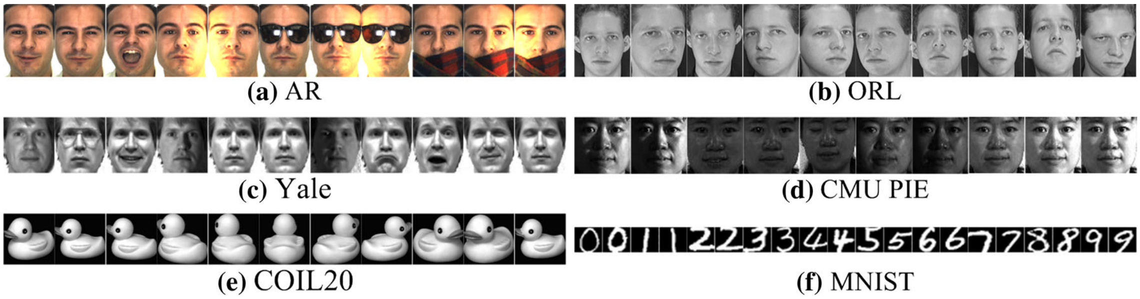
Algorithm 1 Image recognition based on NEDMF_SR

```
Input: Training modular matrices $\mathbf{B}(k) \in R_+^{m_k \times n}$, testing modular matrices $\mathbf{V}(k)$, $k = 1, 2, \ldots, K$.
Factorize $\mathbf{B}(k)$ to obtain $\mathbf{Z}(k) \in R_+^{m_k \times f_k}$ and $\mathbf{H}(k) \in R_+^{f_k \times n}$ by NEDMF_SR.
Calculate the projection coefficient vector $\hat{\mathbf{h}}_j(k) = \left((\mathbf{Z}(k))^T \mathbf{Z}(k)\right)^{-1} (\mathbf{Z}(k))^T \mathbf{v}_j(k)$.
Calculate the recognition accuracies of local classifiers.
 for $k = 1:K$
    for $j = 1:g$
    1) Calculate the distances between $\mathbf{v}_j(k)$ and the $k$-th module of any training sample: $d_{j1}^k, \ldots, d_{j\xi}^k, \ldots, d_{jn}^k$, and
       compose vector $\mathbf{d}^k = [d_{j1}^k, \ldots, d_{j\xi}^k, \ldots, d_{jn}^k]$.
    2) Search for the subscript $\xi$ of training sample corresponding to the minimum in $\mathbf{d}^k$: $\xi = \arg\min_\xi \mathbf{d}^k$.
    3) Determine the label of testing sample $\mathbf{v}_j$: $\text{identity}(\mathbf{v}_j) = \text{identity}(\mathbf{b}_\xi)$.
    end.
    4) Obtain the recognition accuracy $n_k$ of the $k$-th local classifier.
 end.
Obtain the recognition accuracy vector $\mathbf{n} = [n_1, n_2, \ldots, n_K]$ and the normalized vector $\mathbf{n}_{norm} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$.
 Determine the final class label of $\mathbf{v}_j$.
1) Calculate the distances between $\mathbf{v}_j(k)$ and any training sample: $d_{j1}, \ldots, d_{j\xi}, \ldots, d_{jn}$, and compose vector
   $\mathbf{d} = [d_{j1}, \ldots, d_{j\xi}, \ldots, d_{jn}]$.
2) Search for the subscript $\xi$ of training sample corresponding to the minimum in $\mathbf{d}$: $\xi = \arg\min_\xi \mathbf{d}$.
Output: Labels of testing samples $\mathbf{v}_j$: $\text{identity}(\mathbf{v}_j) = \text{identity}(\mathbf{b}_\xi)$, $j = 1, 2, \ldots, g$.
```

3. **Algorithm 1** Image recognition based on NEDMF_SR

4 Experiments and analysis

In this section, by comparing with nine representative algorithms on six different types of image databases, the effectiveness of the proposed NEDMF_SR method is evaluated. And five experiments including parameter selection, weight coefficient selection, convergence study, basis image visualization, and occluded image recognition



(a) AR

(b) ORL

(c) Yale

(d) CMU PIE

(e) COIL20

(f) MNIST

Fig. 2 Example images of 6 databases

are conducted. Firstly, the database introduction is given as follows, and then, the experimental results are presented.

4.1 Databases introduction

Six common databases selected for our experiments include AR [35], ORL [36], Yale [37], CMU PIE [38], COIL20 [25], and MNIST databases (http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html). Figure 2 shows the example images of these databases.

1. AR face image database: The database includes over 4000 facial images with the resolution of 768 × 576, corresponding to 126 individuals (70 males and 56 females). All these images feature frontal view faces with no restriction on facial expressions, illuminations, facial disguises (sun glasses or scarf), make up, hair styles, etc. In the experiments of this paper, a subset of AR face image database provided by Martinez is chosen, which consists of 2600 images corresponding to the faces of 100 individuals (50 men and 50 women). For each subject, 26 images were taken with different illuminations, facial expressions, and occlusions, as shown in Fig. 2a.
2. ORL face image database: The database contains 400 facial images of 40 individuals (10 samples per subject). The facial images of some individuals vary in times, lighting, facial expressions (open or closed eyes and smiling or not smiling) and facial disguises (with or without glasses), as shown in Fig. 2b.
3. Yale face image database: The database is composed of 165 gray scale images corresponding to 15 adults. For each individual, 11 images were taken with varying illumination conditions and facial expressions (normal, happy, sad, surprised, sleepy, and winking), as shown in Fig. 2c.
4. CMU PIE face image database: The database is composed of 41,368 color, face images of 640 × 480 pixels in size captured from 68 subjects across 13 different head poses, under 43 varying illumination conditions, and with four distinct expressions. Since the experiment aims to demonstrate the effects of illumination changes, frontal images chosen from the illumination subset in the CMU PIE dataset are utilized for further analyses. This subset consists of 49 images per individual taken under different illuminations, which is shown in Fig. 2d.
5. COIL20 image database: The database contains 20 objects and each object contains 72 images, which were taken at different degrees with intervals of 5°, as shown in Fig. 2e.
6. MNIST image database: The database is comprised of totally 4000 digit images from 0 to 9 and is equally divided into training set and testing set. Each image is 28 × 28 pixels in size, as shown in Fig. 2f.

In order to decrease the memory consumption, and improve the computational efficiency, each image in AR, ORL, Yale, CMU PIE, and COIL20 databases is uniformly resized to obtain a 256 gray-level image of size 60 × 60 pixels before experiments, while images in MNIST database are resized to 256 gray-level images of size 30 × 30 pixels because of the relatively low resolution. In the experiments, the nearest neighbor (NN) classifiers are adopted for both the local classifier and the global classifier, and the image recognition accuracy is calculated by Eq. (25). Moreover, image classification accuracies are also calculated with different proportions of the labeled data and modular schemes.

$$\mathrm{RecAccuracy} = \frac{Cor}{g} \times 100\% \tag{25}$$

where *Cor* denotes the number of correctly classified testing samples, and *g* denotes the total number of testing samples.

In the following experiments of Sects. 4.2 to 4.5, for the sake of facilitating understanding, only the case that each image is equally divided into three blocks is used as an example to show the experimental results for the optimal selection of the parameters γ, δ, and λ. In fact, the experiments with multiple image modular schemes and a subsequent series of experiments on occluded recognition fully show that optimal parameter selection results with the trisection partition scheme could be applied to other modular schemes as well. Moreover, the experimental results for five kinds of modular schemes are shown in the following Sects. 4.6 and 4.7. All the experiments of this paper are repeated 10 times independently, and the corresponding average values are recorded as the final results. Besides, the software environment of experiments is Matlab R2012a.

4.2 Experiments on parameter selection

In this section, the selections of γ, δ, and λ are investigated to indicate their importance for image recognition. The setting of relevant experiment parameters is as follows: γ and δ are set by searching the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1\ , 10\}$, λ searches the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1\ , 10\ , 10^{2}\}$. The dimension of non-negative matrix factorization for AR and CMU PIE databases is set as $f = 60$, while $f = 15$ for the three databases ORL, Yale, and COIL20, and $f = 10$ for MNIST database. One of the modular schemes, that is, each image is equally divided into three non-overlapping modules from top to bottom, is used as an example to show the optimization results of parameters.

The partitions of relative experiment databases are as follows. (1) AR database: For each subject, 6 non-occluded images are randomly selected as the training set; 6 images occluded by sunglasses are taken as the eye testing set, which is recorded as AR eye, as shown in Fig. 3a; and 6 images occluded by scarves are taken as the mouth testing set, which is recorded as AR mouth, as shown in Fig. 3b; (2) ORL, Yale, and COIL20 databases: 6 images of each subject are randomly chosen as the training set, and the rest images with 30% occlusions added artificially are taken as the testing set, as shown in Fig. 3c–e, respectively; (3) CMU PIE database: 24 images of each subject are randomly selected as the training set, and the rest images with 30% occlusions added artificially are taken as the testing set, as shown in Fig. 3f; (4) MNIST database: 40 images of each class are randomly selected as the training set, and the images with 30% occlusions added artificially in the original testing set are taken as the testing set, as shown in Fig. 3g.

4.2.1 Parameter selection for γ and δ

In this section, the selections of γ and δ used in Sect. 3.4 are investigated to indicate their importance for image recognition. Specifically, λ is set as 0.1, while δ and γ are set by searching the grid $\{10^{-3}, 10^{-2}, 10^{-1},\ 1\ ,\ 10\}$. Figure 4 shows the occluded image classification accuracy of the proposed NEDMF_SR on different databases.

It can be seen from Fig. 4 that the image recognition accuracies with the proposed NEDMF_SR remain to be comparatively stable in regard to $\delta \in [0.001\ ,\ 0.1]$ and $\gamma \in [0.1\ ,\ 10]$. Furthermore, the ranges of δ and γ are slightly different among various databases and could be optimized according to actual situation.

4.2.2 Parameter selection for λ

In this section, the selection of λ used in Sect. 3.4 is evaluated to indicate its importance for image recognition. Specifically, δ and γ are fixed at 0.001 and 0.1, and λ is set by searching the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$. Figure 5 shows the occluded image classification accuracies of the proposed NEDMF_SR method on different databases.

It can be observed from Fig. 5 that the image recognition accuracies of the proposed NEDMF_SR are relatively stable with respect to $\lambda \in [0.001\ ,\ 1]$, and consequently, λ is fixed at 0.1 in all the subsequent experiments of this paper.

