



Projective robust nonnegative factorization



Yuwu Lu^a, Zhihui Lai^b, Yong Xu^{c,*}, Jane You^d, Xuelong Li^e, Chun Yuan^a

^a Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, China

^b College of Computer Science and Software Engineering, Shenzhen University, China

^c Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, China

^d Department of Computing, Hong Kong Polytechnic University, China

^e State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:

Received 7 July 2015

Revised 1 April 2016

Accepted 4 May 2016

Available online 11 May 2016

Keywords:

Robust

Nonnegative matrix factorization

Graph regularization

Face recognition

ABSTRACT

Nonnegative matrix factorization (NMF) has been successfully used in many fields as a low-dimensional representation method. Projective nonnegative matrix factorization (PNMF) is a variant of NMF that was proposed to learn a subspace for feature extraction. However, both original NMF and PNMF are sensitive to noise and are unsuitable for feature extraction if data is grossly corrupted. In order to improve the robustness of NMF, a framework named Projective Robust Nonnegative Factorization (PRNF) is proposed in this paper for robust image feature extraction and classification. Since learned projections can weaken noise disturbances, PRNF is more suitable for classification and feature extraction. In order to preserve the geometrical structure of original data, PRNF introduces a graph regularization term which encodes geometrical structure. In the PRNF framework, three algorithms are proposed that add a sparsity constraint on the noise matrix based on $L_{1/2}$ norm, L_1 norm, and $L_{2,1}$ norm, respectively. Robustness and classification performance of the three proposed algorithms are verified with experiments on four face image databases and results are compared with state-of-the-art robust NMF-based algorithms. Experimental results demonstrate the robustness and effectiveness of the algorithms for image classification and feature extraction.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Nonnegative matrix factorization (NMF) [5] is a popular matrix factorization technique for dimensionality reduction [15,30] and feature extraction [16,27]. NMF represents original input data as the output of two low-rank nonnegative matrix factors. As a parts-based representation of original data, NMF only employs additive representation. In other words, all elements must be equal to or greater than zero. With non-negativity constraints on the two matrix factors, NMF typically yields a sparse representation of data and has been widely used in pattern recognition [28], computer vision [11,31], and image processing [4,13].

Due to non-negativity constraints, NMF favors to sparse representation, but it does not always result in parts-based representation [19]. To achieve localized NMF representation, Stan et al. [20] proposed local NMF (LNMF), which adds further constraints on the two nonnegative factors. To encode discriminant information into NMF, Wang et al. [32] proposed a novel

* Corresponding author. Tel.: +86 75526032458.

E-mail address: yongxu@ymail.com (Y. Xu).

subspace method called Fisher NMF (FNMF), which also produces additive and spatially localized basis images as LNMF. Yuan et al. proposed a novel NMF method that learns spatially localized and parts-based subspace representation of visual patterns [34]. The learned localized features are not only suitable for image compression, but also for object recognition. Yang et al. proposed a novel variant of NMF called projective NMF (PNMF) [35]. Linear and nonlinear extensions of PNMf were introduced in [35].

Other studies try to combine further constraints or discriminant information into NMF to improve performance in classification [17,21], or clustering [6,12]. In [21], the authors proposed discriminant NMF (DNMF), which minimizes within-class scatter and maximizes between-class scatter to learn the nonnegative factorization matrix. Guan et al. [17] encoded manifold regularization and margin maximization into NMF and generated a new method, called manifold regularized discriminative NMF (MD-NMF). In [6], the authors proposed graph regularized NMF (GNMF), which encoded geometrical information of the data space for clustering. Liu et al. [12] proposed a semi-supervised matrix decomposition method, called constrained NMF (CNMF), that encoded label information into NMF for clustering.

Although NMF and its related works have been successfully used in many fields, these methods are sensitive to noise. Therefore, many methods have been proposed that design more robust algorithms to deal with noise. In [7], the authors proposed a robust NMF method (RNMF-21) using the $L_{2,1}$ norm as a loss function that can handle noise and outliers. Zhang et al. [14] proposed a method named robust NMF (RNMF), which assumed that some matrix data entries could be arbitrarily corrupted. Xia et al. [36] proposed a robust kernel NMF using the $L_{2,1}$ norm as a loss function. Shen et al. [3] proposed a robust NMF algorithm (RNMF-1) that jointly approximated a clean data matrix with the product of two nonnegative matrices and estimated the positions and values of outliers or noise.

Recently, robust NMF and its extensions have been successfully applied to various tasks in image processing [18], computer vision [24], and signal processing [13]. However, these algorithms encounter the following practical limitations: 1) in many applications, data contains noise and the proposed robust NMF methods cannot effectively separate them using learning nonnegative low-dimensional representation; thus, learned projections or bases are unsuitable for feature extraction and classification; 2) since samples of data lie on a low dimensional manifold embedded in high-dimensional ambient space, it is necessary to consider the geometrical structure of data to obtain parts-based representation. Existing robust NMF methods do not take this into account. Specifically, existing robust NMF and its extensions do not consider the robustness and manifold structure simultaneously.

Inspired by recent studies in robust NMF, we propose a framework called Projective Robust Nonnegative Factorization (PRNF) to overcome the aforementioned problems. PRNF projects original data on a low-dimensional subspace for robust feature extraction and classification. In the proposed PRNF framework, noise is weakened and the learned projections are more robust to noise, thus PRNF is suitable for classification. To preserve the geometrical structure of original data, PRNF introduces graph regularization to encode the data geometrical structure in the learning steps. In the PRNF framework, we introduce three regularization terms that encode the $L_{1/2}$ norm, L_1 norm, and $L_{2,1}$ norm as a sparsity constraint of the noise matrix, respectively. Experimental results on four public face databases verify robustness and competitive performance against existing robust NMF methods.

The main contributions of this paper are as follows:

- (1) A general framework for robust NMF is proposed to conclude existing robust NMF algorithms.
- (2) A general model of the projective robust NMF algorithm is presented based on the robust NMF model. The proposed framework, referred to as Projective Robust Nonnegative Factorization (PRNF), can not only weaken the influence of noise when learning the optimal projection, but can also maintain the geometrical structure of original data for a better parts-based representation.
- (3) Three algorithms that take different norms as the sparsity constraint on noise data are proposed. We propose three algorithms based on the $L_{1/2}$, L_1 , and $L_{2,1}$ norms for the noise matrix, and analyze their updated rules and prove the algorithms' convergence.

This paper is organized as follows. Section 2 introduces the motivation for the proposed method and framework. Section 3 provides an analysis of different noise distributions. Sections 4, 5, and 6 give details on the three algorithms with the $L_{1/2}$, L_1 , and $L_{2,1}$ norms for the PRNF framework and their theoretical proof of convergence, respectively. Extensive experimental results are presented in Section 7. Section 8 concludes the paper.

2. Framework for projective robust nonnegative factorization

In this section, we introduce the proposed framework, i.e., Projective Robust Nonnegative Factorization (PRNF), for robust face image classification and feature extraction. First, we explain details of the motivation for our method and then present the unified framework with different norms.

2.1. Motivation

Standard NMF is sensitive to outliers and noise [25]. Many algorithms have been proposed to improve the robustness of NMF [3,7,14]. However, existing robust NMF methods have shortcomings as discussed in the Introduction. To the best of our knowledge, there is no algorithm that can separate clean data and noise effectively and maintain the geometrical structure

of the data after projection. In this paper, we propose a robust framework named Projective Robust Nonnegative Factorization (PRNF). PRNF not only projects original data on a low-dimensional subspace, but also encodes graph information in the decomposition. Since PRNF introduces a noise term in the optimization model, it can learn the optimal projection for robust feature extraction, which can enhance classification accuracy by decreasing noise disturbance. PRNF also maintains the geometrical structure of original data to avoid loss of structural information. In the PRNF framework, with different sparsity constraints of noise matrices, we propose three algorithms that solve the problems in existing robust NMF methods. The three optimization algorithms encode the $L_{1/2}$ norm, L_1 norm, and $L_{2,1}$ norm as constraints of the noise matrix, respectively. Our method is the first to encode geometrical information into robust NMF and can separate noise from data in learning the projections. The robust NMF and PRNF frameworks are described in the next two subsections.

2.2. Framework of robust nonnegative matrix factorization

NMF decomposes input matrix data $X \in R^{M \times N}$ into the product of two nonnegative matrix factors $U = (u_1, \dots, u_k) \in R^{M \times K}$ and $V = (v_1, \dots, v_N) \in R^{K \times N}$. However, NMF is sensitive to outliers and noise. In order to improve the robustness of NMF, a common strategy is to add a noise matrix as a nonnegative part in the decomposition. Algorithms of robust NMF can be formulated with the following optimization model

$$\min_{U \geq 0, V \geq 0, E \geq 0} \text{loss}(X, UV, E) + \lambda \Omega(E) \quad (1)$$

where λ is a tradeoff parameter and $E \in R^{M \times N}$ is the noise matrix. The first term in (1) denotes robust NMF, and the last term of (1) denotes the regularized term either on U , V , or E , or any two of these. Most existing robust NMF methods can be derived from the optimization model (1). Details of existing robust NMF methods are as follows.

