This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

Nonnegative Discriminant Matrix Factorization

Yuwu Lu, Zhihui Lai, Yong Xu, *Senior Member, IEEE,* Xuelong Li, *Fellow, IEEE,* David Zhang, *Fellow, IEEE* and Chun Yuan

***Abstract*—Nonnegative Matrix Factorization (NMF), which aims at obtaining the nonnegative low-dimensional representation of data, has been received widely attentions. To obtain more effective nonnegative discriminant bases from the original NMF, a novel method called Nonnegative Discriminant Matrix Factorization (NDMF) is proposed for image classification in this paper. NDMF integrates the nonnegative constraint, orthogonality and discriminant information in the objective function. NDMF considers the incoherent information of both factors in standard NMF and is proposed to enhance the discriminant ability of the learned base matrix. NDMF projects the low-dimensional representation of the subspace of the base matrix to regularize the NMF for discriminant subspace learning. Based on the Euclidean distance metric and the generalized Kullback-Leibler (KL) divergence, two kinds of iterative algorithms are presented to solve the optimization problem. The between- and within-class scatter matrices are divided into positive and negative parts for the update rules and the proofs of the convergence are also presented. Extensive experimental results demonstrate the effectiveness of the proposed method in comparison to the state-of-the-art discriminant NMF algorithms.**

***Index Terms*—Nonnegative matrix factorization, maximum margin criterion, discriminative ability, face recognition.**

I. Introduction

In many data analysis tasks, a fundamental problem is to find a suitable low-representation of the data [1], [2], [3], [4]. A useful representation should discover the latent information embedded in the data set for further processing in the dimensionality reduction [5]. The most popular dimension reduction methods include principal component analysis (PCA) [6] and linear discriminant analysis (LDA) [7]. In recent years, matrix factorization methods have become popular and a number of matrix factorization methods have been proposed. Usually, matrix factorization techniques find two or more lower dimensional matrices whose product provides a good approximation to the original input data matrix [8]. For example, vector quantization (VQ) [9], singular value decomposition (SVD) [10], and nonnegative matrix factorization (NMF) [11] are some of the most popular matrix factorization techniques. VQ maps data vectors into binary vectors by exploiting a minimum distortion rule. SVD represents the original matrix in a low-rank approximation which is optimal in the sense of reconstruction error. Different from PCA, VQ and SVD, NMF aims to find two nonnegative matrices whose product is able to best approximate the original data matrix.

Previous researches [1], [11] indicate that the non-negativity constraint leads to a parts-based representation of the object. Some studies have shown that there are psychological and physiological evidences for parts-based representation in the human brain [12], [13], [14]. NMF only allows additive, not subtractive combination of the original data, and thus it is naturally favor to sparse, parts-based representation which is more robust than non-sparse, global representations [15].

In the past decade, a number of works related to NMF have been proposed. In [16], Li et al. imposed extra constraints to solve the localized and part-based decomposition by extending the standard NMF. Hazan et al. introduced an algorithm for a nonnegative 3D tensor factorization for the purpose of establishing local parts feature decomposition from an object class of images [17]. This algorithm uses nonnegative tensor factorization for handling the data encoded as high-order tensors. To encode discriminant information into NMF, Wang et al. proposed the Fisher-NMF (FNMF) [18], and Zafeiriou et al. extended it by adding an extra term of scatter difference to the objective function of NMF to obtain the discriminant subspace [19], [20]. The authors proposed a discriminative convex NMF for the classification of human brain tumours [21]. In addition, graph regularized NMF (GNMF) [8], constrained NMF (CNMF) [22], and projective nonnegative graph embedding (PNGE) [23], are widely used in nonnegative data factorization for image clustering and recognition, and the variations of the NMF-based methods can also be used in biomedical applications [24], [25]. In [26], the authors proposed a flexible nonnegative patch alignment framework for NMF related dimension reduction methods. Guan et al. [27] proposed an efficient NeNMF for optimizing NMF and its

Manuscript received Jul 6, 2015. This work was supported by the Natural Science Foundation of China (Grant Nos. 61203376, 61300032, 61573248, 61375012, 61362031, 61370163) and the Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20130329151843309, JCYJ20150324141711637, and JCYJ20140904154630436).

Y. Lu and C. Yuan are with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, P.R. China. (e-mail: luyuwu2008@163.com; yuanc@sz.tsinghua.edu.cn).

Z. Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, P.R. China. (e-mail: lai_zhi_hui@163.com).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, P.R. China. (e-mail: yongxu@ymail.com).

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P.R. China. (e-mail: xuelong_li@opt.ac.cn).

D. Zhang is with the Biometrics Research Center, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

extensions.

According to the Ref [28], NMF with additional constraints can be categorized into four classes, including sparse NMF, orthogonal NMF, manifold NMF and discriminant NMF. In this paper, we focus on the discriminant ability of the NMF-based methods. It is clear that integrating the discriminant information into NMF will benefit for improving the classification performance of the NMF-based methods [18], [19], [20]. However, the methods in [18], [19], [20] only used the discriminant information of the coefficient matrix V in NMF for facial expression recognition and thus the discriminant ability of these methods will be limited. Moreover, in the lower dimensional space, since the value of k (see (1)) is much smaller than M and N, the discriminant ability may be weaken when DNMF is used for general image recognition problem with large number of classes. The regularized terms in [18], [19], [20] implicitly depend on the representation coefficients matrix V, the derived base matrix U will emphasize more distinct localized properties instead of discrimination and the performance of DNMF will be finally degraded.

A suitable criterion used as the regularized term of the NMF-based methods should consider both the discriminative and reconstructive properties. Since the ratio form of LDA [7], [29] and its variations [30] [31] is difficult to solve when they are integrated to the NMF model in small sample size problem and the base/projection of the LDA is not orthogonal, it is not suitable to use them as the regularized term. However, maximum margin criterion (MMC) [32], [33], which maximizes the margin between classes and minimizes the within-class scatter, is a good choice to regularize the factors of NMF for enhancing the discriminant ability since the projections of MMC are orthogonal and contain strong discriminant information. Therefore, in order to effectively enhance the discriminant ability of NMF, in this paper, we propose a novel method, called Nonnegative Discriminant Matrix Factorization (NDMF) for image classification.

The contributions or the excellent properties of NDMF can be highlighted as followings:

1) Two iterative algorithms, which are proven to be convergent based on the Euclidean distance metric and the generalized Kullback-Leibler (KL) divergence, are proposed respectively. In the proposed methods, we divide the between-class and within-class scatter matrices into positive and negative parts, which is helpful for the proofs of the convergence.

2) Different from the previous discriminant NMF [18], [19], [20] in which the regularized terms work implicitly depending on the representation coefficient matrix V, NDMF combines the low-dimensional representation of the data with the subspace of U to regularize the NMF for discriminant subspace learning. That is, the base matrix U is directly related to the coefficient matrix.

3) The discriminant and localized properties as well as the orthogonality of the base matrix are fully taken into consideration and incorporated in one model. That is, NDMF combines nonnegative constraint, orthogonality and discriminant information in the objective function, in which both U and V are combined together to construct the regularized term.

The rest of the paper is organized as follows: Section II briefly reviews NMF and its related works. The detailed algorithms of NDMF and theoretical proof of the convergence of the algorithms in two formulations are given in Section III and Section IV, respectively. Extensive experimental results are presented in Section V. Section VI concludes the paper.