4.3 Experiments on convergence study

In this paper, the iterative update rules are adopted to determine the local optimum of objective function for the NEDMF_SR, and the convergence proof of iteration rules is given in the "Appendix". Here, the convergence of the

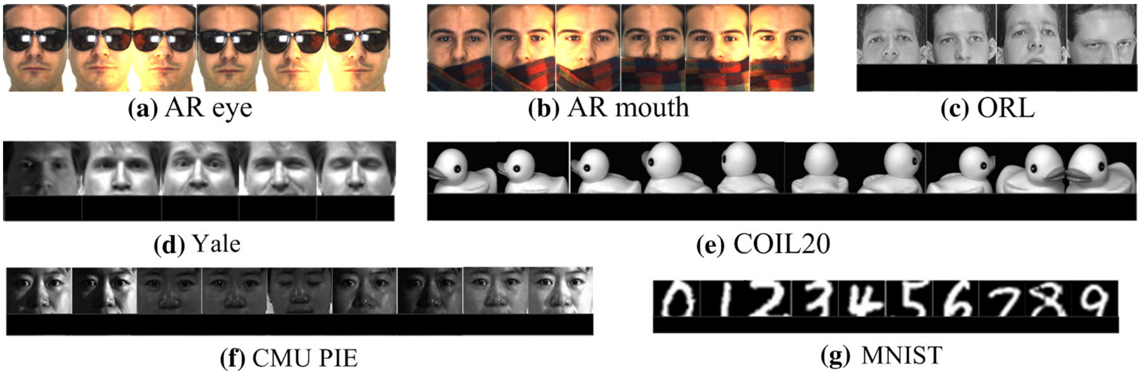
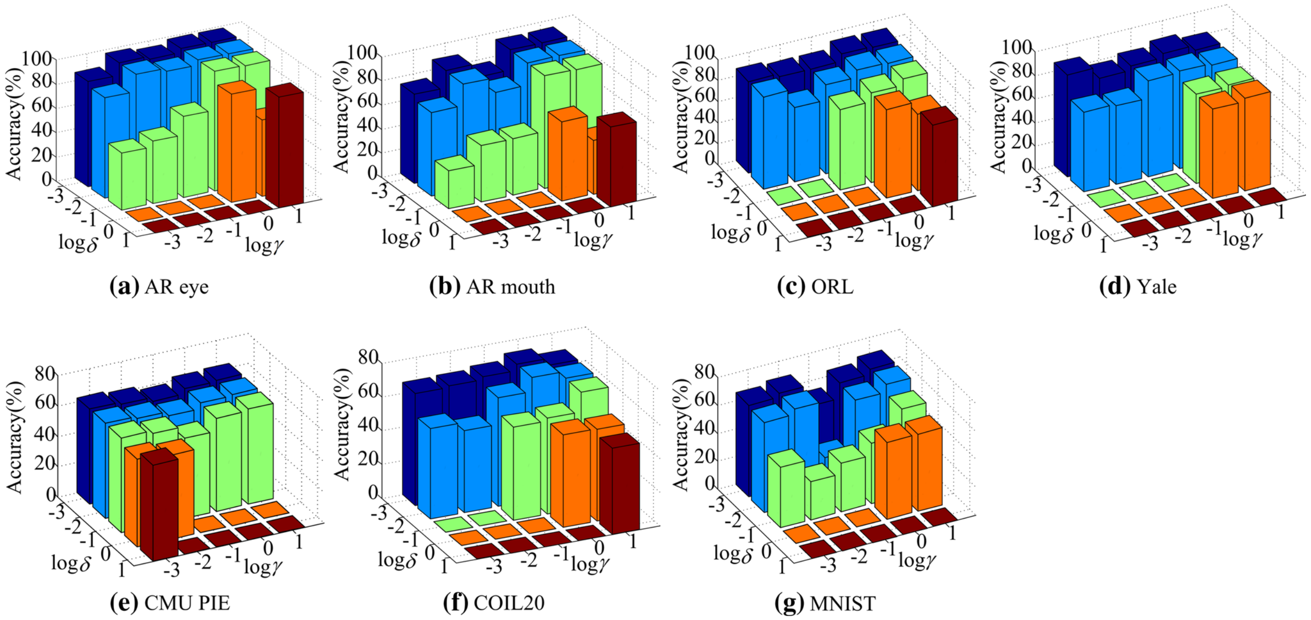

(a) AR eye
(b) AR mouth
(c) ORL
(d) Yale
(e) COIL20
(f) CMU PIE
(g) MNIST

Fig. 3 Examples of testing sets for 6 databases



(a) AR eye **(b)** AR mouth **(c)** ORL **(d)** Yale

(e) CMU PIE **(f)** COIL20 **(g)** MNIST

Fig. 4 Recognition accuracies of NEDMF_SR versus δ and γ on different databases

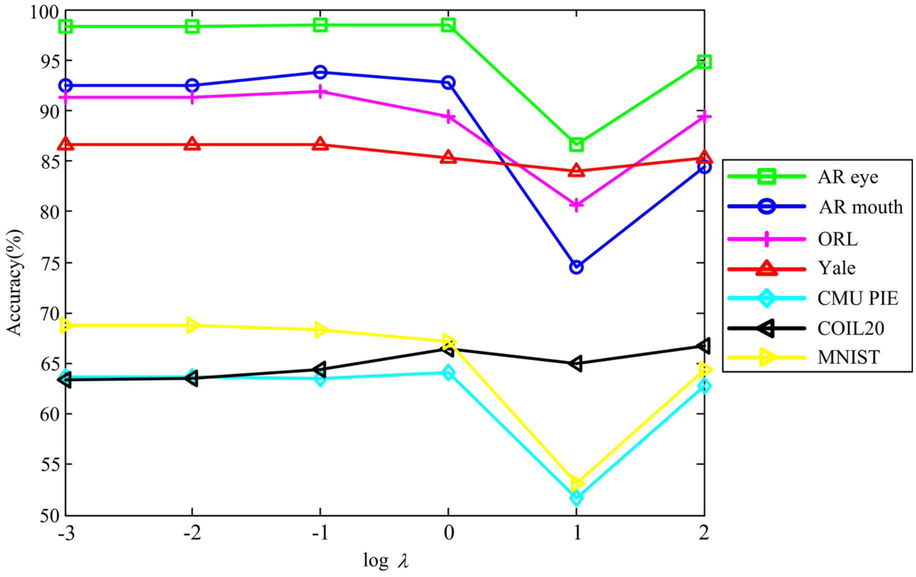


Fig. 5 Recognition accuracy of NEDMF_SR versus λ

proposed NEDMF_SR method is shown through experiment, and the convergence speed comparison between the original NMF and the proposed NEDMF_SR is given.

Figure 6 shows the convergence of each module obtained by trisecting the original image, where the x-axis denotes the iteration number and the y-axis denotes the value of objective function.

It can be seen from Fig. 6 that the proposed NEDMF_SR method has already converged within 20 iterations on ORL, COIL20, and MNNIST databases and 40 iterations on AR database. However, the original NMF will not converge until 200 iterations on all the four databases. In conclusion, the convergence speed of the NEDMF_SR method is superior to that of NMF method.

4.4 Experiments on base visualization

The training set is used to learn base applied for base visualization, and the setting of which is the same as in Sect. 4.2. Figures 7, 8, 9 and 10 present the resulting feature base components of NMF [8], DNMF [20], NDMF [23], SKNMF [24], GDNMF [26], LWSG_NMF [30],