RNMF-21 [7] introduced the $L_{2,1}$ norm as the loss function. The model of RNMF-21 is

$$\min_{U \geq 0, V \geq 0} \|X - UV\|_{2,1} \quad (2)$$

When $\lambda = 0$ and there is no noise matrix E , model (1) becomes model (2) by using the $L_{2,1}$ norm as the loss function. In [14], the authors proposed RNMF with the following objective function

$$\min_{U \geq 0, V \geq 0, E} \|X - UV - E\|_F^2 + \lambda \|E\|_1. \quad (3)$$

Function (3) coincides with (1).

In order to encode partial corruption as additive noise, RNMF-1 [3] was proposed to classify data with noise. The cost function of RNMF-1 is

$$\min_{U \geq 0, V \geq 0} \|X - UV - E\|_F^2 + \lambda \sum_j [\|E_{:,j}\|_1]^2. \quad (4)$$

Function (4) also coincides with (1) when there is a noise matrix with the L_1 norm as the sparse constraint.

2.3. Projective robust nonnegative factorization framework

In practical applications, data usually contains noise or outliers. In the presence of noise, previous robust NMF methods only focus on reducing the influence of outliers (i.e., RNMF-21) or noise (i.e., RNMF, RNMF-1) in factorization instead of learning the projections for feature extraction. On the other hand, existing projection-based NMF methods do not take noise or outliers into account and learned projections are affected by noise, which indicate that they cannot achieve good performance in feature extraction and classification in noisy data. In this study, we use both the geometrical structure information and robustness together for projection-based NMF.

A general framework for projective robust nonnegative factorization is

$$\min_{U \geq 0, E \geq 0} \text{loss}(X, U, E), \quad (5)$$

where U denotes the base matrix, and E denotes the noise matrix.

In order to encode other information for robust nonnegative factorization, one can add a regularization term to the above model, denoted as $\Omega(U, E)$. Then, we obtain a novel model

$$\min_{U \geq 0, E \geq 0} \text{loss}(X, U, E) + \lambda \Omega(U, E). \quad (6)$$

In practical applications, the integrated local geometrical structure on robust nonnegative factorization is usually a regularized term such as [21]

$$\text{tr}(VLV^T), \quad (7)$$

where $L = D - W$ and W is the weight matrix of the nearest neighbor graph constructed by the data point at its vertices (D is a diagonal matrix where $D_{ii} = \sum_j W_{ij}$).

From functions (7) and (5), we encode geometrical information into projective NMF. A new model for the projective robust nonnegative matrix method can be written as

$$\min_{U \geq 0, E \geq 0} \|X - UU^T X - E\|_F^2 + \lambda \text{Tr}(U^T X L X^T U). \quad (8)$$

Similar to [14], we typically assume that the noise matrix is sparse and nonnegative. Thus, a sparsity constraint is added to (8)

$$\min_{U \geq 0, E \geq 0} \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_p, \quad (9)$$

where $\|\cdot\|_p$ represents p -norm of a matrix and $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are regularization parameters.

The proposed PRNF framework (i.e., (9)) has two advantages: 1) the learned projection matrix U avoids the negative influence of noise, thus it is more suitable for robust feature extraction on noisy data; 2) with the graph encoded in the model, learned projections preserve local geometrical structure, which further improves its performance since data lie on a low-dimensional manifold.

In this study, we explore three values of p (i.e., $p = 1/2$, $p = 1$, and $p = 2$, 1) and the corresponding algorithms are named PRNF-1/2, PRNF-1, and PRNF-21. Moreover, the corresponding updating rules and convergence proofs are given in Sections 4–6. By using different norms, we hope that the PRNF algorithms are suitable for different datasets with different distributions of noise.

3. Analysis of different noises distributions

In this section, we analyze different noise distributions of the three proposed algorithms, i.e., PRNF-1/2, PRNF-1, and PRNF-21. First, we analyze PRNF-1/2 and PRNF-1 that follow a Gaussian distribution. Next, we analyze PRNF-21 that follows a Laplacian distribution with zero mean. Assume that the input data x_i is an M -dimensional column vector contaminated by additional noise e_i

$$x_i = \theta_i + e_i, \quad (10)$$

where θ_i is the unobservable true value of x_i . θ_i can be viewed as a point in a M -dimensional space such that

$$\theta_i = UU^T x_i. \quad (11)$$

We assume that the noise e_i follows a Gaussian distribution

$$p(x_i|\theta_i) \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\|x_i - \theta_i\|^2}{2\sigma^2}\right\}. \quad (12)$$

In order to maximize the data log likelihood, we obtain,

$$\max_{\theta_i} \log \prod_{i=1}^N p(x_i|\theta_i) = \max_{\theta_i} \left\{ N \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^N \frac{\|x_i - \theta_i\|^2}{2\sigma^2} \right\}. \quad (13)$$

Maximizing the data log likelihood is equivalent to minimizing the term $\sum_{i=1}^N \|x_i - \theta_i\|^2$ in (13). We have $\arg \min_{\theta_i \geq 0} \sum_{i=1}^N \|x_i - \theta_i\|^2 = \arg \min_{\theta_i \geq 0} \sum_{i=1}^N \|x_i - \theta_i\|_1 = \arg \min_{\theta_i \geq 0} \sum_{i=1}^N \|x_i - \theta_i\|_{1/2}$, thus, the minimization of $\sum_{i=1}^N \|x_i - \theta_i\|^2$ is equal to the minimization of $\sum_{i=1}^N \|x_i - \theta_i\|_1$ and $\sum_{i=1}^N \|x_i - \theta_i\|_{1/2}$. Thus, we obtain (14) and (15) by substituting θ_i with $UU^T x_i$,

$$\min_{\theta_i} \sum_{i=1}^N \|x_i - \theta_i\|_1 = \min_{U \geq 0} \sum_{i=1}^N \|x_i - UU^T x_i\|_1 = \min_{E \geq 0} \|E\|_1, \quad (14)$$

$$\min_{\theta_i} \sum_{i=1}^N \|x_i - \theta_i\|_{1/2} = \min_{U \geq 0} \sum_{i=1}^N \|x_i - UU^T x_i\|_{1/2} = \min_{E \geq 0} \|E\|_{1/2}. \quad (15)$$

This means that assumption of a Gaussian noise model transfers the maximum likelihood problem into $L_{1/2}$ and L_1 problems by imposing the constraints $U \geq 0$ and $E \geq 0$.

Assume that the noise e_i follows a Laplacian distribution with zero mean

$$p(x_i|\theta_i) \sim \exp\left\{-\frac{\|x_i - \theta_i\|}{\gamma}\right\}, \quad (16)$$

where γ is a scalar parameter.

To maximize the data log likelihood, we obtain

$$\max_{\theta_i} \log \prod_{i=1}^N p(x_i | \theta_i) = \max_{\theta_i} -\frac{1}{\mathcal{V}} \sum_{i=1}^N \|x_i - \theta_i\|. \quad (17)$$

Maximizing data log likelihood is equivalent to minimizing the term $\sum_{i=1}^N \|x_i - \theta_i\|$ in (17). Thus we obtain (18) by substituting θ_i with $UU^T x_i$,

$$\min_{\theta_i} \sum_{i=1}^N \|x_i - \theta_i\| = \min_{U \geq 0} \sum_{i=1}^N \|x_i - UU^T x_i\| = \min_{E \geq 0} \|E\|_{2,1}. \quad (18)$$

This means that the assumption of a Laplacian noise model transfers the maximum likelihood problem into an $L_{2,1}$ problem by imposing the constraints $U \geq 0$ and $E \geq 0$.

4. Projective robust nonnegative factorization via $L_{1/2}$ norm

In this section, the $L_{1/2}$ norm is used as the sparsity constraint on the noise matrix. A new optimization model and its iterative rules are proposed. We also prove the convergence of the proposed algorithm.

4.1. Objective function and updating rules

The $L_{1/2}$ regularizer is an unbiased estimator that imposes strong sparsity on the minimization problem [37]. Furthermore, the $L_{1/2}$ regularizer can not only provide sparse solutions, but also provides computation efficiency.

The objective function based on the F-norm of a matrix can be presented as

$$\min_{U \geq 0, E \geq 0} \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_{1/2}, \quad (19)$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are regularization parameters and $\|E\|_{1/2}$ is defined as

$$\|E\|_{1/2} = \sum_{i=1}^M \sum_{j=1}^N e_{ij}^{1/2}. \quad (20)$$

Let ψ_{ik} and ϕ_{ij} be the Lagrange multiplier for constraints $u_{ik} \geq 0$ and $e_{ij} \geq 0$, respectively. Define matrix $\Psi = [\psi_{ik}]$ and $\Phi = [\phi_{ij}]$. The Lagrange L can be presented as

$$\begin{aligned} L &= \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_{1/2} + \text{Tr}(\Psi U^T) + \text{Tr}(\Phi E^T) \\ &= \text{Tr}((X - E)(X - E)^T) - 2\text{Tr}(UU^T X (X - E)^T) + \text{Tr}(UU^T X X^T UU^T) + \lambda_1 \text{Tr}(U^T X L X^T U) \\ &\quad + \lambda_2 \|E\|_{1/2} + \text{Tr}(\Psi U^T) + \text{Tr}(\Phi E^T). \end{aligned} \quad (21)$$

Partial derivatives of L with respect to U and E are

$$\frac{\partial L}{\partial U} = -4(X - E)X^T U + 4UU^T X X^T U + 2\lambda_1 X L X^T U + \Psi, \quad (22)$$

$$\frac{\partial L}{\partial E} = 2(E - X) + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}} + \Phi. \quad (23)$$

$E^{-1/2}$ is given by the sum of element-wise square roots of each entry in matrix E [33].