II. Brief Reviews of Nonnegative Matrix Factorization (NMF) and its Related Works

In order to the ease of reading, in this section, we briefly review some related works including NMF, LNMF, DNMF and MMC.

A. Nonnegative Matrix Factorization

NMF is different from VQ and SVD as it enforces the constraint that the elements of the factor matrices must be nonnegative.

NMF decomposes a matrix $X \in R^{M\times N}$ into two nonnegative matrixes $U=(u_1,\cdots,u_K)\in R^{M\times K}$ and $V=(v_1,\cdots,v_N)\in R^{K\times N}$. In [11], Lee et al. proposed two objective functions: the Euclidean distance and the KL divergence. The Euclidean distance based objective function is defined as:

$$O=\|X-UV\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. It means that each data point x_i is approximated by a linear combination of the columns of U, weighted by the components in V.

Lee and Seung proposed the following iterative update rules to obtain a local minimum of (1):

$$u_{jk}^{t+1}=u_{jk}^{t}\frac{(XV^T)_{jk}}{(UVV^T)_{jk}},\quad v_{ki}^{t+1}=v_{ki}^{t}\frac{(U^TX)_{ki}}{(U^TUV)_{ki}}. \tag{2}$$

B. Local Nonnegative Matrix Factorization

The additive parts learned by NMF are not necessarily localized, thus Li et al. [16] proposed a local nonnegative matrix factorization (LNMF) algorithm for learning spatially localized, parts-based representation of visual patterns.

The model of LNMF is defined as follows:

$$D_{LNMF}=D(X,UV)+\alpha\sum_{i,j}a_{ij}-\beta\sum_{i}b_{ii} \tag{3}$$

where $D(X,UV)$ is the KL divergence between X and UV, $A=[a_{ij}]=U^TU$ and $B=[b_{ij}]=VV^T$, α, β are positive constants. According to the authors, minimizing $\sum_{i,j}a_{ij}$ suppresses over decomposition of the bases matrix U, while maximizing $\sum_{i}b_{ii}$ encourages retaining components with important information.

C. Discriminant Nonnegative Matrix Factorization

Motivated by LNMF, the authors in [19] incorporated discriminant constraints inside the NMF decomposition. The cost function of DNMF is

$$D_{DNMF}=D(X,UV)+\alpha tr[S_w]-\beta tr[S_b] \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

where α , β are positive constants. S_w and S_b are within-class matrix and between-class matrix of the coefficient matrix V , respectively. S_w and S_b are defined as follows:

$$S_w = \sum_{r=1}^{C}\sum_{\rho=1}^{n_i}(\eta_\rho^{(r)} - \mu^{(r)})(\eta_\rho^{(r)} - \mu^{(r)})^T \quad (5)$$

$$S_b = \sum_{r=1}^{C} n_r(\mu^{(r)} - \mu)(\mu^{(r)} - \mu)^T \quad (6)$$

C is the number of class, and n_i is the number of samples in the *ith* class. The j th column of the database X is the ρ th image of the r th class. Thus, $j = \sum_{i=1}^{r-1} n_i + \rho$. The vector v_j that corresponds to the j th column of the matrix V , is the coefficient vector for the ρ th facial image of the r th class and will be denoted as $\eta_\rho^{(r)} = [\eta_{\rho,1}^{(r)} \cdots \eta_{\rho,K}^{(r)}]^T$. The mean vector of the vectors $\eta_\rho^{(r)}$ for the class r is denoted as $\mu^{(r)} = [\mu_1^{(r)} \cdots \mu_K^{(r)}]^T$ and the mean of all classes as $\mu = [\mu_1 \cdots \mu_K]^T$.

The regularized terms in [18], [19], [20] implicitly depend on the representation coefficients matrix V , the derived base matrix U will emphasize more on the reconstruction property instead of discrimination. Thus, the performance of the related works about DNMF will be degraded in the general recognition problems.

D. Maximum Margin Criterion

To enhance the discriminant information and avoid the small sample size problem in LDA, the authors in [33] proposed new feature extractors based on maximum margin criterion (MMC). The model of MMC is defined as below:

$$\max \sum_{k=1}^{d} w_k^T(\tilde{S}_b - \tilde{S}_w)w_k \quad (7)$$

$$\text{s.t. } w_k^T w_k - 1 = 0, \ k = 1, \cdots, d.$$

where $\tilde{S}_b$ and $\tilde{S}_w$ are between- and within-class scatter matrices in classical LDA. Therefore, the projections of MMC have strong discriminant ability. In addition, similar to PCA, the projections of MMC also have strong reconstructive ability since the projections are also orthogonal. This property can be integrated in the NMF model for enhancing the discriminant and reconstructive ability in a certain sense, which will be used as one of the motivations for the proposed method.

III. Nonnegative Discriminant Matrix Factorization

In this section, we introduce the proposed method, i.e. Nonnegative Discriminant Matrix Factorization (NDMF) for image classification. We first give the details of the motivation of the proposed method, and then present the objective function and the update rules. At last, we give the update rules of the proposed object function and analyze the connection of our method with gradient method.

A. The Motivations of the Proposed Method

In NMF methods, the original input data is divided into the product of two nonnegative matrices. Thus, both of the two nonnegative matrices contain the discriminative information of the original data. While, the related discriminative NMF [18], [19], [20] methods are only introduce the discriminative information of the coefficient matrix. That is, the related works of DNMF only introduce the within- and between-class scatter values with respect to the coefficient matrix V . The regularized terms in DNMF works only implicitly depend on the representation coefficients matrix, the derived base matrix emphasizes more on the reconstruction property instead of discrimination. Thus, the performance of the related works about DNMF would be degraded in the general recognition problems and the discriminant information exploited by DNMF is limited. To compensate for these deficiencies of DNMF works, we propose a novel method, which is named NDMF. Our idea is to effectively introduce discrimination information of the base matrix and the coefficient matrix into NMF for obtaining better classification performance, and at the same time to give the distinct localized parts for better representing the original data. Different from [18], [19], [20], the proposed method requires the orthogonality of the base matrix U to enhance the localized parts representation of the original data, and at the same time, explores the nonnegative discriminant subspace U , on which the data can obtain better separability. NDMF combines the base matrix and the coefficient matrix effectively to better use the discriminative information. Besides, NDMF also introduces the orthogonal regularization term of the base matrix to obtain better parts-based representation. The main advantages of NDMF can be concluded as follows:

Firstly, we define the within-class scatter matrix and between-class scatter matrix of the coefficient matrix V and combine them with the subspace (i.e. the base matrix) U together to enhance discriminant ability of NDMF. Secondly, since NMF cannot guarantee to derive the orthogonal base matrix, we directly introduce the orthogonality of the base matrix to enhance the distinct localized parts representation ability of NMF. Thirdly, according to the MMC [32], [33], we also tend to maximize the between-class scatter and minimize the within-class scatter of the low-dimensional representation on the subspace U so as to increase the discriminant ability when U is used for feature extraction and classification. This indicates that when the discriminant coefficient matrix V , which is usually viewed as the low-dimensional representation of the original data, combining with the base matrix U , can obtain stronger discriminant ability. This bridges the gap of the independency between the low-dimensional representation coefficient and the discriminant subspace used for classification. Thus, the relative independency of the DNMF on the coefficients is broken through and the discriminant ability of the base U can be greatly improved.