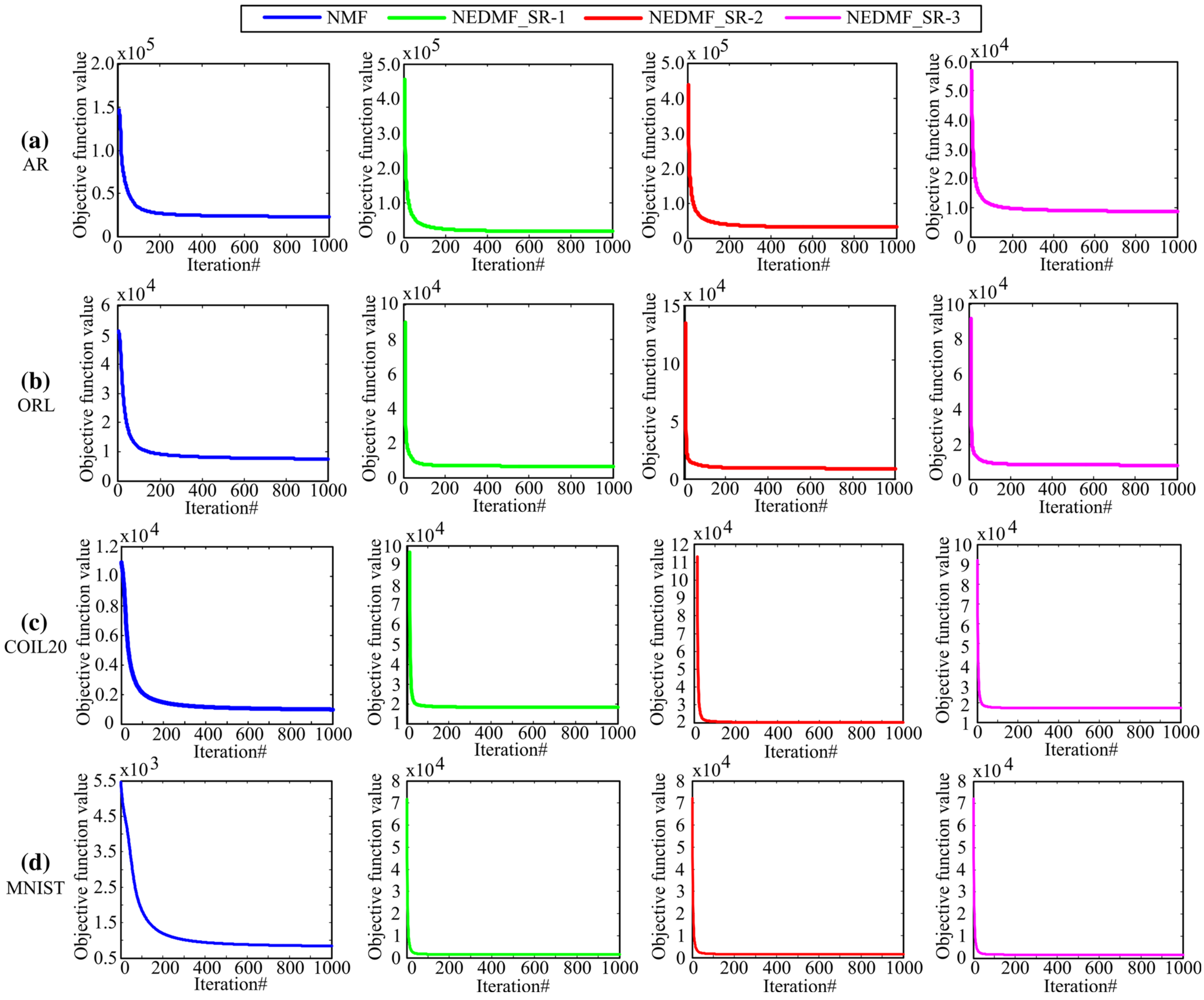


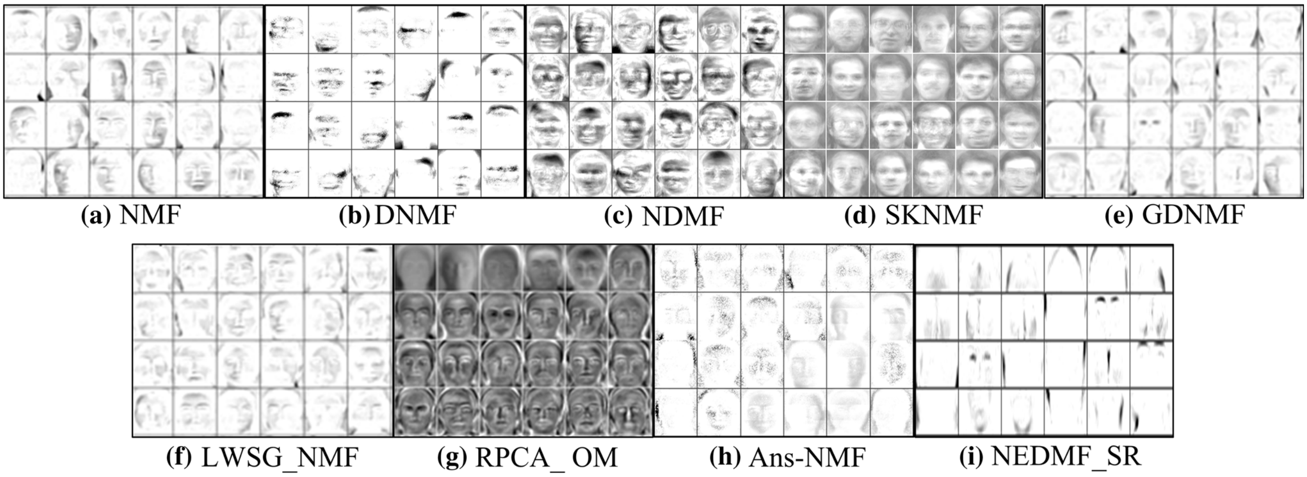
Fig. 6 Convergence of NEDMF_SR-1, NEDMF_SR-2, NEDMF_SR-3, and NMF on different databases. NEDMF_SR-i denotes that the proposed NEDMF_SR algorithm is applied on the i-th module

RPCA_OM [39], Ans-NMF [18], and the proposed NEDMF_SR for 24-dimensional subspace. Here, the contrastive method RPCA_OM introduces $\ell_{2,1}$-norm and mean matrix into the objective function, which enhances the robustness to occlusion and improves the face recognition rate. In this experiment, the parameter γ is suggested with the scale of $m^{1/2}$ (m is the dimension of matrix $\mathbf{Z}$). Meanwhile, RPCA_OM is firstly utilized to reduce the dimension by keeping 95% data energy, and then, the first 24 basis images are selected to conduct visualization.

From the basis images learned from different methods, it can be found that: (1) The bases of NMF, DNMF, NDMF, SKNMF, GDNMF, LWSG_NMF, RPCA_OM, and Ans-NMF are all less sparse than those of the proposed NEDMF_SR; (2) The NEDMF_SR is capable of learning localized regions and possesses higher parts-based learning ability. The main reason lies in that NEDMF_SR applies modular representation approach and adds ℓ_1-norm into the objective function, which can obtain stronger parts-based representation ability, and meanwhile learn discriminant characteristics of each image region (such as hairline and eyes) independently.

4.5 Experiments of weight coefficient learning

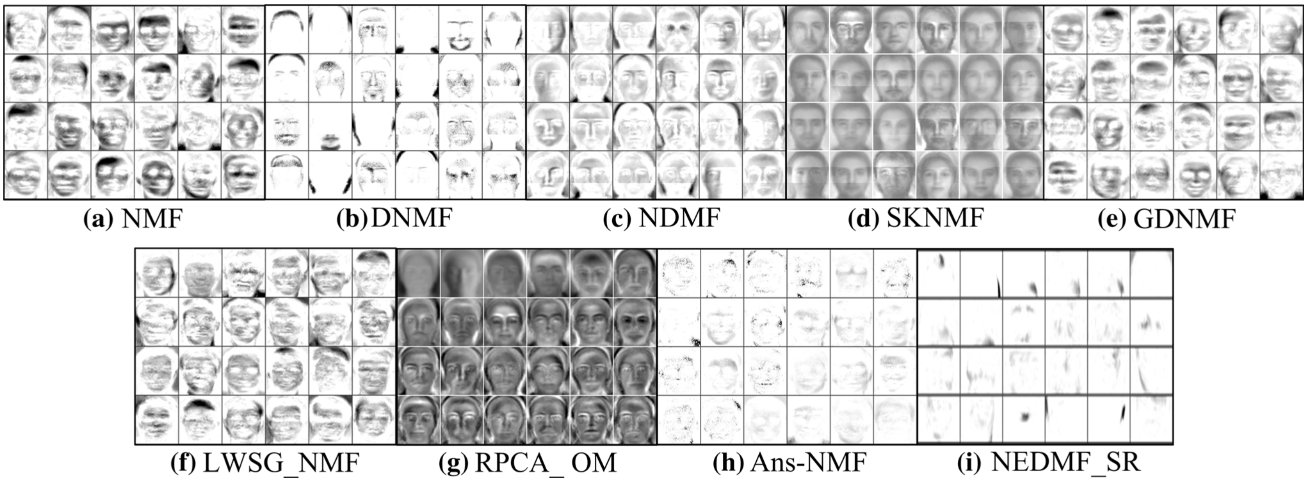
The theoretical analysis of weight coefficients for global classifier is conducted in Sect. 3.6, and the optimization analysis of them is mainly presented in this section. The experiments are performed on six databases, respectively, and the experiment setups are the same as those in Sect. 4.1. Taking one of the modular schemes that each image is equally divided into three non-overlapping local



(a) NMF **(b)** DNMF **(c)** NDMF **(d)** SKNMF **(e)** GDNMF

(f) LWSG_NMF **(g)** RPCA_ OM **(h)** Ans-NMF **(i)** NEDMF_SR

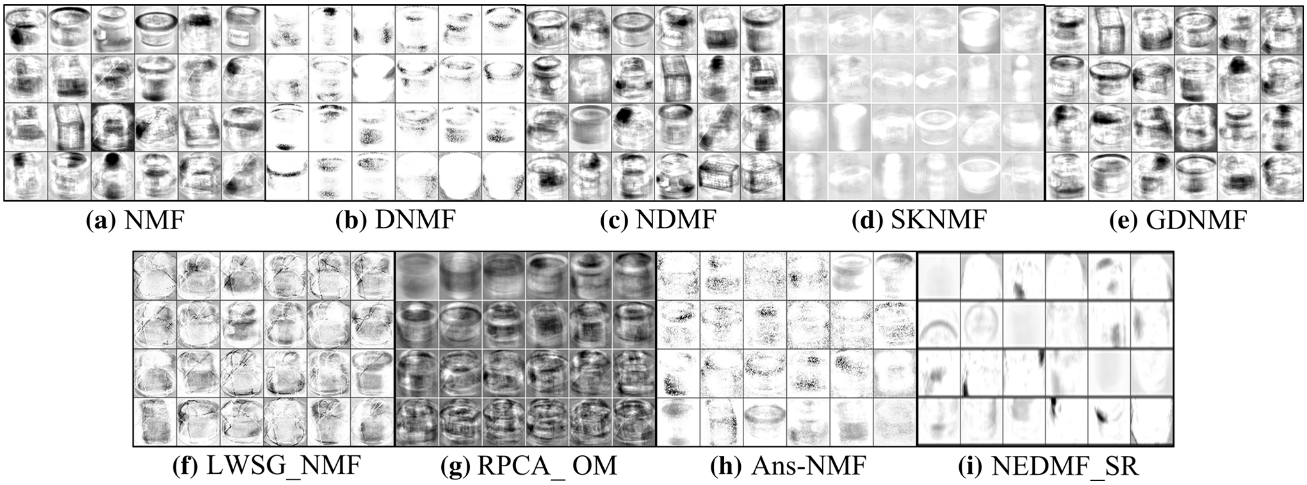
Fig. 7 Base vectors learned from AR database



(a) NMF **(b)** DNMF **(c)** NDMF **(d)** SKNMF **(e)** GDNMF

(f) LWSG_NMF **(g)** RPCA_ OM **(h)** Ans-NMF **(i)** NEDMF_SR

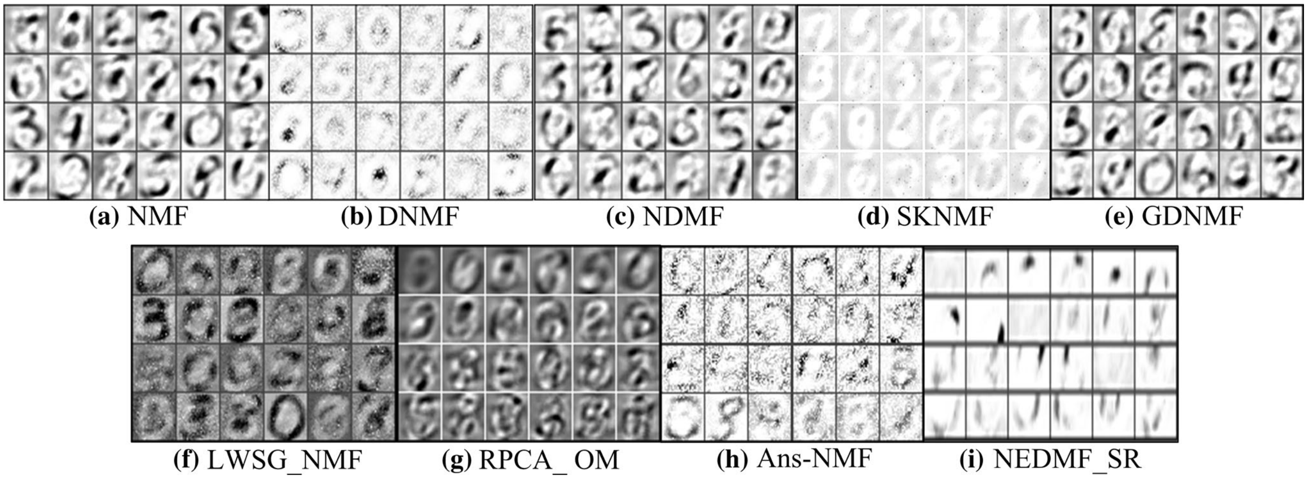
Fig. 8 Base vectors learned from ORL database