According to the Karush–Kuhn–Tucker conditions $\psi_{ik} u_{ik} = 0$ and $\phi_{ij} e_{ij} = 0$, let $\frac{\partial L}{\partial U} = 0$ and $\frac{\partial L}{\partial E} = 0$, so that we obtain

$$(4(X - E)X^T U - 4UU^T X X^T U - 2\lambda_1 X L X^T U)_{ik} u_{ik} = 0, \quad (24)$$

$$\left(2(X - E) - 2UU^T X - \frac{\lambda_2}{2} E^{-\frac{1}{2}} \right)_{ij} e_{ij} = 0. \quad (25)$$

From (24) and (25), we obtain the following

$$u_{ik} \leftarrow u_{ik} \frac{(2X X^T U + \lambda_1 X W X^T U)_{ik}}{(2E X^T U + 2UU^T X X^T U + \lambda_1 X D X^T U)_{ik}}, \quad (26)$$

$$e_{ij} \leftarrow e_{ij} \frac{(2X)_{ij}}{(2E + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}})_{ij}}. \quad (27)$$

4.2. Convergence analysis

In this subsection, we prove the convergence of the optimization (19). Regarding iterative updating rules (26) and (27), we have the following theorem:

Theorem 1. *Objective function (19) is nonincreasing under the updating rules in (26) and (27).*

Similar to [12] and [1], we also introduce an auxiliary function to prove Theorem 1, in which the following definition and lemma are needed:

Definition 1. Function $G(u, u')$ is an auxiliary function for $F(u)$ if the conditions $G(u, u') \geq F(u)$, $G(u, u) = F(u)$ are satisfied.

Lemma 1. *If G is an auxiliary function of F , then F is non-increasing under the update*

$$u^{(t+1)} = \underset{u}{\operatorname{arg\,min}} G(u, u'). \quad (28)$$

We define an appropriate auxiliary function for objective function (19). For any element u_{ab} in U , let $F_{u_{ab}}$ denote the part of (19) relevant to u_{ab} . We prove that $F_{u_{ab}}$ is non-increasing under the updating step of (33) by defining the auxiliary function regarding u_{ab} as follows:

Lemma 2. *Function*

$$G(u, u_{ab}^t) = F_{u_{ab}}(u_{ab}^t) + F'_{u_{ab}}(u_{ab}^t)(u - u_{ab}^t) + \frac{(2EX^T U + 6UU^T XX^T U + \lambda_1 XDX^T U)_{ab}}{u_{ab}^t} (u - u_{ab}^t)^2, \quad (29)$$

is an auxiliary function for $F_{u_{ab}}$, which is the only part of (19) relevant to u_{ab} .

Proof. Since $G(u, u) = F_{u_{ab}}(u)$, we only need to prove that $G(u, u_{ab}^t) \geq F_{u_{ab}}(u)$. We compare $G(u, u_{ab}^t)$ in (36) with the Taylor series expansion of $F_{u_{ab}}(u)$

$$F_{u_{ab}}(u) = F_{u_{ab}}(u_{ab}^t) + F'_{u_{ab}}(u_{ab}^t)(u - u_{ab}^t) + \frac{1}{2} F''_{u_{ab}}(u_{ab}^t)(u - u_{ab}^t)^2, \quad (30)$$

where $F''_{u_{ab}}$ is the second order derivative with respect to U . It is straightforward to check that

$$\begin{aligned} F'_{u_{ab}} &= (-4(X - E)X^T U + 4UU^T XX^T U + 2\lambda_1 XLX^T U)_{ab}, \\ F''_{u_{ab}} &= (-4X(X - E)^T + 12XX^T UU^T + 2\lambda_1 XL^T X^T)_{aa}. \end{aligned} \quad (31)$$

Using (31) in (30) and comparing with (29), we see that to prove $G(u, u_{ab}^t) \geq F_{u_{ab}}(u)$ is equivalent to proving

$$\frac{(2EX^T U + 6UU^T XX^T U + \lambda_1 XDX^T U)_{ab}}{u_{ab}^t} \geq \frac{1}{2} F''_{u_{ab}}(u_{ab}^t). \quad (32)$$

In order to prove the above inequality, we have

$$\begin{aligned} (EX^T U)_{ab} &\geq u_{ab}^t (X(E - X)^T)_{aa}, \\ (UU^T XX^T U)_{ab} &\geq u_{ab}^t (XX^T UU^T)_{aa}, \\ (XLX^T U)_{ab} &\geq u_{ab}^t (XL^T X^T)_{aa}. \end{aligned} \quad (33)$$

Thus, (32) holds and $G(u, u_{ab}^t) \geq F_{u_{ab}}(u)$. \square

Next, we define another auxiliary function for the updating rule in (27). Let $F_{e_{ab}}$ denote the part of (19) relevant to e_{ab} . We prove that $F_{e_{ab}}$ is non-increasing under the updating step of (27) by defining the auxiliary function regarding e_{ab} as follows.

Lemma 3. *Function*

$$G(e, e_{ab}^t) = F_{e_{ab}}(e_{ab}^t) + F'_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t) + \frac{(2E + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}})_{ab}}{e_{ab}^t} (e - e_{ab}^t)^2, \quad (34)$$

is an auxiliary function for $F_{e_{ab}}$, which is the only part of O relevant to e_{ab} .

Proof. Since $G(e, e) = F_{e_{ab}}(e)$, we only need to show that $G(e, e_{ab}^t) \geq F_{e_{ab}}(e)$ by comparing $G(e, e_{ab}^t)$ in (34) with the Taylor series expansion of $F_{e_{ab}}(e)$

$$F_{e_{ab}}(e) = F_{e_{ab}}(e_{ab}^t) + F'_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t) + \frac{1}{2} F''_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t)^2, \quad (35)$$

where $F''_{e_{ab}}$ is the second order derivative with respect to E . It is straightforward to check that

$$\begin{aligned}
F'_{e_{ab}} &= \left(2(E - X) + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}} \right)_{ab}, \\
F''_{e_{ab}} &= \left(2I - \frac{\lambda_2}{4} E^{-\frac{3}{2}} \right)_{aa}.
\end{aligned} \tag{36}$$

Using (36) in (35) and comparing with (34), we see that to prove $G(e, e^t_{ab}) \geq F_{e_{ab}}(e)$ is equivalent to proving

$$\frac{(2E + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}})_{ab}}{e^t_{ab}} \geq \frac{1}{2} F''_{e_{ab}}(e^t_{ab}). \tag{37}$$

Obviously, (37) holds and $G(e, e^t_{ab}) \geq F_{e_{ab}}(e)$. \square

Proof of Theorem 1. Using $G(u, u^t_{ab})$ from (29) in (28) and using $G(e, e^t_{ab})$ from (34) in (28), we obtain

$$\begin{aligned}
u_{ab}^{(t+1)} &= \arg \min_u G(u, u_{ab}^{(t)}) = u_{ab}^t \frac{(2XX^T U + \lambda_1 X W X^T U)_{ab}}{(2EX^T U + 2UU^T X X^T U + \lambda_1 X D X^T U)_{ab}}, \\
e_{ab}^{(t+1)} &= \arg \min_v G(e, e_{ab}^{(t)}) = e_{ab}^t \frac{(2X)_{ab}}{(2E + 2UU^T X + \frac{\lambda_2}{2} E^{-\frac{1}{2}})_{ab}}.
\end{aligned}$$

Since (29) and (34) are auxiliary functions, $F_{u_{ab}}$ and $F_{e_{ab}}$ are nonincreasing under the updating rule. \square

5. Projective robust nonnegative factorization via L_1 norm

In this section, we add the L_1 norm on the noise matrix as the sparsity constraint and propose a new optimization model. Iterative rules are proven to be convergent based on the Euclidean distance as a metric.

5.1. Objective function and updating rules

Standard NMF must satisfy a Gaussian distribution [7]. However, in practice, input data always contains outliers or noise. Standard NMF is sensitive to outliers or noise. The L_0 norm is used to obtain the sparsity solution in computer vision or machine learning, but optimization of the L_0 norm is difficult. The theory of compressive sensing [8] shows that the L_1 norm can still obtain a sparse solution. In the PRNF framework, the L_1 norm is introduced as the constraint of sparsity for noise. Independent of the type of noise, the L_1 norm can effectively weaken the disturbance.

The new optimization model is given as

$$\min_{U \geq 0, E \geq 0} \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_1. \tag{38}$$

The difference between optimization models (19) and (38) is that the noise matrix in (19) is with the $L_{1/2}$ norm, but (38) is with the L_1 norm.

The iterative updating rule of (38) about U is the same as (26). The updating rule of variable U and its proof of convergence are the same as in Section 4. Due to space limitations, we only introduce the updating rule of variable E and prove its convergence. The iterative updating rule of (38) about E is

$$E \leftarrow T_{\frac{\lambda_2}{2}}(X - UU^T X), \tag{39}$$

where T denotes the soft-thresholding operator [9], which is given in the next subsection.

5.2. Convergence analysis

When E is given, the iterative updating rule for (38) is the same as (26) and the proof of convergence is also the same as in Section 4; thus, it is omitted here.