Compared with the existing discriminant NMF works, the proposed method has stronger classification performance and more flexible and generalization ability for image classification. When ignoring the orthogonal term of the base matrix, NDMF will be degraded into DNMF. That is, DNMF can be seen as a special case of NDMF. NDMF further extends the existing discriminant NMF methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

B. Objective Function

To improve the performance of NMF, the common strategy is to regularize the nonnegative parts in the decomposition. The regularized optimization model of NMF can be stated as follows:

$$\min_{U>0,V>0} D(X,UV)+\lambda\Omega(U,V) \tag{8}$$

where $\lambda \geq 0$ is a tradeoff parameter. The first term of (8) represents traditional NMF, and the last term of (8) denotes the regularized term either on U or V, or both U and V. The related works about DNMF [18], [19], [20] introduce the between- and within-class scatter matrices of the coefficient matrix V into NMF as the regularized term. From (5) and (6), we can find that DNMF [18], [19], [20] only introduces the discriminant information of the coefficient matrix V into NMF (i.e. the regularized term is only related to V).

From the motivations presented in the previous section, it is clear that we not only require U to be orthogonal, i.e. $\left\|U^TU-I\right\|^2$ is appended to the proposed model, but also define a new regularized term, which includes the following two items:

$$US_wU^T=\sum_{r=1}^{C}\sum_{\rho=1}^{n_i}(U\eta_\rho^{(r)}-U\mu^{(r)})(U\eta_\rho^{(r)}-U\mu^{(r)})^T \tag{9}$$

$$US_bU^T=\sum_{r=1}^{C}n_r(U\mu^{(r)}-U\mu)(U\mu^{(r)}-U\mu)^T \tag{10}$$

From (9) and (10), our method introduces the discriminant information in a completely different way from [18], [19], [20]. Our method not only introduces the orthogonality of the base matrix U but also combines the discriminant information of matrices V and U.

Using the Frobenius norm as the cost function, NDMF aims to minimize the following objective function

$$\begin{aligned} O&=\left\|X-UV\right\|^2+\lambda_1\left\|U^TU-I\right\|^2+\lambda_2(tr(US_wU^T)-tr(US_bU^T)) \\ &=\left\|X-UV\right\|^2+\lambda_1\left\|U^TU-I\right\|^2+\lambda_2tr(U(S_w-S_b)U^T) \\ &\text{s.t. } U\geq 0,\ V\geq 0 \end{aligned} \tag{11}$$

where λ_1 and λ_2 are regularization parameters, I is the identity matrix. S_w and S_b are within-class matrix and between-class matrix of the coefficient matrix V which are the same as (5) and (6), respectively. The objective function (11) not only introduces the orthogonality of NMF, but also improves the discriminant ability by combining the discriminant information of matrices V and U.

In order to satisfy the non-negativity constrains of NMF, we rewrite (11) as:

$$\begin{aligned} O&=\left\|X-UV\right\|^2+\lambda_1\left\|U^TU-I\right\|^2 \\ &+\lambda_2tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T) \end{aligned} \tag{12}$$

S_w^+ and S_b^+ denote the nonnegative parts of matrices S_w and S_b. $S_{|w|}^-$ and $S_{|b|}^-$ are the element wise absolute of the negative parts of matrices S_w and S_b.

It can be found that DNMF [18], [19], [20] introduces the discriminant information of the coefficient matrix V into NMF to enhance the discriminant ability of NMF. However, DNMF only regularizes the coefficient matrix V. Thus, the discriminant information exploited by DNMF is limited. Different from [18], [19], [20], the proposed method requires the orthogonality of the base matrix U to enhance the localized parts representation of the original data, and at the same time, explores the nonnegative discriminant subspace U, on which the data can obtain better separability.

The above optimization problem introduces the orthogonality of the base matrix U, which can enhance the localized parts representation ability of the original data. The key novelty of our proposed method is that the discriminant information of the coefficient matrix V is combined with the base matrix U as the regularized term to measure the discriminant information. Thus, we not only take into account the separability of the low-dimensional representation V but also the discriminant ability of subspace U.

C. Update Rules

The objective function in (12) is not convex in both variables U and V. It is hard to find the global minima for O. In the following, we describe an iterative updating algorithm to obtain the local optima solution of O following the similar way as [8].

Note that $tr(AB)=tr(BA)$, the objective function O can be rewritten as follows:

$$\begin{aligned} O&=tr\left((X-UV)(X-UV)^T\right)+\lambda_1\left\|U^TU-I\right\|^2 \\ &+\lambda_2tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T) \\ &=tr(XX^T)-2tr(XV^TU^T)+tr(UVV^TU^T) \\ &+\lambda_1tr(U^TU-I)^2+\lambda_2tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T) \end{aligned}$$

Let ψ_{ij} and ϕ_{ij} be the Lagrangian multipliers for constraints $u_{ij}\geq 0$ and $v_{ij}\geq 0$, respectively. We define matrix $\Psi=[\psi_{ij}]$, $\Phi=[\varphi_{ij}]$, then the Langrange function

$$\begin{aligned} L&=tr(XX^T)-2tr(XV^TU^T)+tr(UVV^TU^T)+\lambda_1tr(U^TU-I)^2 \\ &+\lambda_2tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T)+tr(\Psi U^T)+tr(\Phi V^T) \end{aligned}$$

The partial derivatives of L with respect to U and V are:

$$\begin{aligned} \frac{\partial L}{\partial U}&=2UVV^T-2XV^T+4\lambda_1UU^TU-4\lambda_1U \\ &+2\lambda_2U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))+\Psi \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{\partial L}{\partial V}&=2U^TUV-2U^TX+ \\ &2\lambda_2\nabla_v tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T)+\Phi \end{aligned} \tag{14}$$

Using the Karush-Kuhn-Tucker condition $\psi_{ij}u_{ij}=0$ and $\phi_{ij}v_{ij}=0$, we obtain the following equations for u_{ij} and v_{ij}:

$$\begin{aligned} &(XV^T)_{ij}u_{ij}-(UVV^T)_{ij}u_{ij}-2\lambda_1(UU^TU)_{ij}u_{ij}+2\lambda_1(U)_{ij}u_{ij} \\ &-\lambda_2(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-)))_{ij}u_{ij}=0 \end{aligned} \tag{15}$$

$$\begin{aligned} &(U^TX)_{ij}v_{ij}-(U^TUV)_{ij}v_{ij}- \\ &(\lambda_2\nabla_v tr(U((S_w^+-S_{|w|}^-)-(S_b^+-S_{|b|}^-))U^T))_{ij}v_{ij}=0 \end{aligned} \tag{16}$$

Equations (15) and (16) lead to the following update rules:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

$$u_{ij} \leftarrow u_{ij} \frac{(XV^T + 2\lambda_1 U + \lambda_2 U(S_{|w|}^- + S_b^+))_{ij}}{(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ij}} \quad (17)$$

$$v_{ij} \leftarrow v_{ij} \frac{(U^T X + \lambda_2 \nabla_v tr(U(S_{|w|}^- + S_b^+)U^T))_{ij}}{(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}} \quad (18)$$

We have the following theorem regarding the iterative update rules (17) and (18). We will show that the update rules of U and V in (17) and (18) will converge and the final solution will be a local optimum. Appendix A gives the detailed proof of Theorem 1.