(a) NMF **(b)** DNMF **(c)** NDMF **(d)** SKNMF **(e)** GDNMF

(f) LWSG_NMF **(g)** RPCA_ OM **(h)** Ans-NMF **(i)** NEDMF_SR

Fig. 9 Base vectors learned from COIL20 database


(a) NMF (b) DNMF (c) NDMF (d) SKNMF (e) GDNMF

(f) LWSG_NMF (g) RPCA_ OM (h) Ans-NMF (i) NEDMF_SR

Fig. 10 Base vectors learned from MNIST database

Table 1 Results of weight coefficient learning with NEDMF_SR on AR database (%)

The *k*-th module, $k = 1, 2, 3$	Sunglasses occlusion			Scarf occlusion		
	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy
1	96.00	0.46	98.50	87.00	0.58	93.83
2	27.83	0.13		61.67	0.41	
3	84.33	0.41		2.17	0.01	

Table 2 Results of weight coefficient learning with NEDMF_SR on ORL and Yale databases (%)

The *k*-th module, $k = 1, 2, 3$	ORL database			Yale database		
	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy
1	84.37	0.52	91.87	81.33	0.46	86.67
2	76.87	0.47		85.33	0.48	
3	1.87	0.01		10.67	0.06	

modules from top to bottom as an example, Tables 1, 2, 3, and 4 show the learning results.

Actually, for frontal face recognition, the hairline, facial outline, and mouth are more important than nose among the salient facial features; moreover, features of the upper part are more useful than those of the lower part. As a consequence, in the case of eyes occlusion, larger weights should be assigned to the first and third modules, and the weight of the first module should be slightly larger than that of the third module, that is α_1 and α_3 should be larger than α_2; furthermore, α_1 should be slightly larger than α_3. Meanwhile, Table 1 shows that the best combination of weight coefficients is $\alpha_1 = 0.46$, $\alpha_2 = 0.13$, and $\alpha_3 = 0.41$, which is completely in accordance with the analysis. Therefore, the optimization result is reasonable. The above experimental results demonstrate that the proposed method can automatically assign smaller weights to occlusion regions and greater weights to occlusion-free regions, which reduces the effects of occlusion region on image classification, and consequently, the method has strong robustness to occlusion. Likewise, the optimization analyses with mouth occlusion and random occlusion are able to be analyzed. In addition, from Tables 1, 2, 3, and 4, it can be seen that the global classification results of the proposed NEDMF_SR method in six databases all have different degrees of improvement compared with the best local ones, which validates that the proposed global classifier makes full use of the discriminative features of each module,

Table 3 Results of weight coefficient learning with NEDMF_SR on CMU PIE and COIL20 databases (%)

The k-th module, $k = 1, 2, 3$	CMU PIE database			COIL20 database		
	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy
1	58.45	0.49	62.68	80.38	0.48	82.96
2	57.47	0.48		78.18	0.46	
3	2.57	0.03		9.09	0.06	

Table 4 Results of weight coefficient learning with NEDMF_SR on MNIST database (%)

The k-th module, $k = 1, 2, 3$	MNIST database		
	Local recognition accuracy	Weight coefficient $\alpha_k \in [0, 1]$	Global recognition accuracy
1	54.7	0.43	66.5
2	61.0	0.49	
3	10.0	0.08	

assigns corresponding weight according to its contribution, and consequently, becomes an enhanced discriminant classifier.

4.6 Experiments on occluded face recognition

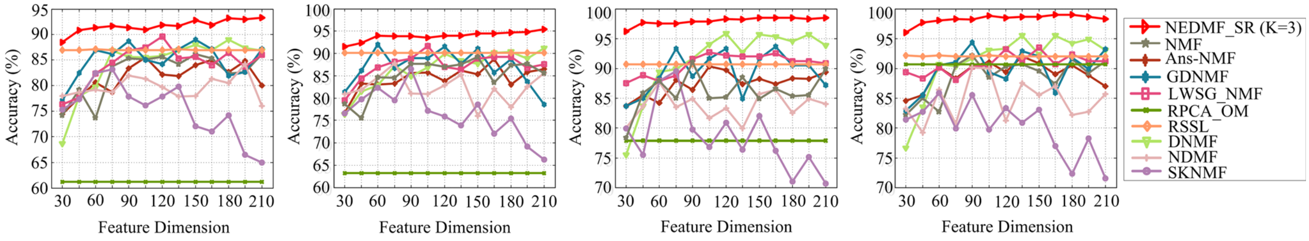
In many actual face recognition circumstances, the testing image may suffer from partial corruption or occlusions. Thus, the robustness of the proposed NEDMF_SR method to different kinds of occlusions, such as real disguise and varying degrees of random block occlusion, is tested in this section.

4.6.1 Real-world malicious occlusion

The face recognition experiments under real-world malicious occlusion are conducted using the proposed NEDMF_SR method in this part. A subset chosen from the AR database is utilized to conduct this experiment. The subset includes 2600 images from 100 individuals (26 samples per class), namely 50 men and 50 women. Experiments are carried out in two cases: image with scarf or sunglasses occlusions.

1. Sunglasses occlusion

In this experiment, for each subject, 4, 5, 6, and 7 non-occluded frontal view facial expression images are randomly selected as the training set, and 600 images (6 samples per person) with sunglasses occlusion are used as the testing set. One of the modular schemes, that is, each facial image is equally divided into three non-overlapping modules from top to bottom, is taken as an example. Here, RPCA_OM is firstly utilized to reduce the dimension by keeping 95% data energy, and the dimensionality of RSSL [40] is $f = 5 \times \lfloor (\upsilon - 1)/5 \rfloor$, where f is the number of concepts. Each experiment is, respectively, repeated 10 times, and the average accuracy is reported. Figure 11



(a) N=4 **(b)** N=5 **(c)** N=6 **(d)** N=7

Fig. 11 Recognition rates under sunglasses occlusion versus dimensions on AR database with N images per individual randomly selected for training

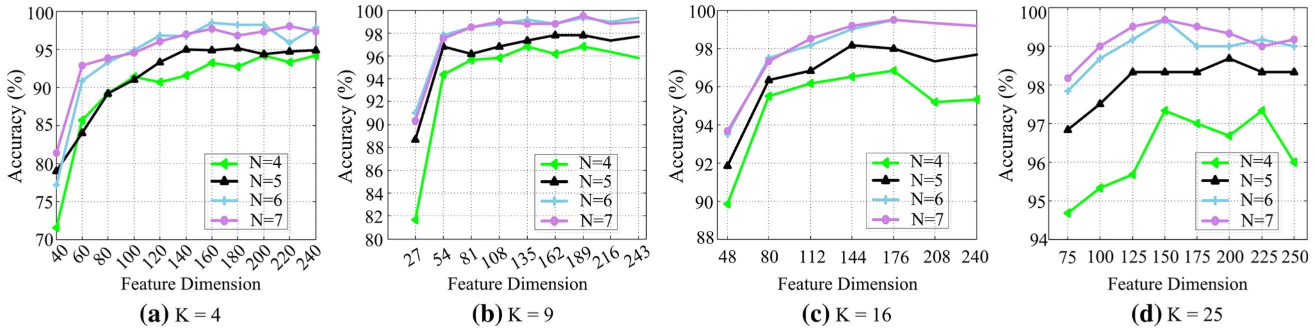


Fig. 12 Recognition accuracies under sunglasses occlusion versus dimensions on AR database with K modules

shows the variations of average accuracies with the reduction of subspace dimensions.

It can be observed that the face recognition performance of the proposed NEDMF_SR method outperforms all the contrastive methods, which indicates that NEDMF_SR is able to learn a more effective data representation, and therefore shows stronger robustness to occlusion. Furthermore, it can also be seen that the recognition accuracy of the proposed NEDMF_SR method steadily improves as the proportion of the labeled data increases.

In order to further analyze the influence of different modular schemes on the recognition accuracy of the proposed NEDMF_SR method, the corresponding experimental results are given below, respectively. For different values of K, the maximum of feature dimensions is kept at about 240. Figure 12 shows the average recognition accuracies versus the feature dimensions with respect to different values of K.

It can be seen that: (1) The image recognition accuracy of the proposed NEDMF_SR method improves stably with increasing number of modules. Through analysis, the reasons are that the more the number of modules is, the finer the block division is, and the more discriminative local feature the proposed method can capture. Meanwhile, the increase of module number is beneficial to the separation of occlusion regions, and the occlusion regions are assigned to the smaller weights, which alleviates the passive impacts of occlusion on face recognition and improves the robustness to occlusion. In addition, the information complementarity among different modules is utilized. All of these improve the recognition accuracy. (2) When $N = 7$, the recognition accuracy reaches the highest value with the proposed method, that is, the recognition accuracy steadily improves as the number of training samples increases, which is consistent with the conclusion drawn from Fig. 12.

2. Scarf occlusion

The experiments of this part aim to conduct the face recognition experiments with more complex and larger area of occlusion. For each individual, 4, 5, 6, and 7 non-occluded frontal view facial expression images are randomly chosen as the training set, and 600 images (6 samples per subject) under scarf occlusion (more complex and larger area of occlusion) are used for testing. The same trisection partition scheme is taken as an example. Figure 13 shows the variations of average accuracies with the feature dimensions.

It can be seen that the proposed NEDMF_SR method achieves the best performance, and the average recognition accuracy reaches up to 87.98% and the highest up to 93.83% when $N = 7$, which further demonstrates the robustness of our method to large area of occlusion. In addition, the performance of the proposed method steadily improves as the number of training samples increases, which is entirely consistent with the experiment conclusion in Sect. 4.6.1.