When U is given, the iterative updating rule (39) can be represented as the following L_1 minimization problem

$$\min_v \frac{1}{2} \|x - v\|_F^2 + \lambda_1 \|v\|_1. \tag{40}$$

The unique solution v^* of (40) is solved by the soft-thresholding operator [9] using the following theorem:

Theorem 2. The optimal solution of (40) can be solved by the soft-thresholding operator, which is defined as

$$T_v(x) = \begin{cases} x - v, & x > v \\ x + v, & x < -v \\ 0, & \text{otherwise} \end{cases}, \tag{41}$$

where $x \in \mathbb{R}$ and $v > 0$.

According to Theorem 2, the optimal solution of problem (38) is $T_{\frac{\lambda_2}{2}}(X - UU^T X)$. Thus, the updating rules of model (38) are (26) and (39).

6. Projective robust nonnegative factorization via $L_{2,1}$ norm

In this section, we introduce the third algorithm of PRNF (i.e., PRNF-21) with the $L_{2,1}$ norm as a constraint on the noise matrix. We describe the algorithm model and its iterative updating rules, and then prove the convergence.

6.1. Objective function and updating rules

From a probability point of view, the $L_{2,1}$ norm follows a Laplacian distribution with zero mean [7]. The $L_{2,1}$ regularization penalizes each matrix row and enhances sparsity among matrix rows, thus selecting the most prominent morphometric variables [10].

The objective function of PRNF-21 is defined as

$$\min_{U \geq 0, E \geq 0} \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_{2,1}, \quad (42)$$

where $\|E\|_{2,1} = \sum_{i=1}^N \sqrt{\sum_{j=1}^M e_{ji}^2} = \sum_{i=1}^N \|E_i\|$ and E_i is the i th column of E .

Model (42) adds the $L_{2,1}$ norm as a sparsity constraint on noise, which is different from (19) and (38). The updating rule of variable U and its proof of convergence are the same as in Section 4. Due to space limitations, we only introduce the updating rule of variable E and prove its convergence.

Let ϕ_{ij} be the Lagrange multiplier for constraints $e_{ij} \geq 0$. Define matrix $\Phi = [\phi_{ij}]$. The Lagrange L can be presented as

$$\begin{aligned} L &= \|X - UU^T X - E\|_F^2 + \lambda_1 \text{Tr}(U^T X L X^T U) + \lambda_2 \|E\|_{2,1} + \text{Tr}(\Phi E^T) \\ &= \text{Tr}((X - E)(X - E)^T) - 2\text{Tr}(UU^T X (X - E)^T) + \text{Tr}(UU^T X X^T UU^T) + \lambda_1 \text{Tr}(U^T X L X^T U) \\ &\quad + \lambda_2 \|E\|_{2,1} + \text{Tr}(\Phi E^T). \end{aligned} \quad (43)$$

The partial derivative of L with respect to E is

$$\frac{\partial L}{\partial E} = 2(E - X) + 2UU^T X + \frac{\lambda_2}{2} EG + \Phi, \quad (44)$$

where $G_{ij} = 1/\|E_i\|_2$.

According to the Karush–Kuhn–Tucker condition $\phi_{ij} e_{ij} = 0$, we obtain the following

$$e_{ij} \leftarrow e_{ij} \frac{(2X)_{ij}}{(2E + 2UU^T X + \frac{\lambda_2}{2} EG)_{ij}}. \quad (45)$$

6.2. Convergence analysis

In order to prove convergence for the updating rule of (45), we first give the following lemma. We define another auxiliary function for the updating rule in (45). Let $F_{e_{ab}}$ denote the part of (42) relevant to e_{ab} . We prove that $F_{e_{ab}}$ is non-increasing under the updating rule (45) by defining the auxiliary function regarding e_{ab} as follows:

Lemma 4. *Function*

$$G(e, e_{ab}^t) = F_{e_{ab}}(e_{ab}^t) + F'_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t) + \frac{(2E + 2UU^T X + \frac{\lambda_2}{2} EG)_{ab}}{e_{ab}^t} (e - e_{ab}^t)^2, \quad (46)$$

is an auxiliary function for $F_{e_{ab}}$, which is the only part of O relevant to e_{ab} .

Proof. Since $G(e, e) = F_{e_{ab}}(e)$, we only need to show that $G(e, e_{ab}^t) \geq F_{e_{ab}}(e)$ by comparing $G(e, e_{ab}^t)$ in (46) with the Taylor series expansion of $F_{e_{ab}}(e)$

$$F_{e_{ab}}(e) = F_{e_{ab}}(e_{ab}^t) + F'_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t) + \frac{1}{2} F''_{e_{ab}}(e_{ab}^t)(e - e_{ab}^t)^2, \quad (47)$$

where $F''_{e_{ab}}$ is the second order derivative with respect to E . It is straightforward to check that

$$\begin{aligned} F'_{e_{ab}} &= \left(2(E - X) + 2UU^T X + \frac{\lambda_2}{2} EG \right)_{ab}, \\ F''_{e_{ab}} &= \left(2I + \frac{\lambda_2}{4} G \right)_{aa}. \end{aligned} \quad (48)$$

Using (48) in (47) and comparing with (46), we see that to prove $G(e, e_{ab}^t) \geq F_{e_{ab}}(e)$ is equivalent to proving

$$\frac{(2E + 2UU^T X + \frac{\lambda_2}{2} EG)_{ab}}{e_{ab}^t} \geq \frac{1}{2} F''_{e_{ab}}(e_{ab}^t). \quad (49)$$

Obviously, (49) holds and $G(e, e^{t_{ab}}) \geq F_{e_{ab}}(e)$. \square

To prove that the objective function of (42) decreases monotonically when updating G while fixing E and U , we give the lemma as follows:

Lemma 5. Under the updating rule of (45), the following holds

$$\|E^{t+1}\|_{2,1} \leq \|E^t\|_{2,1}. \quad (50)$$

Proof. First, note that

$$\|E^{t+1}\|_{2,1} - \|E^t\|_{2,1} = \sum_i^N (\|E_i^{t+1}\| - \|E_i^t\|) = \sum_i^N \left(\|E_i^{t+1}\| - \frac{1}{G_{ii}} \right), \quad (51)$$

and

$$\frac{1}{2} [\text{Tr}E^{t+1}G(E^{t+1})^T - \text{Tr}E^tG(E^t)^T] = \frac{1}{2} \sum_i^N (\|E_i^{t+1}\|^2 G_{ii} - \|E_i^t\|^2 G_{ii}) = \frac{1}{2} \sum_i^N \left(\|E_i^{t+1}\|^2 G_{ii} - \frac{1}{G_{ii}} \right), \quad (52)$$

since

$$\begin{aligned} & \|E^{t+1}\|_{2,1} - \|E^t\|_{2,1} - \frac{1}{2} [\text{Tr}E^{t+1}G(E^{t+1})^T - \text{Tr}E^tG(E^t)^T] \\ &= \sum_i^N \left(\|E_i^{t+1}\| - \frac{1}{2} \|E_i^{t+1}\|^2 G_{ii} - \frac{1}{2G_{ii}} \right) \\ &= \sum_i^N \frac{G_{ii}}{2} \left(\frac{2\|E_i^{t+1}\|}{G_{ii}} - \|E_i^{t+1}\|^2 - \frac{1}{G_{ii}^2} \right) = \sum_i^N \frac{-G_{ii}}{2} \left(\|E_i^{t+1}\| - \frac{1}{G_{ii}} \right)^2 \leq 0. \end{aligned}$$

That is

$$\|E^{t+1}\|_{2,1} - \|E^t\|_{2,1} \leq \frac{1}{2} [\text{Tr}E^{t+1}G(E^{t+1})^T - \text{Tr}E^tG(E^t)^T] = \frac{1}{2} (\|E^{t+1}\|_{2,1} - \|E^t\|_{2,1}).$$

Since

$$\|E^{t+1}\|_{2,1} - \|E^t\|_{2,1} \leq 0.$$

Thus, Lemma 5 is proven. \square

From Lemma 5, objective function (42) decreases monotonically when updating G while fixing E and U .

7. Experimental results

In this section, we systematically evaluate the proposed Projective Robust Nonnegative Factorization (PRNF) framework for robust face recognition. We evaluate the performance of PRNF-1/2, PRNF-1, and PRNF-21 in terms of outliers, noise, random corruption, and occlusions.

7.1. Databases

Four different publicly available databases are used in our experiments, i.e., YALE [38], ORL [26], CMU PIE [23,41], and AR [2,40]. The YALE face database contains 165 images of 15 individuals (each person provided 11 different images) with various facial expressions and lighting conditions. The ORL face database contains 40 distinct subjects. All subjects are in the up-right, frontal position (with some tolerance for side movement). The CMU PIE database contains 41,368 face images collected from 68 subjects. Each subject has 13 images with different poses, 43 different illumination conditions, and 4 different expressions. In our experiment, a subset of 5 near frontal poses (C05, C07, C09, C25, and C29) and illuminations indexed as 08 and 11 were used. Therefore, each subject has ten images. The AR face database contains over 4000 color face images of 126 people (70 men and 56 women), including frontal face views with different facial expressions, lighting conditions, and occlusions. Pictures of most individuals were taken at two instances (separated by two weeks). Each instance contains 13 color images and 120 individuals (65 men and 55 women) that participated in both sessions. In the experiments, we exploit color face images of 100 subjects (50 men and 50 women) as shown in [29]. Color images were converted to gray-level images. Fig. 1 shows sample images from the above four databases. Each row shows seven images captured at different conditions for one subject. All images in our experiments were resized to 56×46 pixels.