Theorem 1. The objective function O in (1) is nonincreasing under the update rules in (17) and (18).

D. Connection with Gradient Method

In the proposed method, i.e. NDMF, the objective function (12) can be minimized by gradient descent algorithm [36]. Using gradient descent method, the additive update rules for (12) are:

$$u_{ij} \leftarrow u_{ij} + \eta_{ij} \frac{\partial O}{\partial p_{ij}},\ v_{ij} \leftarrow v_{ij} + \delta_{ij} \frac{\partial O}{\partial v_{ij}}.$$

η_{ij} and δ_{ij} are the parameters to control the step size of gradient descent. Let $\eta_{ij} = -u_{ij} / 2(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ij}$, we have

$$u_{ij} + \eta_{ij} \frac{\partial O}{\partial u_{ij}} = u_{ij} - \frac{u_{ij}}{2(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ij}} \frac{\partial O}{\partial u_{ij}}$$

$$= u_{ij} - \frac{u_{ij}(UVV^T - XV^T + 2\lambda_1 UU^T U - 2\lambda_1 U + \lambda_2 U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-)))_{ij}}{(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ij}}$$

$$= u_{ij} \frac{(XV^T + 2\lambda_1 U + \lambda_2 U(S_{|w|}^- + S_b^+))_{ij}}{(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ij}}$$

Similarly, let $\delta_{ij} = -v_{ij} / 2(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}$, we have

$$v_{ij} + \delta_{ij} \frac{\partial O}{\partial v_{ij}} = v_{ij} - \frac{v_{ij}}{2(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}} \frac{\partial O}{\partial v_{ij}}$$

$$= v_{ij} - \frac{v_{ij}((U^T UV)_{ij} - (U^T X)_{ij} + (\lambda_2 \nabla_v tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T))_{ij})}{(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}}$$

$$= v_{ij} \frac{(U^T X + \lambda_2 \nabla_v tr(U(S_{|w|}^- + S_b^+)U^T))_{ij}}{(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}}$$

The multiplicative updating rules in (17) and (18) are special cases of gradient descent [8], [37].

IV. Algorithm Minimizing the KL Divergence Cost

As described in [37], NMF can also be measured by the KL divergence. In this section, we also present the algorithm of NDMF based on the KL divergence.

A. The Objective Function and the Update Rules

The objective function of NDMF based the KL divergence is defined as:

$$O_{KL} = D(X \| UV) + \lambda_1 D(I \| U^T U) + \lambda_2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T)$$

$$= \sum_{i,j} (x_{ij} \log \frac{x_{ij}}{(UV)_{ij}} - x_{ij} + (UV)_{ij}) + \lambda_1 \sum_{i,j} (I_{ij} \log \frac{I_{ij}}{(U^T U)_{ij}} - I_{ij} + (U^T U)_{ij})$$

$$+ \lambda_2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T) \quad (19)$$

After some simplifications and elimination of pure data terms, we have

$$O_{KL} = \sum_{i,j} ((UV)_{ij} - x_{ij} \log(UV)_{ij}) + \lambda_1 \sum_{i,j} ((U^T U)_{ij} - I_{ij} \log(U^T U)_{ij})$$

$$+ \lambda_2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T) \quad (20)$$

Taking the derivative with respect to U and V, we have the following update rules which can achieve a local minima of (20):

$$u_{ij} \leftarrow u_{ij} \frac{\sum_b v_{jb} x_{ib} / \sum_k u_{ik} v_{kb} + 2\lambda_1 \sum_k u_{ik} / \sum_p u_{pi} u_{ip} + 2\lambda_2 (U(S_{|w|}^- + S_b^+))_{ij}}{\sum_b v_{jb} + 2\lambda_1 \sum_k u_{ik} + 2\lambda_2 (U(S_w^+ + S_{|b|}^-))_{ij}} \quad (21)$$

$$v_{ij} \leftarrow v_{ij} \frac{\sum_p u_{pi} x_{pi} / \sum_k u_{pk} v_{kj} + (\lambda_2 \nabla_v tr(U(S_{|w|}^- + S_b^+)U^T))_{ij}}{\sum_p u_{pi} + (\lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}} \quad (22)$$

Theorem 2. The objective function O_{KL} in (20) is nonincreasing under the update rules in (21) and (22). The objective function is invariant under these updates if and only if U and V are at a stationary point.

The proof of the Theorem 2 is presented in Appendix B.

B. Connection with Gradient Method

The objective function of NDMF based the KL divergence can also be minimized by gradient descent algorithm as we have shown in Section III.

Let the step size of gradient descent as follows:

$$\eta_{ij} = -u_{ij} / (\sum_b v_{jb} + 2\lambda_1 \sum_k u_{ik} + 2\lambda_2 (U(S_w^+ + S_{|b|}^-))_{ij})$$

$$\delta_{ij} = -v_{ij} / (\sum_p u_{pi} + (\lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij})$$

Then the additive update rules becomes

$$u_{ij} + \eta_{ij} \frac{\partial O_{KL}}{\partial u_{ij}} = u_{ij} + \eta_{ij} (\sum_b v_{jb} + 2\lambda_1 \sum_k u_{ik} +$$

$$2\lambda_2 (U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-)))_{ij} - (\sum_b (v_{jb} x_{ib} / \sum_k u_{ik} v_{kb}) + 2\lambda_1 \sum_k u_{ik} / \sum_p u_{pi} u_{ip}))$$

$$= u_{ij} \frac{\sum_b v_{jb} x_{ib} / \sum_k u_{ik} v_{kb} + 2\lambda_1 \sum_k u_{ik} / \sum_p u_{pi} u_{ip} + 2\lambda_2 (U(S_{|w|}^- + S_b^+))_{ij}}{\sum_b v_{jb} + 2\lambda_1 \sum_k u_{ik} + 2\lambda_2 (U(S_w^+ + S_{|b|}^-))_{ij}}$$

$$v_{ij} + \delta_{ij} \frac{\partial O_{KL}}{\partial v_{ij}} = v_{ij} + \delta_{ij} (\sum_p u_{pi} - \sum_p (u_{pi} x_{pj} / \sum_k u_{pk} v_{kj})$$

$$+ (\lambda_2 \nabla_v tr(U(S_w^+ - S_{|w|}^- + S_{|b|}^- - S_b^+)U^T))_{ij}) \quad .$$

$$= v_{ij} \frac{\sum_p u_{pi} x_{pi} / \sum_k u_{pk} v_{kj} + (\lambda_2 \nabla_v tr(U(S_{|w|}^- + S_b^+)U^T))_{ij}}{\sum_p u_{pi} + (\lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ij}}$$

We can obtain the multiplicative update rules in (21) and (22).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

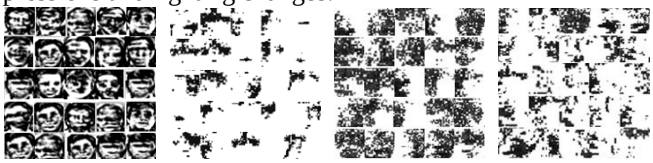
>

V. Experimental Results

In this section, we systematically evaluate the proposed Nonnegative Discriminant Matrix Factorization (NDMF) for face recognition and handwritten digit recognition tasks. We evaluate the performance of the proposed NDMF with six representative algorithms including LDA [7], NMF [11], LNMF [16], DNMF [19], PGDNMF [20], and manifold regularized discriminative NMF (MD-NMF) [34]. The experiments were conducted on the YALE [7], ORL [38], FERET [39], and CMU PIE [40] databases to evaluate the effectiveness of our algorithms for face recognition. And the MNIST [41] database is used to evaluate the proposed method for handwritten digit recognition task. All the images of these five databases with 256 gray levels per pixel, and the pixel values were scaled to be within [0, 1]. Fig. 1 shows example images of the YALE, ORL, FERET, PIE and MNIST databases.