To further analyze the impact of different modular schemes on the recognition accuracy of the proposed

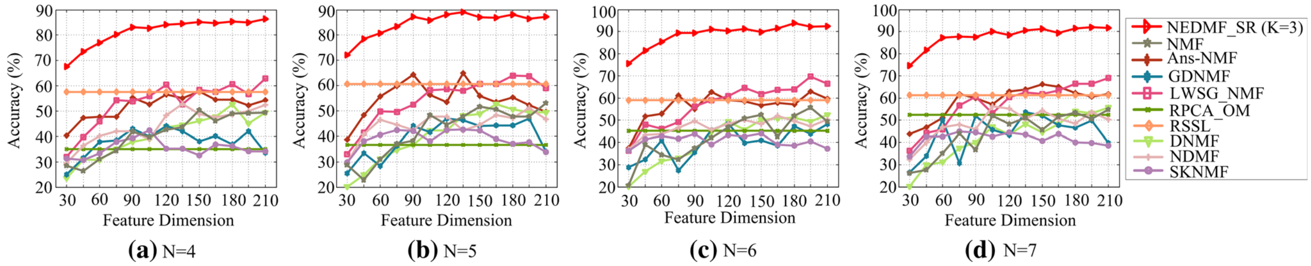
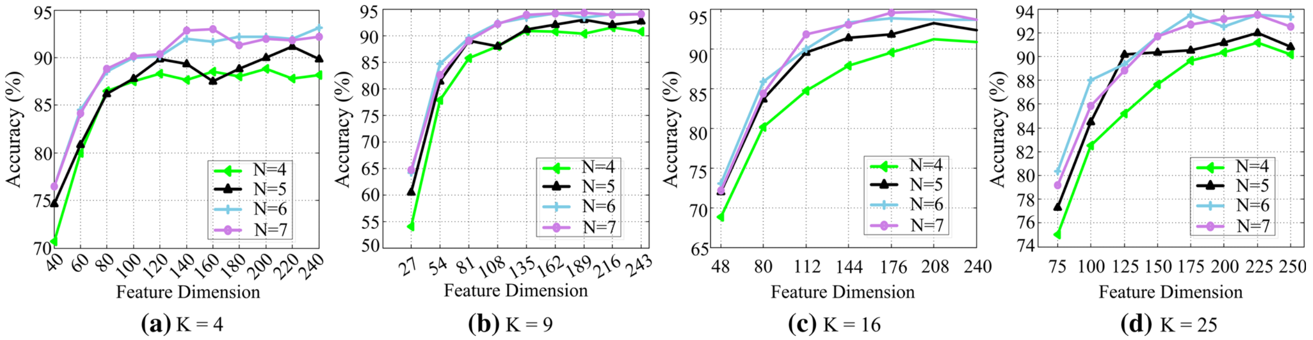


Fig. 13 Recognition accuracies under scarf occlusion versus dimensions on AR database with N images per individual randomly selected for training



(a) K = 4 **(b)** K = 9 **(c)** K = 16 **(d)** K = 25

Fig. 14 Recognition accuracies under scarf occlusion versus dimensions on AR database with K modules

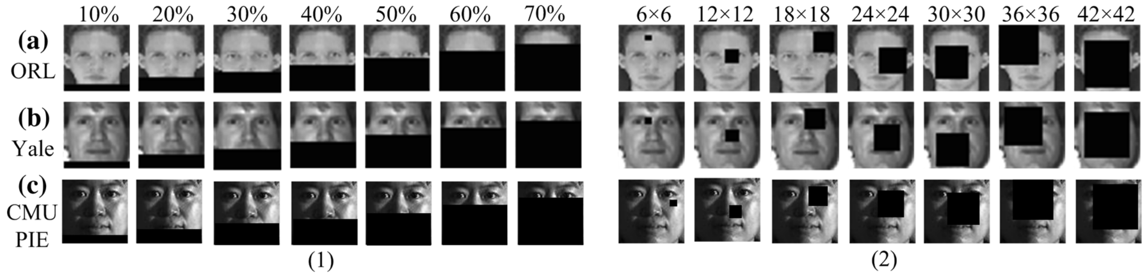
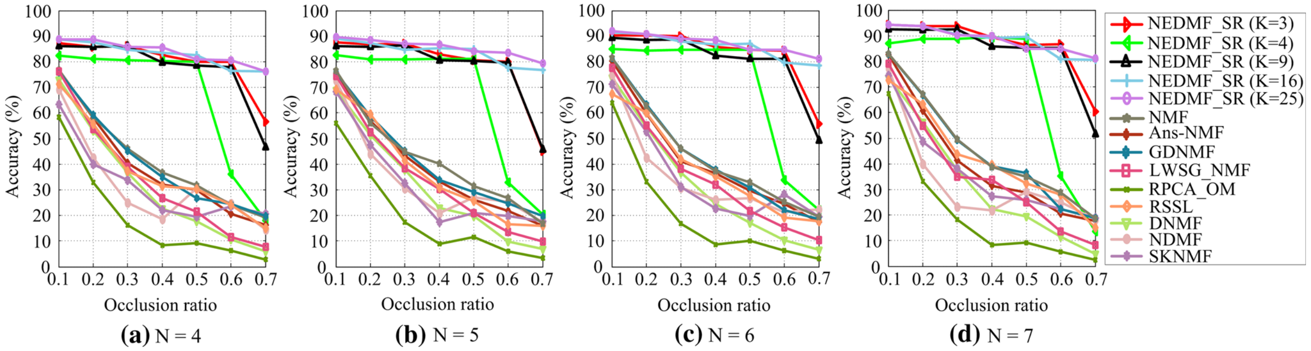


Fig. 15 Examples of partial occlusion. (**1**) Facial images with 10–70% partial occlusion areas. (**2**) Facial images with 6 × 6 to 42 × 42 pixels of random block occlusion areas

NEDMF_SR method, similar to the experiment under sunglasses occlusion, the average recognition accuracies versus the feature dimensions are given, respectively, for different K, as shown in Fig. 14. It can be seen that the recognition accuracy of the proposed NEDMF_SR method presents a growth tendency with the increase of K, which is the same as the conclusion drawn from the experiment under sunglasses occlusion.

4.6.2 Contiguous occlusions of random block

This experiment aims to evaluate the robustness of the NEDMF_SR method to random occlusion. Experimental data are, respectively, chosen from ORL, Yale, and CMU PIE databases. In the first setting, 10–70% areas of face images are partially occluded, as illustrated in Fig. 15(1). In the second setting, the face images are randomly occluded by a pixel block of size 6 × 6 to 42 × 42, at intervals of 6 × 6 pixels, as illustrated in Fig. 15(2). The database partitions in relevant experiments: (1) ORL database: 4, 5, 6, and 7 non-occluded images are randomly chosen as the training set for each individual, and the rest with artificially added occlusions are used for testing; (2) Yale database: 5, 6, 7, and 8 non-occluded images are taken as the training set, and the rest are for testing; (3) CMU PIE database: 12, 16, 20, and 24 non-occluded images are used as the training set for each individual, and the rest are for testing.



(a) N = 4 **(b)** N = 5 **(c)** N = 6 **(d)** N = 7

Fig. 16 Recognition comparison with different occlusion degree and N on ORL database

1. Faces with 10–70% partial occlusion areas

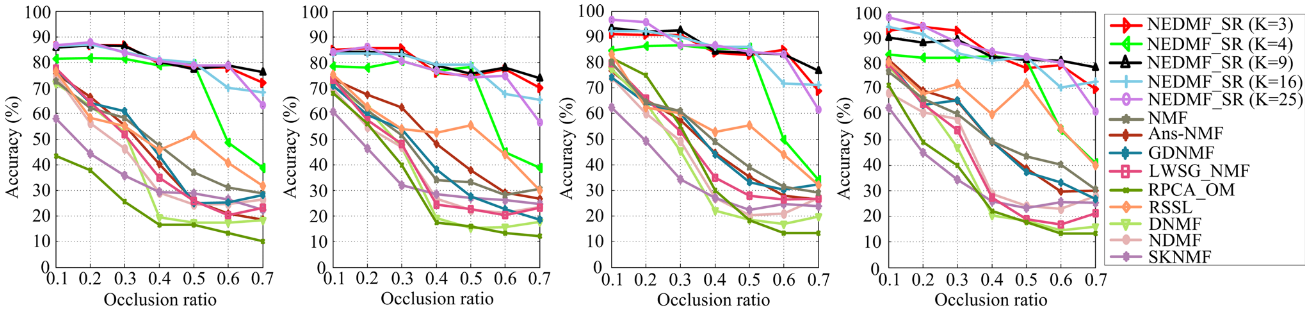
Figure 16 compares the average recognition accuracies of different methods with varying degrees of partial occlusion on ORL database. Meanwhile, the face recognition results of NEDMF_SR with five different modular schemes ($K = 3, 4, 9, 16, 25$) are given simultaneously. Specifically, the variation range of feature dimension is set as [8:2:32] when $K = 3$; subsequently, [3:2:25] when $K = 4$; [2:2:11] when $K = 9$; [2:1:6] when $K = 16$; and [2:1:4] when $K = 25$. In order to make fair comparison, the variation ranges of feature dimension for other methods are all set as [24:6:96]. Besides the feature dimension of RPCA_OM is set to 50.

It can be observed from Fig. 16 that: (1) The performance decreases slowly in a large range of 10–60% occlusion using the proposed NEDMF_SR method while drops dramatically in a slight range of 10–20% occlusion using other NMF methods. (2) To our relief, when 50% of faces are maliciously occluded, the face recognition accuracies of NEDMF_SR are always maintained at 80% or more under different modular schemes. (3) When K is 16 and 25, respectively, even with more than 50–70% occlusion, the recognition accuracies still keeps the accuracies over 80% with the NEDMF_SR, while basically in the range of 5–20% with other contrastive methods. In summary, the above experimental results show that the recognition accuracy of this paper is superior to those of all the contrastive methods under large area of occlusion, that is, the proposed method has strong robustness to the large area of continuous occlusion. Figures 17 and 18 compare the average recognition accuracies of different methods with varying degrees of partial occlusion on Yale and CMU PIE databases, respectively, and the same conclusions can be drawn.

2. Faces with random block occlusion

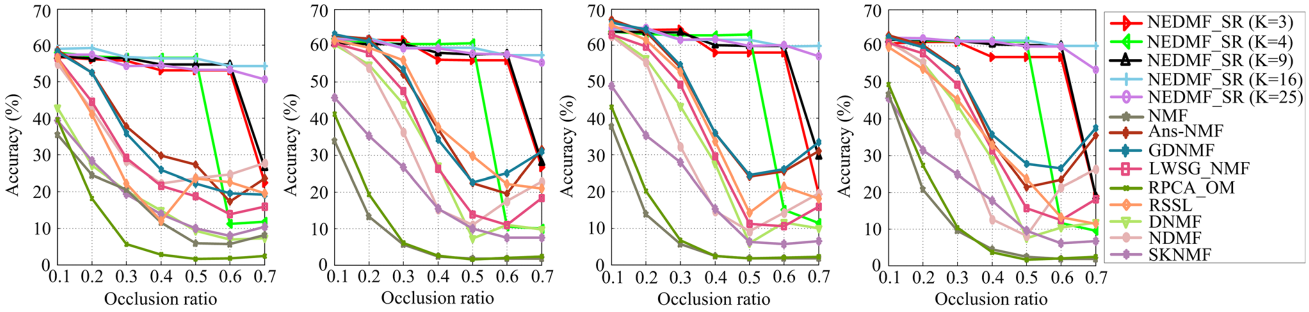
Under random block occlusion, the face recognition accuracies of the NEDMF_SR method and other contrastive methods are also, respectively, calculated to demonstrate the robustness of NEDMF_SR, where the block size is $n \times n$, and $n = 6$ to 42. The setup is challenging as the locations of occluded regions from two faces remained to be recognized have some certain differences. Thus, images of one individual would seem to be quite different in terms of pixel distances.