Fig. 1. Sample images from YALE (first row), ORL (second row), PIE (third row), and AR (fourth row). Each row shows seven images captured under different situations for one subject.



Fig. 2. Images with “salt & pepper” noise for YALE, ORL, and PIE, respectively.

Table 1

The performance (recognition rate (%) and standard deviation) of testing noises with different methods on YALE, ORL, and PIE face databases.

Databases	NMF	RNMF-21	RNMF-1	RNMF	PRNF-1/2	PRNF-1	PRNF-21
YALE	86.89 ± 1.24	88.77 ± 2.19	82.16 ± 2.52	85.11 ± 3.47	92.78 ± 2.18	90.33 ± 3.72	93.86 ± 2.43
ORL	78.50 ± 1.55	77.65 ± 2.57	79.20 ± 3.13	81.65 ± 3.77	85.00 ± 3.36	84.20 ± 2.98	87.50 ± 3.51
PIE	71.67 ± 4.44	73.21 ± 3.75	75.81 ± 4.19	76.69 ± 3.26	83.12 ± 4.15	84.88 ± 5.14	85.76 ± 2.81

7.2. Baselines and settings

Our experiments are divided into four parts, which include testing of outliers, noise, random corruption, and occlusions. We use YALE, ORL, and PIE databases to test the performance of PRNF-1/2, PRNF-1, and PRNF-21 with data containing outliers, noise, and random corruption. The AR database was used for testing classification accuracy of the PRNF framework for face images with occlusions. We compared our method with the algorithms of NMF [5] and related works of robust NMF, i.e., RNMF-21 [7], RNMF [14], and RNMF-1 [3]. We use a training set to learn the basis/projection used for feature extraction and a test set to report the accuracy of face recognition. The NN classifier is used to calculate the percentage of samples in the test set that were correctly classified. We set the maximum iteration number for NMF-related methods as 500 and keep it constant in all experiments. NMF and RNMF-21 have no parameters. RNMF-1 and RNMF both have only one parameter λ . We select it via cross validation from [0.0001, 0.001, 0.01, 0.1, 1, 10]. The proposed PRNF framework has two parameters, λ_1 and λ_2 , that we select from the range [0, 1]. According to the experiments, we obtain the best classification accuracy when the parameters are chosen in the range [0.01, 0.2].

7.3. Face recognition with noise

In this subsection, we add “salt & pepper” noise to the YALE, ORL, and PIE databases to test the robustness of the PRNF framework. The density of “salt & pepper” noise was set to 0.1. Fig. 2 shows images of “salt & pepper” noise in the three databases. Fig. 3 shows experimental results for the YALE, ORL, and PIE databases. In these three databases, we randomly selected five images from each subject to construct the training set, and the remaining images made up the test set. Each experiment was conducted ten times over different feature dimensions and the average accuracy is reported.

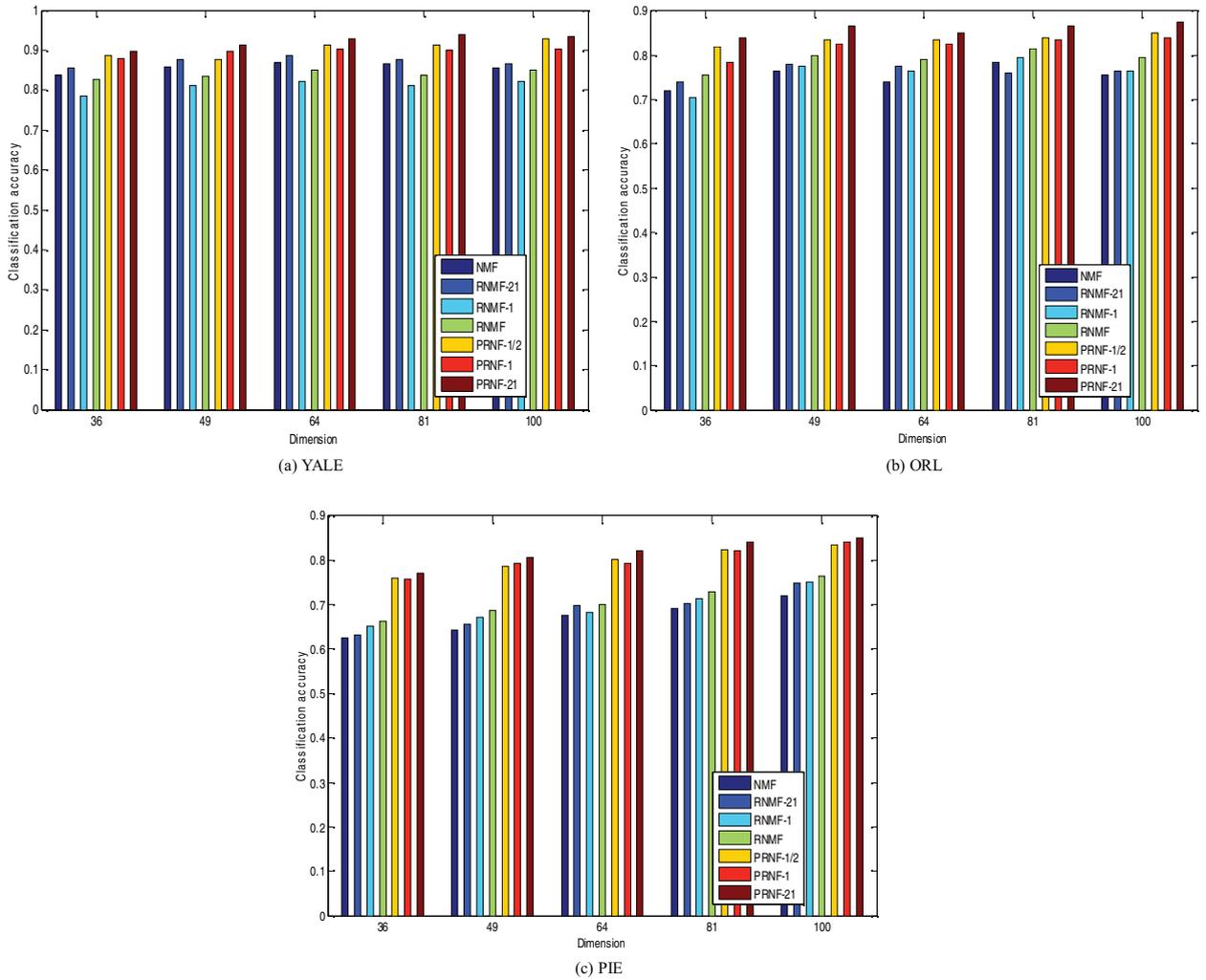


Fig. 3. Face recognition accuracy with “salt & pepper” noise over different feature dimensions for NMF, RNMF-21, RNMF-1, RNMF, PRNF-1/2, PRNF-1, and PRNF-21 on three databases: (a) YALE, (b) ORL, (c) PIE.

Table 2

The performance (recognition rate (%) and standard deviation) of testing corruption with different methods on YALE, ORL, and PIE face databases.

Databases	NMF	RNMF-21	RNMF-1	RNMF	PRNF-1/2	PRNF-1	PRNF-21
YALE	77.78 ± 2.13	78.89 ± 2.69	76.67 ± 4.12	80.30 ± 3.07	82.68 ± 2.18	81.88 ± 1.82	84.32 ± 2.08
ORL	30.00 ± 3.54	30.04 ± 1.15	31.10 ± 2.29	46.84 ± 3.56	56.80 ± 2.25	54.70 ± 4.04	57.60 ± 2.10
PIE	56.81 ± 4.05	57.80 ± 3.68	58.11 ± 1.38	59.33 ± 2.26	67.59 ± 2.67	71.10 ± 3.05	69.26 ± 2.26

From Fig. 3 and Table 1, we see that the three algorithms in the PRNF framework obtain better recognition rates than other robust NMF methods, which shows the proposed methods’ robustness to noise.

7.4. Face recognition with random pixel corruption

In this subsection, we randomly add blocks to images and keep the remaining part unchanged. The size of the block occlusion used is 14×14. Fig. 4 shows images with random corruption in our experiments for the YALE, ORL, and PIE databases. The experimental procedure is the same as in Section 7.3. Fig. 5 and Table 2 show the experimental results conducted on the YALE, ORL, and PIE databases.

Recognition rates as a function of the variation in dimensions are shown in Fig. 5. As seen from Fig. 5 and Table 2, the three algorithms of the PRNF framework obtain better recognition rates than other robust NMF methods, which shows the robustness to corruption in data.



Fig. 4. Images with random corruption: (a) YALE, (b) ORL, (c) PIE.

Table 3

The performance (recognition rate (%) and standard deviation) of testing outliers with different methods on the YALE, ORL, and PIE face databases.

Databases	NMF	RNMF-21	RNMF-1	RNMF	PRNF-1/2	PRNF-1	PRNF-21
YALE	78.78 ± 2.87	79.22 ± 4.12	80.33 ± 4.55	82.11 ± 2.97	83.44 ± 4.66	84.67 ± 3.49	84.11 ± 5.15
ORL	65.05 ± 3.61	65.75 ± 4.11	66.75 ± 6.18	68.50 ± 5.19	70.75 ± 2.84	72.65 ± 4.27	71.05 ± 3.86
PIE	68.21 ± 3.54	69.54 ± 1.36	71.22 ± 2.78	72.56 ± 3.94	79.96 ± 1.86	81.38 ± 2.66	81.11 ± 3.52

Table 4

The performance (recognition rate (%)) of testing on occlusions with different methods on the AR face database.