Fig. 1. Example images from the YALE (first row), ORL (second row), FERET (third row), CMU PIE (fourth row) and MNIST databases (fifth row). Each row shows seven images captured at different situations.

We use the training set to learn basis/projection used for feature extraction and the test set is to report the accuracy of image classification. The NN classifier is used to calculate the percentage of samples in the test set that were correctly classified. Each experiment was conducted ten times and the averaged accuracy is reported. We set the maximum iterations number of the NMF-related methods as 200 and keep it constant in all the experiments. Fig. 2 presents the resulting feature basis components of NMF, LNMF, DNMF, and NDMF for subspace of dimension 100. From the basis images learned from different algorithms, we can find that: 1) NMF basis are less sparse than other algorithms; 2) LNMF can produce localized and parts-based components; 3) DNMF and NDMF are also can produce localized regions; 4) NDMF has better parts-based learning ability than DNMF, i.e. the learned base matrix of NDMF is sparser than the basis of DNMF. The main reason is that NDMF encodes the orthogonal property into NMF which can obtain stronger parts-based representation ability. The sparsity property of NDMF makes it potentially more robust to expressions and lighting changes.



(a) NMF (b) LNMF (c) DNMF (d) NDMF

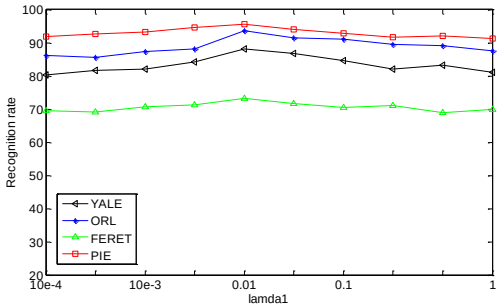
Fig. 2 Basis images of (a) NMF, (b) LNMF, (c) DNMF, and (d) NDMF. The data used are from the ORL database.

A. Parameter Selections

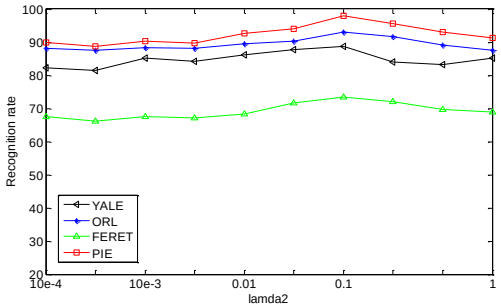
The compared methods and the related parameter selections of these methods are given as follows.

- Linear discriminant analysis (LDA) [7]. The number of feature dimension of LDA is set as the same one reported in [7].
- Discriminant nonnegative matrix factorization (DNMF) [19]. We set the parameters as reported in [19], and the best results are reported by choosing the two parameters in the range of [0.1, 0.5].
- Manifold regularized discriminative NMF (MD-NMF) [34]. There are three tradeoff parameters in MD-NMF. It is time consuming to select these parameters based on the grid search. We also set them the same as the ones in [34], i.e. $\alpha = 0.01$, $\beta = 0.1$, and $\gamma = 100$.
- Projected gradient discriminant NMF (PGDNMF) [20]. The selection of parameters in PGDNMF is all the same as DNMF.
- The proposed Nonnegative Discriminant Matrix Factorization (NDMF). In our experiments, we have tested values for λ_1 and λ_2 in the range [0, 1]. The best results have been obtained when choosing values in the range [0.01, 0.5]. We simply set $\lambda_1 = 0.01$, and $\lambda_2 = 0.1$ in our experiments.

In order to explore the variations of the performance of NDMF against the parameters, we randomly selected 5 images of each subject as training samples and the rest as test samples in the experiment. Fig. 3 (a) and (b) shows the variation of the recognition rate versus the values of lamda1 (λ_1) and lamda2 (λ_2), respectively. For all above learning algorithms, the performances are evaluated on the subspace with the dimension of 36, 49, 64, 81 and 100.



(a)



(b)

Fig. 3 Face recognition rates versus (a) λ_1 with λ_2 fixed, and (b) λ_2 with λ_1 fixed on the YALE, ORL, FERET and PIE database.

B. YALE Database

The YALE database [7] has 165 frontal view face images of 15 individuals. Each subject has 11 images with various facial expressions and lighting conditions. All images were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

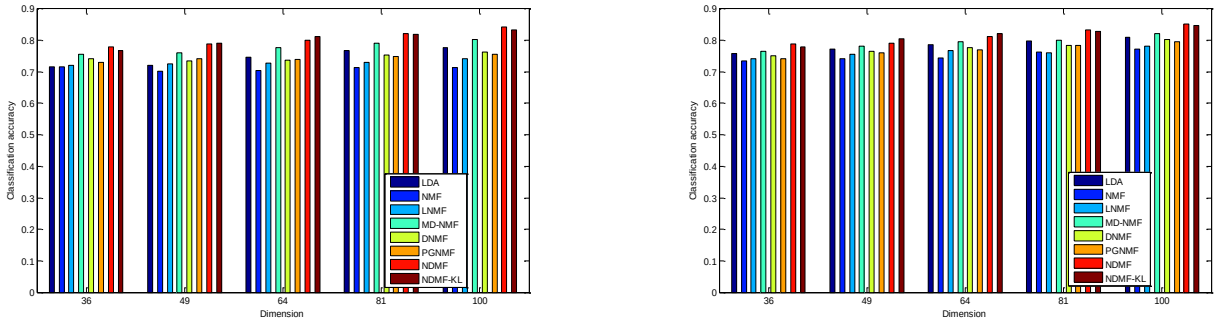
>

TABLE I THE PERFORMANCE (RECOGNITION RATE AND STANDARD DEVIATION) OF DIFFERENT METHODS ON THE YALE DATABASE

Training samples	LDA	NMF	LNMF	DNMF	PGDNMF	MD-NMF	NDMF	NDMF-KL
4	76.22 ±6.68	71.33 ±3.72	74.25 ±3.13	76.21 ±2.15	75.66 ±4.23	80.18 ±2.65	84.31 ±4.09	83.22 ±5.40
5	80.22 ±3.18	77.33 ±4.12	78.25 ±2.13	80.21 ±3.25	79.66 ±1.63	82.18 ±5.25	85.31 ±3.29	84.82 ±4.80