Figure 19 shows the average recognition accuracies of different methods with varying degrees of partial occlusion on ORL database. The experiment setting is the same as that of the first kind of occlusion. It can be observed that the proposed method has stronger robustness to random occlusion than other methods. Furthermore, when K is 16



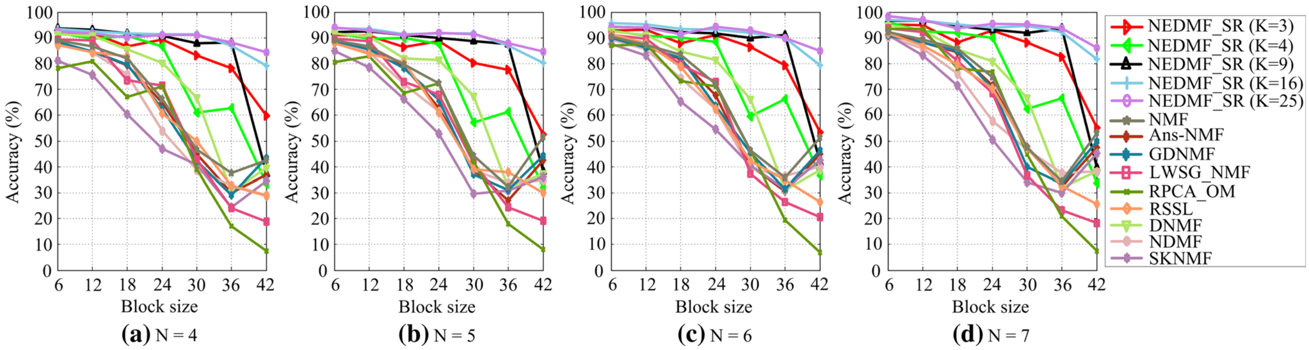
(a) N = 5 **(b)** N = 6 **(c)** N = 7 **(d)** N = 8

Fig. 17 Recognition comparison with different occlusion degree and N on Yale database



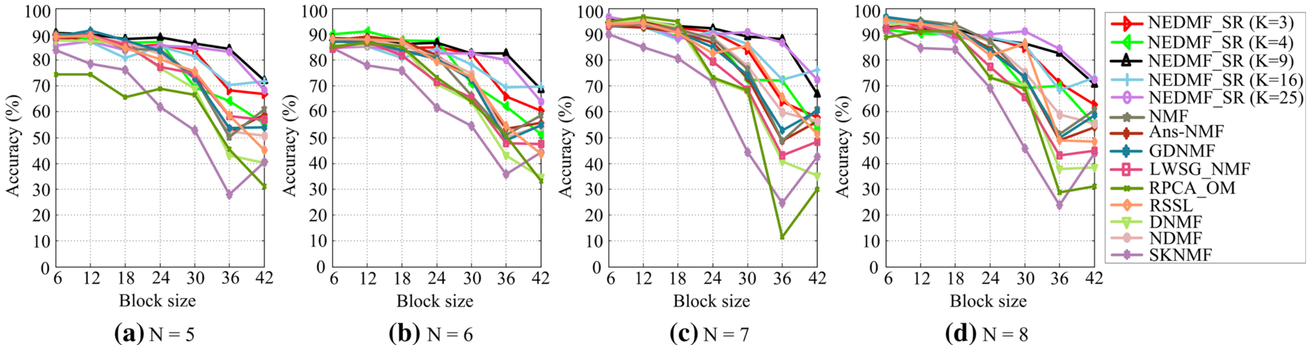
(a) N = 12 **(b)** N = 16 **(c)** N = 20 **(d)** N = 24

Fig. 18 Recognition comparison with different occlusion degree and N on CMU PIE database



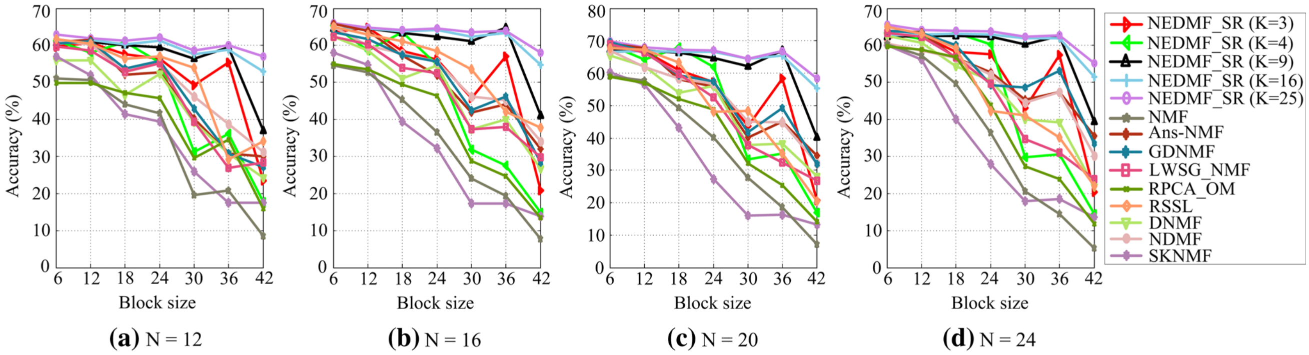
(a) N = 4 **(b)** N = 5 **(c)** N = 6 **(d)** N = 7

Fig. 19 Recognition comparison with different random occlusion degree and *N* on ORL database



(a) N = 5 **(b)** N = 6 **(c)** N = 7 **(d)** N = 8

Fig. 20 Recognition comparison with different random occlusion degree and *N* on Yale database



(a) N = 12 **(b)** N = 16 **(c)** N = 20 **(d)** N = 24

Fig. 21 Recognition comparison with different random occlusion degree and *N* on CMU PIE database

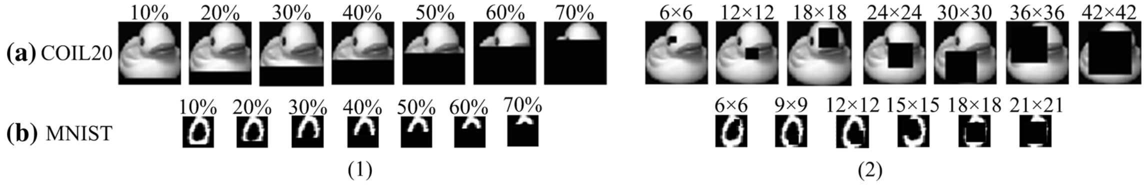


Fig. 22 Example images with occlusions. (**1**) Images with 10–70% partial occlusion areas; (**2**) Images with 6 × 6 to 42 × 42 pixels of random block occlusion areas in COIL20 database and 6 × 6 to 42 × 42 pixels in MNIST database

and 25, respectively, even with size of 30 × 30 block occlusion, the NEDMF_SR still keeps an accuracy over 90%. However, when the block size is greater than 24 × 24 and *K* is equals to 4, similar to other contrastive methods, the performance of the proposed method begins to decrease quickly. Obviously, this modular scheme is not suitable for processing the face recognition under occlusion and is not advised to be adopted in practice. Figures 20 and 21 show the curves of average recognition accuracies with the proposed method and other contrastive methods versus varying degrees of partial occlusion on Yale and CMU PIE databases, respectively, and the similar conclusions can be drawn.

Based on the results presented in the previous series of experiments, the following conclusions can be obtained: (1) Compared with other methods, the proposed method is more robust to the real-world occlusion and contiguous occlusions of random block; (2) As the number of labeled samples increases, the recognition accuracy of the proposed method and other contrastive methods all have a certain degree of improvement, but our method significantly outperforms other methods, which indicates that the NEDMF_SR method can learn a more effective data representation; (3) Compared with other modular schemes, the recognition accuracy of NEDMF_SR decreases rapidly when K is 4 and the occlusion ratio is larger than 40%, which shows that this modular scheme is not suitable for processing the face recognition with large area of occlusion; (4) The recognition accuracies of other methods begin to decrease rapidly when the occlusion ratio is larger than 20%.

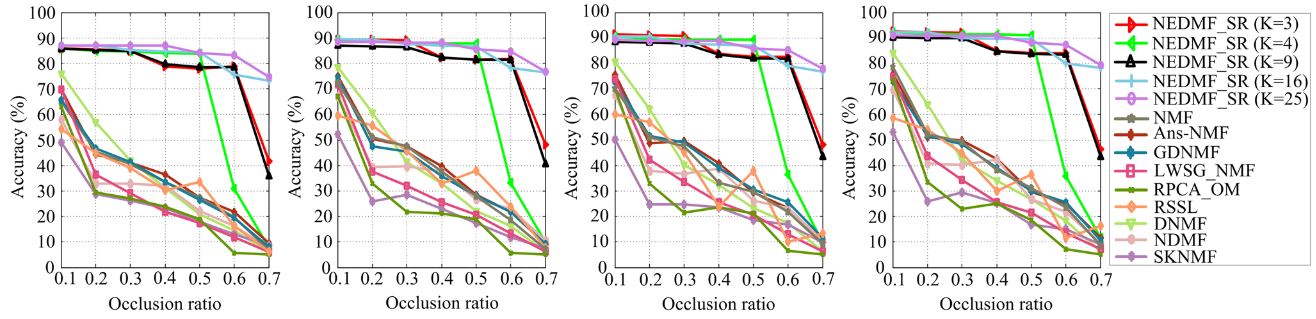
4.7 Experiments on occluded non-face recognition

In order to demonstrate the effectiveness of this paper on the non-face databases, the non-face image classification experiments under occlusions are conducted on COIL20 and MNIST databases. The occlusion schemes are also divided into two types. Specifically, one setting is that images are partially occluded by 10–70% areas, as illustrated in Fig. 22(1); and the other is that images in COIL20 database are occluded by random blocks of 6 × 6 to 42 × 42 pixels in size, at intervals of 6 × 6, while images in MNIST database are occluded by random blocks of 6 × 6 to 21 × 21 pixels in size, at intervals of 3 × 3, as illustrated in Fig. 22(2).