Occlusion	NMF	RNMF-21	RNMF-1	RNMF	PRNF-1/2	PRNF-1	PRNF-21
Sunglasses	59.25	62.25	62.50	63.75	71.50	72.50	70.00
Scarf	50.25	55.25	53.00	54.00	65.50	67.50	64.50

7.5. Face recognition with outliers

In our experiments for testing outliers, we selected three face images in the AR database as outliers. We used the selected images to substitute three images of each subject in the YALE, ORL, and PIE databases. Fig. 6 shows the images of one subject used in our experiments. We use the three outlier images and randomly selected two images of each subject to construct the training set; the remaining images made up the test set in the YALE, ORL, and PIE databases. Each experiment was conducted ten times over different feature dimensions and the average accuracy is reported. Fig. 7 and Table 3 show the experimental results for face recognition with outliers.

As seen in Fig. 7 and Table 3, the three PRNF algorithms obtain better recognition rates than other robust NMF methods, which shows the robustness of the PRNF framework to outliers.

7.6. Face recognition with occlusions

In this subsection, we test the robustness of the proposed method to occlusions in face images. Experiments were conducted on the AR database. Fig. 8 shows face images of one subject used in our experiments. Experiments were divided into two parts (denoted as Exp 1 and Exp 2). Exp 1 tests the robustness of sunglasses occlusion and Exp 2 tests the robustness of scarf occlusion. In Exp 1, Fig. 8a, b, c, and d are used for training, and Fig. 8e, f, g, and h are used for testing. In Exp 2, Fig. 8o, p, q, and r are used for training, and Fig. 8s, t, u, and v are used for testing. Fig. 9 and Table 4 illustrate classification accuracy for sunglasses occlusion and scarf occlusion, respectively.

7.7. Feature selection experiments

In this section, we conducted experiments on the YALE and ORL databases with “salt & pepper” noise to demonstrate the effectiveness of feature selection for the proposed methods. We selected two other feature selection methods for comparison. The first is the filter method and the other is the wrapper method. We computed the Fisher criterion C of each feature t in the filter method. That is $C = t^T S_{bt} / (t^T S_w t)$, where S_b and S_w are the between class and within class scatter matrices, respectively [39]. We select features corresponding to the first largest C . This method is named FISH. The wrapper method is a genetic algorithm (GA) [22]. There are two parameters in GA, i.e., the number of iterations G and the number of individuals T in each population (further details in [39]). In our experiments, we use $G = 300$ and $T = 300$. Fig. 10 shows feature

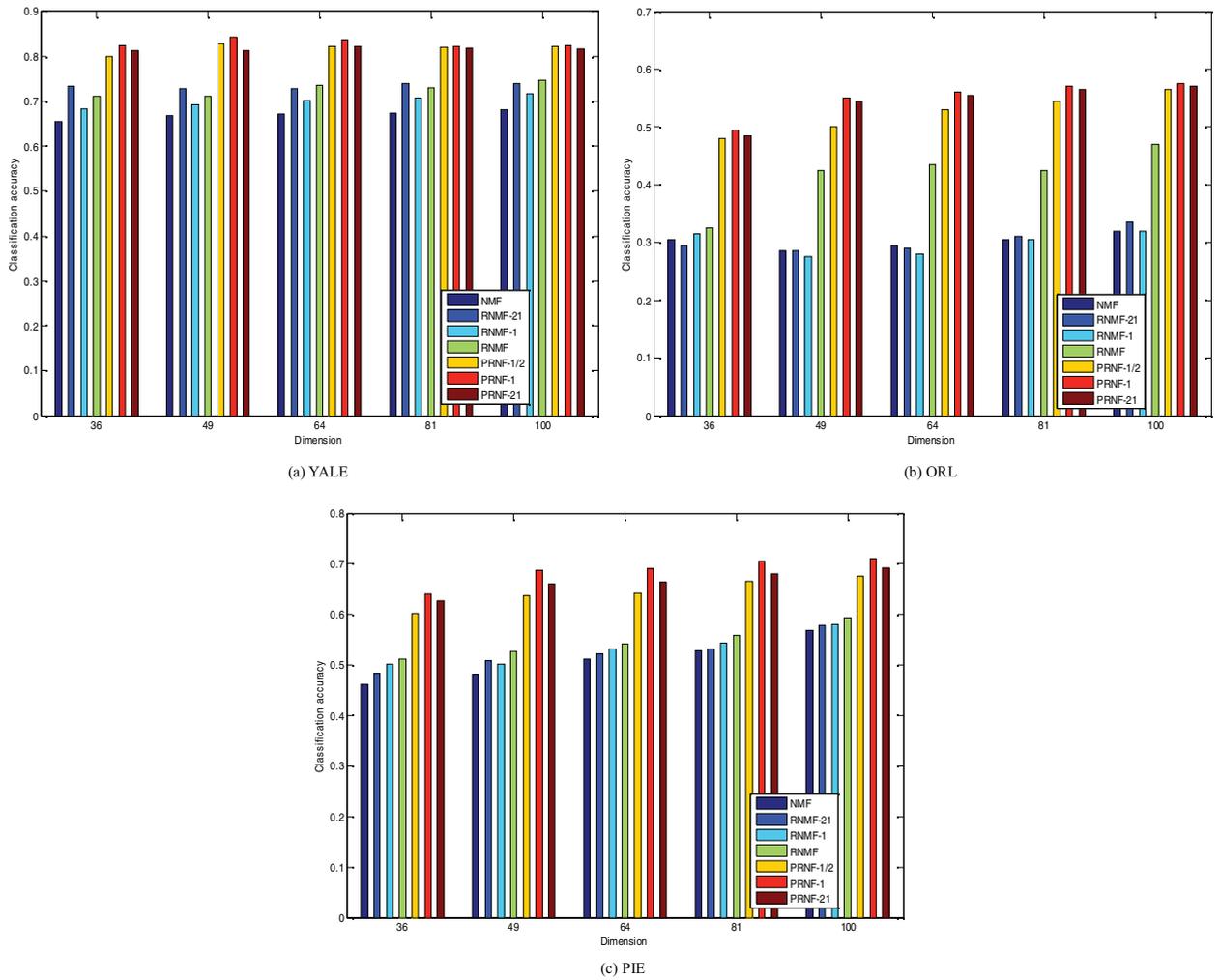


Fig. 5. Face recognition accuracy with corruption for different feature dimensions in NMF, RNMF-21, RNMF-1, RNMF, PRNF-1/2, PRNF-1, and PRNF-21 on three databases: (a) YALE, (b) ORL, (c) PIE.



Fig. 6. Single subject images with outliers. The first three images are outliers. (a) YALE, (b) ORL, (c) PIE.

selection performance of the proposed algorithms. From Fig. 10, we observe that FISH and GA are worse than PRNF-1/2, PRNF-1, and PRNF-21.