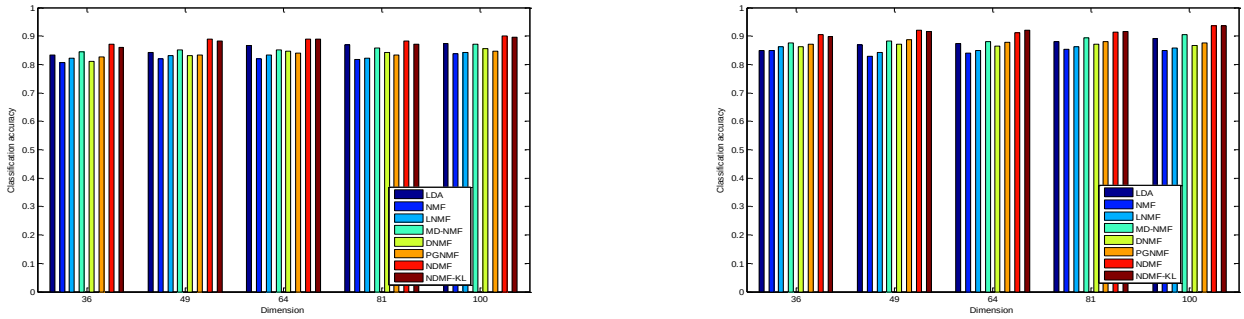
TABLE II THE PERFORMANCE (RECOGNITION RATE AND STANDARD DEVIATION) OF DIFFERENT METHODS ON THE ORL DATABASE

Training samples	LDA	NMF	LNMF	DNMF	PGDNMF	MD-NMF	NDMF	NDMF-KL
4	87.08 ±3.74	83.75 ±2.38	84.16 ±4.16	85.66 ±2.63	84.64 ±3.36	87.18 ±3.12	90.00 ±4.18	89.61 ±4.52
5	88.15 ±3.22	84.90 ±1.87	85.80 ±2.66	86.77 ±5.11	87.64 ±4.61	90.41 ±2.88	93.30 ±3.45	93.25 ±1.07



(a) (b)

Fig. 4. Face recognition rates over different feature dimensions for LDA, NMF, LNMF, MD-NMF, DNMF, PGDNMF, NDMF and NDMF-KL on the YALE database. (a) 4 training samples; (b) 5 training samples.



(a) (b)

Fig. 5. Face recognition rates over different feature dimensions for LDA, NMF, LNMF, MD-NMF, DNMF, PGDNMF, NDMF and NDMF-KL on the ORL database. (a) 4 training samples; (b) 5 training samples.

normalized to 40×40 pixel array and reshaped to a vector. We randomly selected 4 and 5 images from each subject to construct the training set, and the rest images were used for the test set.

The average recognition rates of the test set and the standard deviations of each method are shown in Table I. Fig. 4 (a) and (b) shows the variations of the recognition rates of all the methods with different subspace dimension when 4 and 5 images of each subject were selected as training samples, and the rest images as test samples, respectively. For each dimension, the mean accuracy calculated from ten random splits is reported. As can be seen from Table I and Fig. 4, NDMF obtains the best recognition rates in the experiments, which shows the robustness for the variations on facial expressions and lighting conditions.

C. ORL Database

The ORL database [38] has 40 distinct subjects, and each subject has 10 different images. All the subjects are in up-right, frontal position (with tolerance for some side movement). All images were taken in the same dark background and normalized to 40×40 pixel. We random selected different numbers (4, 5) of images from each subject to construct the training set, and the rest images were used for the test set. The performances of each method conducted on the ORL database are shown in Table II and Fig. 5. Table II lists the recognition rate of each method with 4 and 5 training samples. Fig. 5 (a) is the experimental results of all the comparison methods with 4 training samples of each subject. Fig. 5 (b) is the experimental results of all the comparison methods with 5 training samples of each subject. Again, NDMF performs better than the other comparison methods.

D. FERET Database

The FERET database [39] is used for evaluating face recognition algorithms displays diversity across gender, ethnicity, and age. The image sets were acquired without any restrictions imposed on facial expression and with at least two frontal images shot at different times during the same photo

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

>

TABLE III THE PERFORMANCE (RECOGNITION RATE AND STANDARD DEVIATION) OF DIFFERENT METHODS ON THE FERET DATABASE

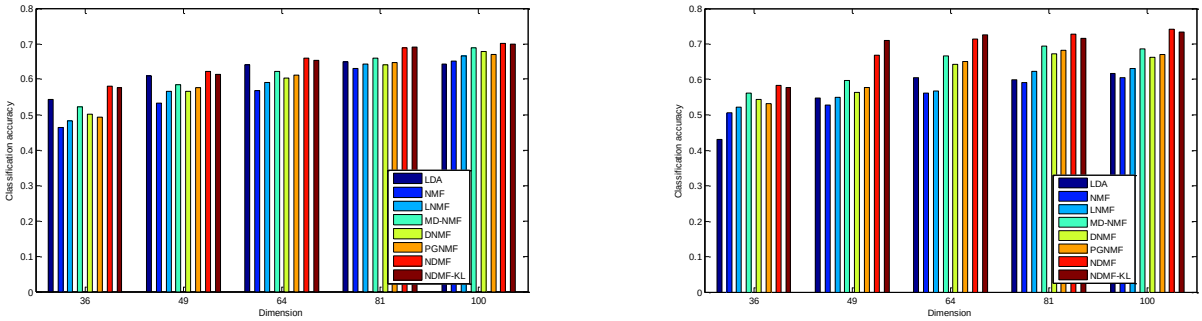
Training samples	LDA	NMF	LNMF	DNMF	PGDNMF	MD-NMF	NDMF	NDMF-KL
4	64.17 ±2.05	65.10 ±3.43	66.47 ±3.99	67.89 ±3.62	67.01 ±4.57	68.84 ±2.48	70.11 ±3.62	69.81 ±3.48
5	61.75 ±3.33	60.45 ±2.44	63.15 ±2.65	66.20 ±3.89	67.10 ±4.36	68.70 ±2.95	74.20 ±3.94	73.45 ±4.10

TABLE IV THE PERFORMANCE (RECOGNITION RATE AND STANDARD DEVIATION) OF DIFFERENT METHODS ON THE CMU PIE DATABASE

Training samples	LDA	NMF	LNMF	DNMF	PGDNMF	MD-NMF	NDMF	NDMF-KL
4	84.51 ±2.87	78.18 ±3.41	80.11 ±6.25	85.90 ±5.89	86.00 ±5.21	86.33 ±6.67	90.11 ±3.78	89.88 ±4.01
5	91.33 ±2.33	82.16 ±3.31	84.81 ±2.80	91.38 ±4.67	92.61 ±5.52	94.26 ±5.81	97.85 ±4.16	96.14 ±5.22

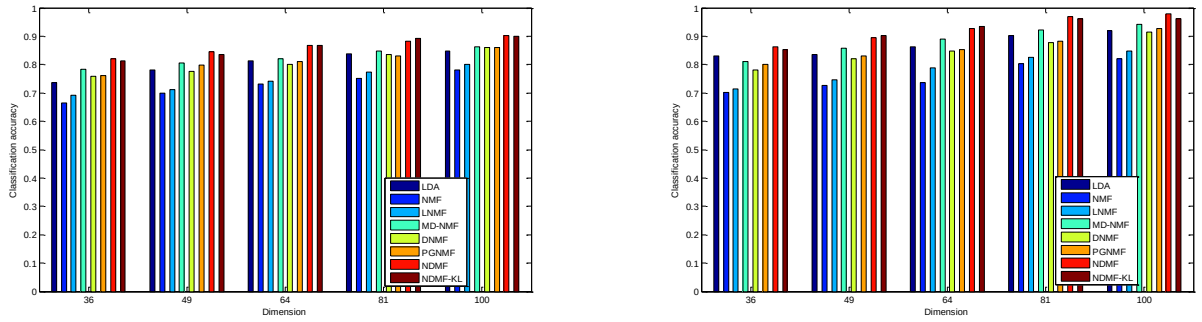
TABLE V THE PERFORMANCE (RECOGNITION RATE AND STANDARD DEVIATION) OF DIFFERENT METHODS ON THE MNIST DATABASE