The partitions of training set and testing set in relevant experiments: (1) COIL20 database: 4, 5, 6, and 7 non-occluded images per subject are randomly chosen as the training set, and the rest with artificially added occlusions are used for testing; (2) MNIST database: 20, 30, 40, and 50 non-occluded images per subject are randomly chosen as the training set, and the original 2000 test images with artificially added occlusions are used for testing.

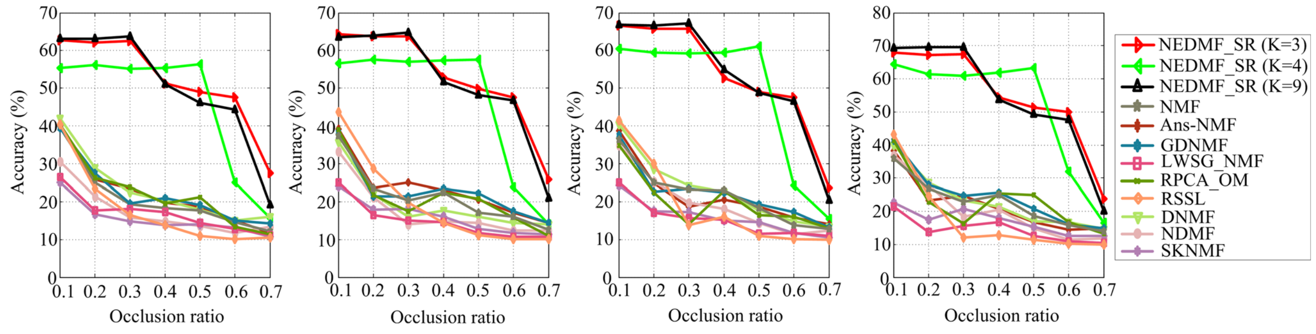
1. Non-faces with 10–70% partial occlusion areas

Figures 23 and 24 show the curves of average recognition accuracies with the proposed method and other contrastive



(a) N = 4 **(b)** N = 5 **(c)** N = 6 **(d)** N = 7

Fig. 23 Recognition comparison with different occlusion degree and N on COIL20 database



(a) N = 20 **(b)** N = 30 **(c)** N = 40 **(d)** N = 50

Fig. 24 Recognition comparison with different occlusion degree and N on MNIST database

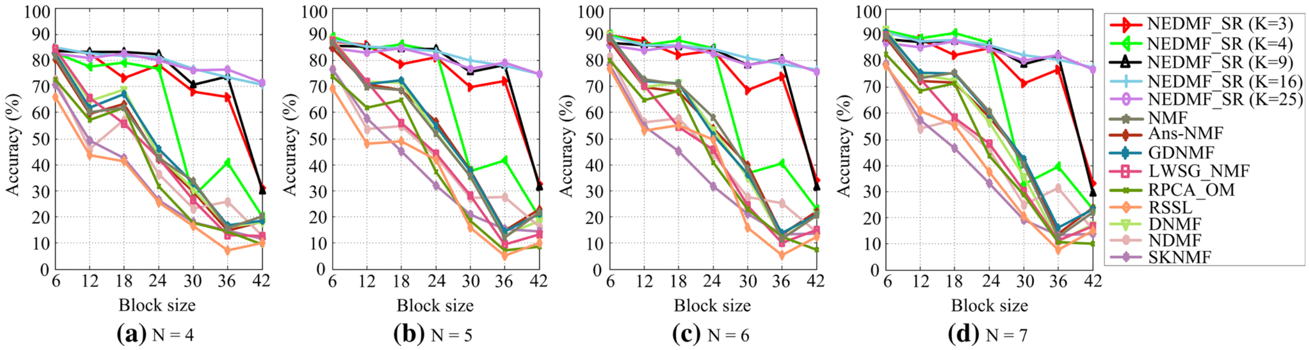


Fig. 25 Recognition comparison with different random occlusion degree and *N* on COIL20 database

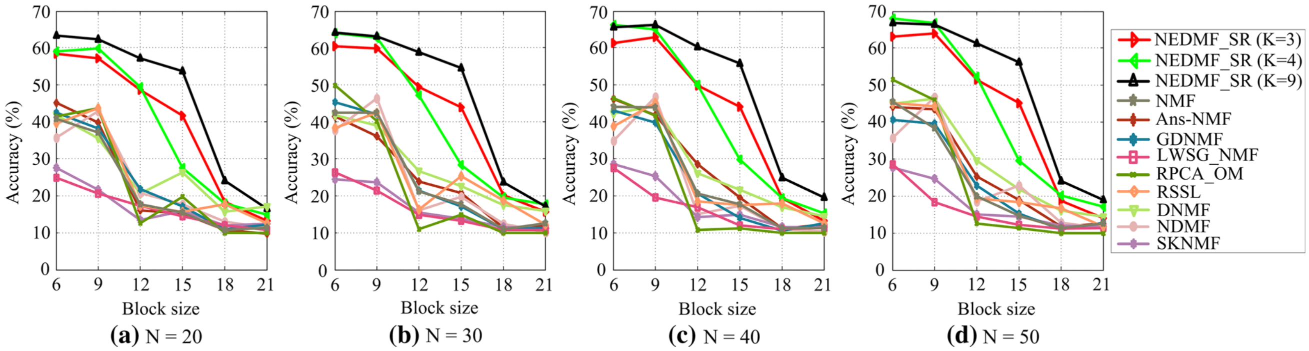


Fig. 26 Recognition comparison with different random occlusion degree and *N* on MNIST database

methods versus varying degrees of partial occlusion on COIL20 and MNIST databases, respectively. It can be seen that the proposed method has better classification performance than others, but when K is 4 and the occlusion ratio is larger than 50%, the recognition accuracy of it decreases rapidly.

2. Non-faces with random block occlusion

The non-face image recognition accuracies of different methods with random block occlusion are also evaluated to verify the robustness of the NEDMF_SR method, and the block size is $n \times n$, where $n = 6$ to 42 for COIL20 database and $n = 6$ to 21 for MNIST database, respectively. Figures 25 and 26 show the curves of average recognition accuracies with the proposed method and other contrastive methods versus varying random block occlusion on COIL20 and MNIST databases, respectively. From the results, it can be clearly observed that the proposed method shows the optimal performance and is more robust to random occlusion.

From Sects. 4.6 and 4.7, it can be seen that the proposed NEDMF_SR method has significantly better performance than other methods and is the only one that achieves uniformly good performance on all of the six databases, which demonstrates that ours is able to learn a more effective information representation. Intuitively, the experiments above show that it is indispensable and very useful to learn the local invariance, discriminativeness, and sparse representation for mining the discriminative and compact representation of data, which is beneficial to image recognition.

5 Conclusion

In this paper, a novel NEDMF_SR method called non-negative enhanced discriminant matrix factorization with sparsity regularization is proposed. Different from other discriminant analysis methods based on NMF, not only does the proposed method introduce the sparsity of coefficient matrix into NMF, but it also combines the within-class and between-class discriminant information of coefficient matrix as the regularized term to enhance the discriminant ability of low-dimensional representation. This method maximizes the between-class discrete penalty term, and at the same time, minimizes the within-class compact encouragement term and the sparse constraint term of low-dimensional representation. Moreover, the update rules and convergence proof of NEDMF_SR are also presented.

When the proposed method is applied to the image recognition under occlusions, by means of modular scheme, the local classifier based on NEDMF_SR is constructed to identify the contribution degree of different modules to image recognition. And on this basis, the global classifier is further constructed to fuse the discriminant information from each module in a weighted way, which better overcomes the performance degradation caused by serious occlusions.

The extensive experiments demonstrate that the proposed NEDMF_SR method has strong robustness to various image damage: real disguise and random block occlusion, and possesses better performances than contrastive methods on six standard image databases.

Since the deep learning for data representation is getting popularity, it has many successful applications. In the future, we will dedicate ourselves to studying the NMF method with new characteristics. On this basis, we will further focus on the combination of deep learning and NMF method and import the pre-trained deep learning representations into an additional NMF layer to construct a deep network architecture with local invariance, discriminativeness, and so on, which will improve the performance of face recognition under large area of occlusion.

Acknowledgements This work was supported partially by National Natural Science Foundation of China (Grant No. 61072110), Science and Technology Overall Innovation Project of Shaanxi Province (Grant 2013KTZB03-03-03), Industrial Research Project of Shaanxi Province (Grant BD12015020001), and International Cooperation Project of Shaanxi Province (Grant BD18015050001).

Compliance with ethical standards

Conflict of interest All the authors of the manuscript declared that there are no potential conflicts of interest.

Human and animal rights All the authors of the manuscript declared that there is no research involving human participants and/or animal.

Informed consent All the authors of the manuscript declared that there is no material that required informed consent.

Appendix: Proof of Theorem 1

To prove Theorem 1, it is necessary to show the nonincreasing property of objective function in Eq. (14) under the iteration rules in Eqs. (20) and (21). As the iteration rule of Eq. (20) is identical to original NMF, the convergence proof of NMF can be adopted to manifest that objective function is nonincreasing under the iteration rule in Eq. (20), and it is only needed to prove that the objective function is nonincreasing under Eq. (21). An auxiliary function approximate to that utilized in the Expectation Maximization (EM) algorithm is adopted in the proof process.

Definition 1 If the conditions $G(M, M^{(t)}) \geq F(M)$, $G(M, M) = F(M)$ hold, then $G(M, M^{(t)})$ is an auxiliary function of $F(M)$.

Lemma 1 *If $G(M, M^{(t)})$ is an auxiliary function of $F(M)$, then $F(M)$ is nonincreasing under the following update condition.*

$$M^{(t+1)} = \arg\min_{M} G(M, M^{(t)}) \tag{26}$$

Proof $F(M^{(t+1)}) \leq G(M^{(t+1)}, M^{(t)}) \leq G(M^{(t)}, M^{(t)}) = F(M^{(t)})$.