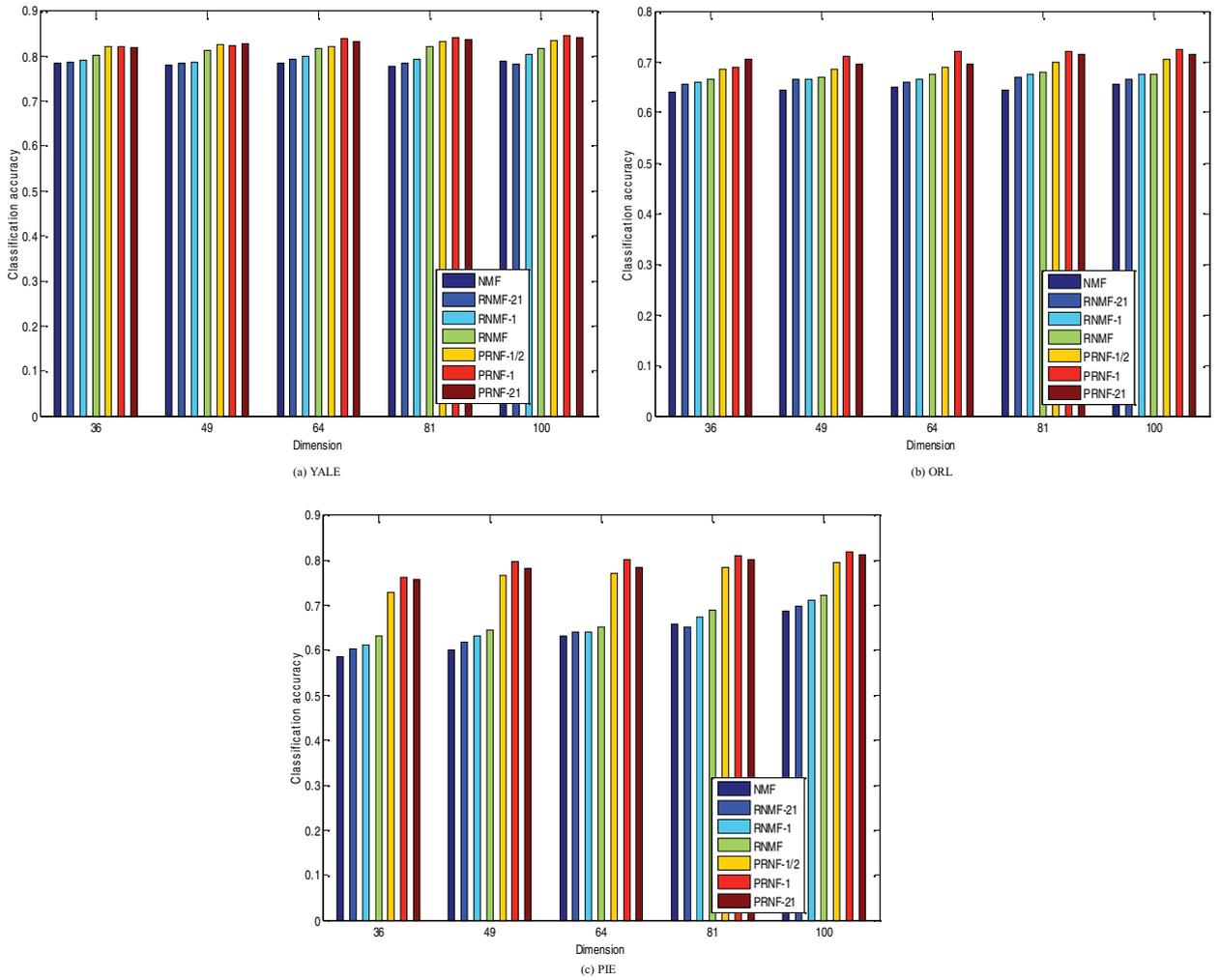


Fig. 7. Face recognition accuracy with outliers for different feature dimensions in NMF, RNMF-21, RNMF-1, RNMF, PRNF-1/2, PRNF-1, and PRNF-21 on three databases: (a) YALE, (b) ORL, (c) PIE.

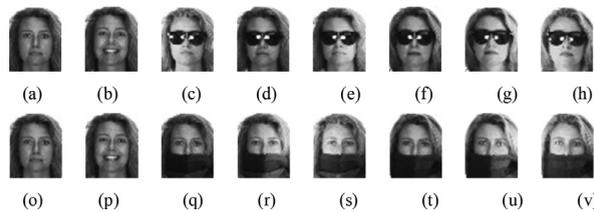


Fig. 8. Images for one subject in the AR database.

7.8. Convergence study

As proven in previous sections, we used iterative updating rules to obtain the local optimum of PRNF-1/2, PRNF-1, and PRNF-21. In this subsection, we experimentally show the convergence speed of our algorithms on the YALE database with outliers.

We compare the convergence speed of the proposed method, i.e., PRNF-1/2, PRNF-1, and PRNF-21. Fig. 11 shows the convergence rate of the three algorithms on the YALE database with outliers. In Fig. 11, the number of iterations is shown on the x-axis and the objective function value is shown on the y-axis.

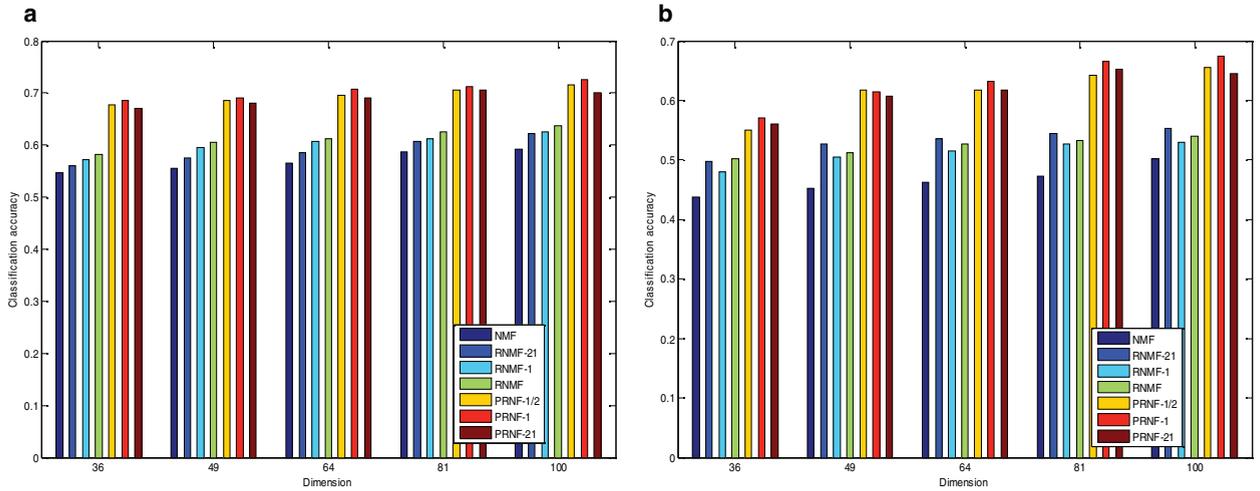


Fig. 9. Experimental results for occlusions on the AR database. (a) Sunglasses, (b) Scarf.

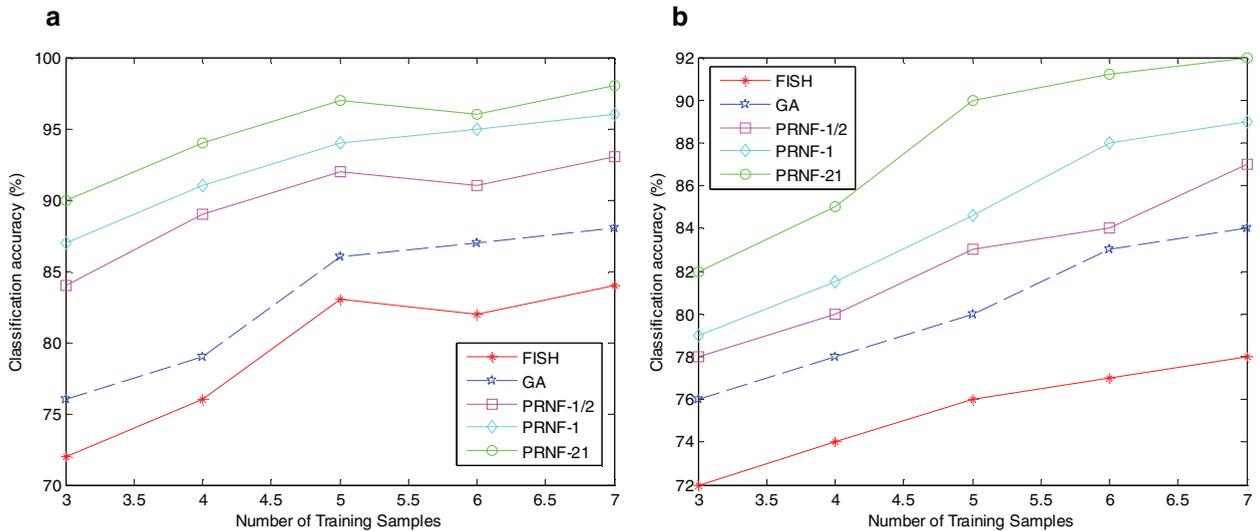


Fig. 10. Classification results of feature selection approaches. (a) Results on the YALE database. (b) Results on the ORL database.

7.9. Observations and discussions

Observations and discussions based on the experimental results are provided below:

- (1) As shown in Fig. 3 and Table 1, RNMF-21 is more robust than NMF with respect to the recognition rate of data with noise. In the PRNF framework, PRNF-21 has the highest recognition rate. This indicates that PRNF-21 can effectively weaken the disturbance of random corruption in order to preserve local geometric structure and thus improve performance.
- (2) In the experiments of random corruption, we see that RNMF obtains a higher recognition rate than RNMF-21 in most cases (Fig. 5 and Table 2). PRNF-1 also has a higher recognition rate than PRNF-21. The reason may be that the L_1 norm based method is more suitable for classifying data with random corruption than the one based on the $L_{2,1}$ norm.
- (3) As shown in Fig. 7 and Table 3, RNMF-21, RNMF-1, and RNMF perform better than NMF in the experiments containing outliers. RNMF has a higher accuracy than RNMF-21 and RNMF-1, and PRNF-1 has a higher accuracy than PRNF-21 and PRNF-1/2; these experiments show that the L_1 norm is more robust than the $L_{2,1}$ to outliers. Recognition rates of the three PRNF algorithms are higher than NMF, RNMF-21, RNMF-1, and RNMF. This indicates that introducing a sparsity constraint and local geometrical structure into the objective function simultaneously can effectively reduce the negative influence of outliers.

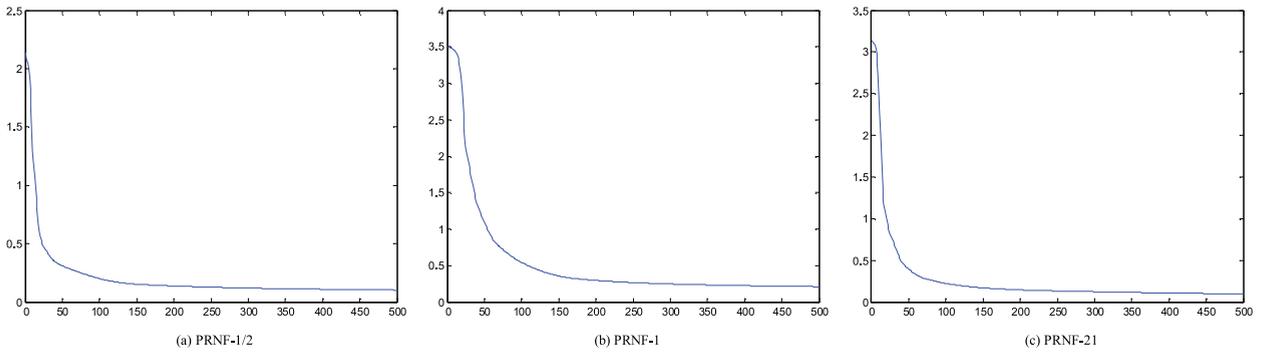


Fig. 11. Convergence rate of PRNF on the YALE image database with outliers.