Training samples	LDA	NMF	LNMF	DNMF	PGDNMF	MD-NMF	NDMF	NDMF-KL
3000	64.22 ±2.05	60.28 ±3.11	61.24 ±4.15	63.90 ±1.62	62.10 ±4.56	65.32 ±5.12	69.12 ±2.38	68.68 ±4.16
4000	67.34 ±1.69	62.06 ±4.51	64.24 ±3.52	66.18 ±3.68	64.20 ±2.24	68.06 ±4.16	72.80 ±2.22	71.04 ±5.21



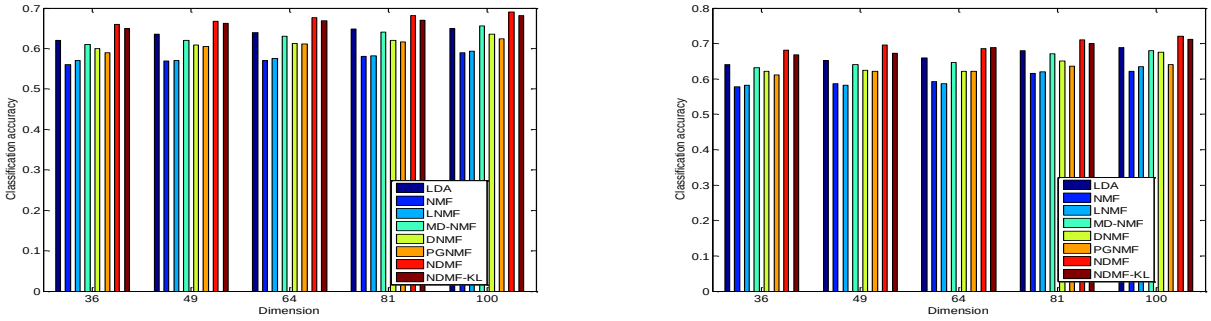
(a) (b)

Fig. 6. Face recognition rates over different feature dimensions for LDA, NMF, LNMF, MD-NMF, DNMF, PGDNMF, NDMF and NDMF-KL on the FERET database. (a) 4 training samples; (b) 5 training samples.



(a) (b)

Fig. 7. Face recognition rates over different feature dimensions for LDA, NMF, LNMF, MD-NMF, DNMF, PGDNMF, NDMF and NDMF-KL on the PIE database. (a) 4 training samples; (b) 5 training samples



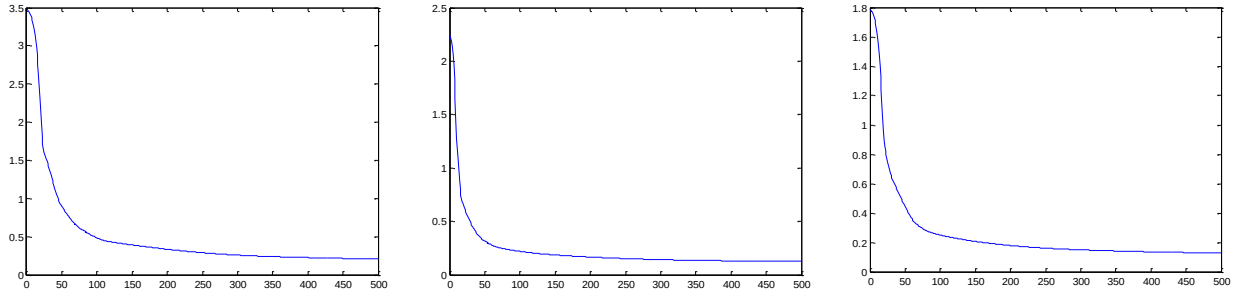
(a) (b)

Fig. 8. Digit recognition rates over different feature dimensions for LDA, NMF, LNMF, MD-NMF, DNMF, PGDNMF, NDMF and NDMF-KL on the MNIST database. (a) 3000 training samples; (b) 4000 training samples.

session. For the FERET face database, we used a subset made

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

>



(a) NMF (b) NDMF (c) NDMF-KL

Fig. 9. Convergence of NMF, NDMF and NDMF-KL on the YALE image database.

up of 1400 images from 200 individuals with each subject providing seven images [42]. Each image was normalized to 40 ×40 pixel array. We randomly selected different numbers (4, 5) of images from each subject to construct the training set, and the rest images were used for the test set.

Table III and Fig. 6 show the classification performance of the comparison methods conducted on the FERET database. Fig. 6 (a) and (b) are the experimental results of all the comparison methods with 4 and 5 training samples of each subject. Table III lists the classification rate of each method with different training samples. From Table III and Fig. 6, we can see that NDMF has higher classification rate than other compared methods.

E. CMU PIE Database

The CMU PIE [40] database contains 41368 face images collected from 68 subjects. Each subject has 13 images of different poses, 43 different illumination conditions, and with 4 different expressions. In our experiment, we selected a subset of 5 near frontal poses (C05, C07, C09, C25, and C29) and illuminations indexed as 08 and 11. Therefore, each subject has ten images. All images were normalized to 32×32 pixel array and reshaped to a vector.

Table IV lists the performance of different methods on the CMU PIE database. The recognition rates vs. the variations of the dimension are shown in Fig. 7. In the experiments, 4 and 5 images of each individual were randomly selected and used as training set, and the rest of images were used as test set. The experimental procedures are the same as Section V -B. As can be seen from Table IV and Fig. 7, NDMF obtains the best recognition rates in all the cases when there are variations in pose and illumination.

F. MNIST Database

In this subsection, in order to verify the performance of the proposed methods for handwritten digit recognition task, we conduct experiments on the MNIST database [41]. The MNIST database contains 60000 training images and 10000 test images, both drawn from the same distribution. All these images are size normalized and the size of each image is 28×28. The task is to classify each image into one of the ten digits and the writers of the training set and test set are different. In our experiments, we randomly select 3000 and 4000 images from 60000 training samples to construct the training set and randomly select 5000 test samples to construct the test set, respectively.

Fig. 8 and Table V illustrate the classification accuracy of the MNIST database. From Fig. 8 and Table V, we can see that the proposed methods still has the best recognition rate among all the compared methods.

G. Observations and Discussions

We obtain some observations based on the experimental results presented in the above sections:

(1) From experiments, we can find that DNMF and PGDNMF perform better than NMF and LNMF. Since DNMN and PGDNMF combine discriminant information in nonnegative factorization, both of them have higher recognition rates than NMF and LNMF. Our experiments also verify that LNMF has better classification performance than NMF, which is consists with [16].

(2) DNMF, PGDNMF and MD-NMF all encode discriminant information for classification. However, we can see that MD-NMF performs better than DNMF and PGDNMF in our experiments, due to the reason that MD-NMF not only introduces marginal information to NMF, but also introduces manifold structure of the data in the learning steps.