Lemma 2 *Function $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))$, namely an auxiliary function of $F(\mathbf{H}(k))$, and $F(\mathbf{H}(k))$ are given by Eqs. (27) and (28), respectively:*

$$\begin{aligned}
G(\mathbf{H}(k), \mathbf{H}^{(t)}(k)) &= \sum_{i,j} \left(B_{i,j}(k)\log B_{i,j}(k) - B_{i,j}(k)\right) \\
&- \sum_{i,j} B_{i,j}(k) \sum_{m} \left(\frac{Z_{i,m}(k)H^{(t)}_{m,j}(k)}{\sum_{m} Z_{i.m}(k)H^{(t)}_{m,j}(k)} \right. \\
&\times \left. \left(\log(Z_{i,m}(k)H_{m,j}(k)) - \log \frac{Z_{i,m}(k)H^{(t)}_{m,j}(k)}{\sum_{m} Z_{i.m}(k)H^{(t)}_{m,j}(k)} \right) \right) \\
&+ \sum_{i,j,k} Z_{i,m}(k)H_{m,j}(k) \\
&+ \gamma \sum_{i=1}^{C} \sum_{j \neq m}^{N_i} \frac{rat_i}{N_i(N_i-1)} \left\| \boldsymbol{\eta}^{(i)}_j(k) - \boldsymbol{\eta}^{(i)}_m(k) \right\|_2^2 + \lambda \sum_{i,j} H_{i,j}(k) \\
&- \frac{\delta}{C(C-1)} \sum_{i \neq j}^{C} W'_{i,j}(k) \left\| \boldsymbol{\mu}^{(i)}(k) - \boldsymbol{\mu}^{(j)}(k) \right\|_2^2
\end{aligned} \tag{27}$$

$$\begin{aligned}
F(\mathbf{H}(k)) &= D_{NEDMF\_SR} \\
&= \sum_{i,j} \left(B_{i,j}(k) \log \frac{B_{i,j}(k)}{(\mathbf{Z}(k)\mathbf{H}(k))_{i,j}} - B_{i,j}(k) + (\mathbf{Z}(k)\mathbf{H}(k))_{i,j} \right) \\
&+ \gamma \sum_{i=1}^{C} \sum_{j \neq m}^{N_i} \frac{rat_i}{N_i(N_i-1)} \left\| \boldsymbol{\eta}^{(i)}_j(k) - \boldsymbol{\eta}^{(i)}_m(k) \right\|_2^2 \\
&- \frac{\delta}{C(C-1)} \sum_{i \neq j}^{C} W'_{i,j}(k) \left\| \boldsymbol{\mu}^{(i)}(k) - \boldsymbol{\mu}^{(j)}(k) \right\|_2^2 \\
&+ \lambda \sum_{i,j} H_{i,j}(k)
\end{aligned} \tag{28}$$

Proof It is easy to find that $G(\mathbf{H}(k), \mathbf{H}(k)) = F(\mathbf{H}(k))$. According to Lemma 1, it is only need to show $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k)) \geq F(\mathbf{H}(k))$ to prove that $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))$ is an auxiliary function of $F(\mathbf{H}(k))$.

Due to the convexity of $\log\left(\sum_m Z_{i,m}(k)H_{m,j}(k)\right)$, the following inequality holds for each non-negative element a_m, all of which subject to the condition of $\sum_m a_m = 1$.

$$-\log\left(\sum_m Z_{i,m}(k)H_{m,j}(k)\right) \leq -\sum_m a_m \log\frac{Z_{i,m}(k)H_{m,j}(k)}{a_m} \tag{29}$$

Assume $a_m = Z_{i,m}(k)H^{(t)}_{m,j}(k)\Big/\sum_m Z_{i,m}(k)H^{(t)}_{m,j}(k)$, Eq. (29) can be transformed as follows:

$$-\log\left(\sum_m Z_{i,m}(k)H_{m,j}(k)\right) \leq -\sum\nolimits_k \frac{Z_{i,m}(k)H^{(t)}_{m,j}(k)}{\sum\limits_k Z_{i,m}(k)H^{(t)}_{m,j}(k)} \left(\log\left(Z_{i,m}(k)H_{m,j}(k)\right) - \log\frac{Z_{i,m}(k)H^{(t)}_{m,j}(k)}{\sum_m Z_{i,m}(k)H^{(t)}_{m,j}(k)}\right) \tag{30}$$

From Eq. (30), it is easily observed that $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k)) \geq F(\mathbf{H}(k))$. Consequently, $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))$ can be viewed as an auxiliary function of $F(\mathbf{H}(k))$.

Proof of Theorem 1 The minimum of $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))$ in regard to $H_{m,l}(k)$ is obtained by setting the gradient to zero:

$$\begin{aligned}\frac{\partial G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))}{\partial H_{m,l}(k)} &= -\sum\nolimits_i B_{i,l}(k)\frac{Z_{i,m}(k)H^{(t)}_{m,l}(k)}{\sum_n Z_{i,n}(k)H^{(t)}_{n,l}(k)}\frac{1}{H_{m,l}(k)} \\ &\quad + \sum\nolimits_i Z_{i,m}(k) + \frac{4rat_r\gamma}{N_r - 1}\left(H_{m,l}(k) - \mu^{(r)}_m(k)\right) \\ &\quad - \frac{4\delta W'_{i,r}(k)}{N_r C(C-1)}\sum_{i\neq r}^{C}\left(\mu^{(r)}_m(k) - \mu^{(i)}_m(k)\right) + \lambda \\ &= 0\end{aligned} \tag{31}$$

The above equation is a quadratic equation of $H_{m,l}(k)$, and by solving it, the iterative rule of Eq. (20) can be obtained. According to Lemma 1, now that $G(\mathbf{H}(k), \mathbf{H}^{(t)}(k))$ is an auxiliary function, and then, the function $F(\mathbf{H}(k))$, i.e., D_{NEDMF_SR}, is nonincreasing under the iterative rule in Eq. (20).

References

1. Jolliffe IT (1989) Principal component analysis. Springer, New York
2. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7(2):179–188
3. Shahid N, Kalofolias V, Bresson X et al (2015) Robust principal component analysis on graphs. In: IEEE conference on computer vision, pp 2812–2820
4. Hu Z, Pan G, Wang Y et al (2016) Sparse Principal Component Analysis via rotation and truncation. IEEE Trans Neural Netw Learn Syst 27(4):875–890
5. Lu M, Huang JZ, Qian X (2016) Sparse exponential family Principal Component Analysis. Pattern Recogn 60:681–691
6. Khalid MI, Alotaiby T, Aldosari SA et al (2016) Epileptic MEG spikes detection using common spatial patterns and linear discriminant analysis. IEEE Access 4:4629–4634
7. Ye Q, Yang J, Liu F et al (2016) L1-norm distance linear discriminant analysis based on an effective iterative algorithm. IEEE Trans Circuits Syst Video Technol 99:1–1
8. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
9. Trigeorgis G, Bousmalis K, Zafeiriou S et al (2017) A deep matrix factorization method for learning attribute representations. IEEE Trans Pattern Anal Mach Int 39(3):417–429
10. Sun F, Xu M, Hu X et al (2016) Graph regularized and sparse nonnegative matrix factorization with hard constraints for data representation. Neurocomputing 173(2):233–244
11. Zhang R, Hu Z, Pan G et al (2016) Robust discriminative non-negative matrix factorization. Neurocomputing 173(3):552–561
12. Wang D, Gao X, Wang X (2016) Semi-supervised nonnegative matrix factorization via constraint propagation. IEEE Trans Cybern 46(1):233–244
13. Li J, Bioucas-Dias JM, Plaza A et al (2016) Robust collaborative nonnegative matrix factorization for hyperspectral unmixing. IEEE Trans Geosci Remote Sens 54(10):6076–6090
14. Tepper M, Sapiro G (2016) Compressed nonnegative matrix factorization is fast and accurate. IEEE Trans Sig Process 64(9):2269–2283
15. Liu JX, Wang D, Gao YL et al (2017) Regularized non-negative matrix factorization for identifying differential genes and clustering samples: a survey. IEEE/ACM Trans Comput Biol Bioinform 99:1-1
16. Li SZ, Hou XW, Zhang HJ et al (2001) Learning spatially localized, parts-based representation. In: IEEE conference on computer vision and pattern recognition, pp 207–212
17. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 5:1457–1469
18. Yang Z, Xiang Y, Xie K et al (2016) Adaptive method for nonsmooth nonnegative matrix factorization. IEEE Trans Neural Netw Learn Syst 28(4):948–960
19. Jia YWY, Turk CHM (2004) Fisher non-negative matrix factorization for learning local features. In: Proceedings of Asian conference on computer vision, pp 27–30
20. Zafeiriou S, Tefas A, Buciu I et al (2006) Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. Neural Netw 17(3):683–695
21. Nikitidis S, Tefas A, Pitas I (2014) Projected gradients for subclass discriminant nonnegative subspace learning. IEEE Trans Cybernet 44(12):2806–2819
22. Guan N, Tao D, Luo Z et al (2011) Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. IEEE Trans Image Process 20(7):2030–2048
23. Lu Y, Lai Z, Xu Y et al (2016) Nonnegative Discriminant Matrix Factorization. IEEE Trans Circuits Syst Video Technol 99:1–1
24. Chen WS, Zhao Y, Pan B et al (2016) Supervised kernel nonnegative matrix factorization for face recognition. Neurocomputing 205:165–181
25. Cai D, He X, Han J et al (2011) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Int 33(8):1548–1560
26. Long X, Lu H, Peng Y et al (2014) Graph regularized discriminative non-negative matrix factorization for face recognition. Multimed Tools Appl 72(3):2679–2699

27. Liao Q, Zhang Q (2016) Local coordinate based graph-regularized NMF for image representation. IEEE Trans Sig Process 124:103–114
28. Li X, Cui G, Dong Y (2016) Graph regularized non-negative low-rank matrix factorization for image clustering. IEEE Trans Cybern 99:1–14
29. Zheng WS, Lai JH, Liao S et al (2012) Extracting non-negative basis images using pixel dispersion penalty. Pattern Recogn 45(8):2912–2926
30. Feng Y, Xiao J, Zhou K et al (2015) A locally weighted sparse graph regularized Non-Negative Matrix Factorization method. Neurocomputing 169:68–76
31. Cai D, Wang X, He X (2009) Probabilistic dyadic data analysis with local and global consistency. In: Proceedings of the 26th annual international conference on machine learning, pp 105–112
32. He X, Niyogi P (2003) Locality preserving projections. Adv Neural Inf Process Syst 16:153–160
33. Sprechmann P, Bronstein AM, Sapiro G (2015) Learning efficient sparse and low rank models. IEEE Trans Pattern Anal Mach Int 37(9):1821–1833
34. Naseem I, Togneri R, Bennamoun M (2010) Linear regression for face recognition. IEEE Trans Pattern Anal Mach Int 32(11):2106–2112
35. Martinez AR, Benavente R (1998) The AR face database. CVC technical report 24, Barcelona, Spain
36. Samaria FS, Harter AC (1994) Parameterisation of a stochastic model for human face identification. In: Proceedings of the second IEEE workshop on applications of computer vision, pp 138–142
37. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Int 19(7):711–720
38. Lyons MJ, Budynek J, Akamatsu S (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Int 21(12):1357–1362
39. Nie F, Yuan J, Huang H (2014) Optimal mean robust principal component analysis. In: Proceedings of the 31st international conference on machine learning, pp 1062–1070
40. Li Z, Liu J, Tang J et al (2015) Robust structured subspace learning for data representation. IEEE Trans Pattern Anal Mach Int 37(10):2085–2098