- (4) In the occlusion experiments shown in Fig. 9 and Table 4, the recognition rate of RNMF-21 is still higher than NMF, RNMF-1, and RNMF, which indicates that RNMF-21 is more robust than NMF, RNMF-1, and RNMF to image occlusions. NMF is very sensitive to occlusions, thus it obtains the lowest recognition rate. In terms of recognition rate, the proposed PRNF framework is the most robust among the compared methods. Usually, PRNF-1 obtains better performance among these methods.

8. Conclusions

In this paper, a novel framework named Projective Robust Matrix Factorization (PRNF) is proposed for robust face recognition. In this framework, we derive three methods for robust classification and feature extraction. We introduced the $L_{1/2}$, L_1 , and $L_{2,1}$ norms as sparsity constraints on the noise matrix, respectively. PRNF can reduce the negative influence of noise, occlusions, and random corruptions to learn the optimal projections that preserve manifold structure. The nonnegative base matrix is more suitable for feature extraction and classification than traditional NMF and its existing robust variations. Experimental results on four public face databases verified that PRNF obtains better performance than state-of-the-art robust NMF methods when data contains noise or outliers.

Acknowledgments

This work was supported by the [Natural Science Foundation of China](#) (Grant Nos. [61203376](#), [61375012](#), [61362031](#), [61300032](#), [61170253](#), [U1433112](#)) and the National Significant Science and Technology Projects of China (No. 2013ZX01039001-002-003), the Project funded by [China Postdoctoral Science Foundation](#) (No. [2016M590100](#)), Shenzhen Municipal Science and Technology Innovation Council (Nos. [JCYJ20130329151843309](#), [JCYJ20140904154630436](#), and [JCYJ20150330155220591](#)). The authors would like to thank the associate editor and the anonymous reviewers for their constructive comments of this manuscript.

References

- [1] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, *J. R. Stat. Soc.* 39 (1977) 1–38.
- [2] A. Martinez and R. Benavente, The AR face database, Univ. Purdue, CVC Tech. Rep 24 (1998).
- [3] B. Shen, L. Si, R. Ji, and B. Liu, Robust nonnegative matrix factorization via L_1 norm regularization, arXiv:1204.2311, 2012.
- [4] B. Qin, C. Hu, S. Huang, Target/background classification regularized nonnegative matrix factorization for fluorescence unmixing, *IEEE Trans. Instrum. Meas.* 65 (2016) 874–889.
- [5] D. Lee, H. Seung, Learning the parts of objects by nonnegative matrix factorization, *Nature* 401 (6755) (1999) 788–791.
- [6] D. Cai, X. He, J. Han, T. Huang, Graph regularized non-negative matrix factorization for data representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (2011) 1548–1560.
- [7] D. Kong, C. Ding, H. Huang, Robust Nonnegative Matrix Factorization using $L_{2,1}$ -norm, in: *Proceedings of 20th ACM International Conference on Information and knowledge management*, 2011, pp. 673–682.
- [8] D. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* 52 (2006) 1289–1306.
- [9] E.T. Hale, W. Yin, Y. Zhang, Fixed-point continuation for L_1 -minimization: methodology and convergence, *SIAM J. Optim.* 19 (2008) 1107–1130.
- [10] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $L_{2,1}$ -norms minimization, *Adv. Neural Inf. Process. Syst.* 23 (2010) 1813–1821.
- [11] G. Gasalino, N.D. Buono, C. Mencar, Subtractive clustering for seeding non-negative matrix factorizations, *Inf. Sci.* 257 (2014) 369–387.
- [12] H. Liu, Z. Wu, D. Cai, T.S. Huang, Constrained non-negative matrix factorization for image representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (2012) 1299–1311.
- [13] J. Zhou, R. Liang, L. Zhao, L. Tao, C. Zou, Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutive non-negative matrix factorization, *Inf. Sci.* 257 (2014) 115–126.
- [14] L. Zhang, Z. Chen, M. Zheng, X. He, Robust non-negative matrix factorization, *Front. Electr. Electron. Eng. China* 6 (2011) 192–200.
- [15] M. Lu, X. Zhao, L. Zhang, F. Li, Semi-supervised concept factorization for document clustering, *Inf. Sci.* 331 (2016) 86–98.
- [16] N.D. Buono, G. Pio, Non-negative matrix tri-factorization for co-clustering: an analysis of the block matrix, *Inf. Sci.* 301 (2015) 13–26.
- [17] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, *IEEE Trans. Image Process.* 20 (2011) 2030–2048.

- [18] N. Gillis, Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices, *SIAM J. Mat. Anal. Appl.* 34 (2013) 1189–1212.
- [19] P.O. Hoyer, Non-negative sparse coding, in: *Proceeding of 12th IEEE Workshop Neural Network Signal Process.*, 2002, pp. 557–565.
- [20] S. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: *Proceeding of IEEE International Conference on Computer Vision Pattern Recognition.*, 2001, pp. 207–212.
- [21] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, *IEEE Trans. Neural Netw.* 17 (2006) 683–695.
- [22] S.J. Russell, P. Norvig, *Artificial Intelligence A Modern Approach*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, USA, 2003.
- [23] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2003) 1615–1618.
- [24] V. Monga, M.K. Mihçak, Robust and secure image hashing via non-negative matrix factorizations, *IEEE Trans. Inf. Forensics Secur.* 2 (2007) 376–390.
- [25] W. Liu, N. Zheng, Q. You, Nonnegative matrix factorization and its applications in pattern recognition, *Chin. Sci. Bull.* 51 (2006) 7–18.
- [26] W. Yang, C. Sun, L. Zhang, A multi-manifold discriminant analysis method for image feature extraction, *Pattern Recognit.* 44 (2011) 1649–1657.
- [27] X. Liu, S. Yan, H. Jin, Projective nonnegative graph embedding, *IEEE Trans. Image Process* 19 (2010) 1126–1137.
- [28] X. Huang, X. Zheng, W. Yuan, F. Wang, S. Zhu, Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization, *Inf. Sci.* 181 (2011) 2293–2302.
- [29] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, C. Yuan, Low-rank preserving projections, *IEEE Trans. Cybern.* (2015), doi:10.1109/TCYB.2015.2457611.
- [30] Y. Chen, J. Zhang, D. Cai, W. Liu, X. He, Nonnegative local coordinate factorization for image representation, *IEEE Trans. Image Process* 22 (2013) 969–979.
- [31] Y. Zhang, D. Yi, B. Wei, Y. Zhuang, A GPU-accelerated non-negative sparse latent semantic analysis algorithm for social tagging data, *Inf. Sci.* 281 (2014) 687–702.
- [32] Y. Wang, Y. Jia, Fisher non-negative matrix factorization for learning local features, in: *Proceeding of Asian Conference on Computer Vision*, 2004, pp. 1–6.
- [33] Y. Qian, S. Jia, J. Zhou, A. Robles-Kelly, Hyperspectral unmixing via $L_{1/2}$ sparsity-constrained nonnegative matrix factorization, *IEEE Trans. Geosci. Remote Sens.* 49 (2011) 4282–4297.
- [34] Z. Yuan, E. Oja, Projective Nonnegative Matrix Factorization for Image Compression and Feature Extraction, in: *Proceeding of 14th Scandinavian Conference on Image Analysis*, June 19 - 22, 2005, pp. 333–342.
- [35] Z. Yang, E. Oja, Linear and nonlinear projective nonnegative matrix factorization, *IEEE Trans. Neural Netw.* 21 (2010) 734–749.
- [36] Z. Xia, C. Ding, E. Chow, Robust Kernel Nonnegative Matrix Factorization, in: *Proceeding of IEEE 12th International Conference on Data Mining Workshops (ICDMW)*, 2012, pp. 522–529.
- [37] Z. Xu, H. Zhang, Y. Wang, Y.L.X.Y. Chang, $L_{1/2}$ regularizer, *Sci. China Ser. F* 53 (2010) 1159–1169.
- [38] Z. Lai, Y. Xu, Q. Chen, J. Yang, D. Zhang, Multilinear sparse principal component analysis, *IEEE Trans. Neural Netw. Learn. Syst.* 25 (2014) 1942–1950.
- [39] Z. Fan, Y. Xu, D. Zhang, Local linear discriminant analysis framework using sample neighbors, *IEEE Trans. Neural Netw.* 22 (2011) 1119–1132.
- [40] Z. Lai, W.K. Wong, Z. Jin, J. Yang, Y. Xu, Sparse approximation to the eigensubspace for discrimination, *IEEE Trans. Neural Netw. Learn. Syst.* 23 (2012) 1948–1960.
- [41] Z. Lai, Y. Xu, J. Yang, D. Zhang, Sparse tensor discriminant analysis, *IEEE Trans. Image Process* 22 (2013) 3904–3915.