(3) Although NDMF does not introduce manifold regularization, it has higher classification accuracies than MD-NMF. The main reason is that NDMF combines the discriminant information both of the nonnegative matrices V and U as the regularized term. The experimental results show that NDMF which combines the discriminant information of the base matrix and the coefficient matrix could obtain good classification performance. This indicates that considering the separability of the low-dimensional representation V and the discriminant ability of subspace U does enhance the performance of the proposed algorithms.

H. Convergence Study

As proved in the previous sections, we used iterative update rules to obtain the local optima of NDMF no matter the cost measurement is the Frobenius norm or KL divergence. In this subsection, we experimentally show the convergent speed of our algorithms on the YALE database.

We compare the convergent speed of the original NMF algorithm and NDMF minimizing the F-norm cost (NDMF) and minimizing the KL divergence cost (NDMF-KL). Fig. 9 shows the convergence rates of the three algorithms on the YALE image database. In Fig. 9, the number of iterations is shown on the x-axis and the value of objective function is shown on the y-axis. We can see that both NDMF and NDMF-KL algorithms converge very fast.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

>

VI. Conclusion

In this paper, we proposed a novel method called Nonnegative Discriminant Matrix Factorization (NDMF) for image classification. Different from the NMF-based discriminant analysis methods, the proposed method not only introduces the orthogonality into NMF, but also combines the discriminant information of the coefficient matrix V and the base matrix U together as the regularized term to enhance the discriminant information. Our method maximizes the between-class scatter, and at the same time, minimizes the within-class scatter of the low-dimensional representation on the base matrix of NMF. We showed the NDMF method in two formulations and proposed update rules for both optimization algorithms. The classification performances of NDMF and NDMF-KL were tested on five standard image databases. The experimental results have demonstrated the effectiveness of the proposed method.

Appendix A

Proof of Theorem 1

To prove Theorem 1, we use an auxiliary function similar to the one used in Expectation-Maximization algorithm [35] and [22]. Definition 1 gives the definition of auxiliary function and Lemma 1 proves $F(u)$ is nonincreasing.

Definition 1[23]: Function $G(u,u')$ is an auxiliary function for $F(u)$ if the conditions $G(u,u') \geq (u), G(u,u) = F(u)$ are satisfied.

Lemma 1[22]: If G is an auxiliary function of F, then F is nonincreasing under the update

$$u^{(t+1)} = \arg\min_{u} G(u, u^{(t)}) \tag{23}$$

For any element u_{ab} in U, let $F_{u_{ab}}$ denote the part of O relevant to u_{ab}. We prove that $F_{u_{ab}}$ is nonincreasing under the update step of (17) by defining an auxiliary function. Lemma 2 defines an auxiliary function of $F_{u_{ab}}$.

Lemma 2: Function

$$G(u,u^t_{ab}) = F_{u_{ab}}(u^t_{ab}) + F'_{u_{ab}}(u^t_{ab})(u-u^t_{ab}) + \frac{1}{2}F''_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab})^2 + \frac{1}{3!}F^{(3)}_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab})^3 + \frac{(\lambda_1 U)_{ab}}{u^t_{ab}}(u-u^t_{ab})^4 \tag{24}$$

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

Proof: $G(u,u) = F_{u_{ab}}(u)$ is obvious, so we only need to show that $G(u,u^t{}_{ab}) \geq F_{u_{ab}}(u)$. For doing this, we compare $G(u,u^t{}_{ab})$ in (24) with the Taylor series expansion of $F_{u_{ab}}(u)$:

$$F_{u_{ab}}(u) = F_{u_{ab}}(u^t{}_{ab}) + F'_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab}) + \frac{1}{2}F''_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab})^2 + \frac{1}{3!}F^{(3)}_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab})^3 + \frac{1}{4!}F^{(4)}_{u_{ab}}(u^t{}_{ab})(u-u^t{}_{ab})^4 \tag{25}$$

where $F''_{u_{ab}}$, $F^{(3)}_{u_{ab}}$ and $F^{(4)}_{u_{ab}}$ are the second, third and fourth order derivative with respect to U, respectively. It is easy to check that

$$\begin{aligned} F'_{u_{ab}} &= \left(\frac{\partial O}{\partial U}\right)_{ab} = (2UVV^T - 2XV^T + 4\lambda_1 UU^T U - 4\lambda_1 U \\ &\quad + 2\lambda_2 U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-)))_{ab}, \\ F''_{u_{ab}} &= (2VV^T + 12\lambda_1 U^T U - 4\lambda_1 I + 2\lambda_2((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))^T)_{bb} \\ F^{(3)}_{u_{ab}} &= 24\lambda_1 U \\ F^{(4)}_{u_{ab}} &= 24\lambda_1 \end{aligned} \tag{26}$$

Putting (26) into (25) and comparing with (24), we can see that, to show $G(u,u^t_{ab}) \geq F_{u_{ab}}(u)$, it is equivalent to prove

$$\frac{(\lambda_1 U)_{ab}}{u^t_{ab}} \geq \frac{1}{4!}F^{(4)}_{u_{ab}}(u^t{}_{ab}) \tag{27}$$

To prove the above inequality, we have

$$(\lambda_1 U)_{ab} \geq \lambda_1 u^t_{ab} \tag{28}$$

Thus, (27) holds and $G(u,u^t_{ab}) \geq F_{u_{ab}}(u)$. □

Let $F_{v_{ab}}$ denotes the part of O relevant to v_{ab}. Lemma 3 defines an auxiliary function regarding $F_{v_{ab}}$ which proves that $F_{v_{ab}}$ is nonincreasing under the update step of (18).

Lemma 3: Function

$$G(v,v^t_{ab}) = F_{v_{ab}}(v^t_{ab}) + F'_{v_{ab}}(v^t_{ab})(v-v^t_{ab}) + \frac{(U^TUV + \lambda_2\nabla_v^2 tr(U(S_w^+ + S_{|w|}^- + S_b^+ + S_{|b|}^-)U^T)V)_{ab}}{v^t_{ab}}(v-v^t_{ab})^2 \tag{29}$$

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

Proof: $G(v,v) = F_{v_{ab}}(v)$ is obvious, so we only need to show that $G(v,v^t{}_{ab}) \geq F_{v_{ab}}(v)$ by comparing $G(v,v^t{}_{ab})$ in (29) with the Taylor series expansion of $F_{v_{ab}}(v)$:

$$F_{v_{ab}}(v) = F_{v_{ab}}(v^t{}_{ab}) + F'_{v_{ab}}(v^t{}_{ab})(v-v^t{}_{ab}) + \frac{1}{2}F''_{v_{ab}}(v^t{}_{ab})(v-v^t{}_{ab})^2 \tag{30}$$

where $F''_{v_{ab}}$ is the second order derivative with respect to V. It is easy to check that

$$\begin{aligned} F'_{v_{ab}} &= \left(\frac{\partial O}{\partial V}\right)_{ab} = (2U^TUV - 2U^TX + \\ &\quad 2\lambda_2\nabla_v tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T))_{ab} \\ F''_{v_{ab}} &= (2U^TU + 2\lambda_2\nabla_v^2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T))_{aa} \end{aligned} \tag{31}$$

Putting (31) into (30) and comparing it with (29), we can see that, to show $G(v,v^t{}_{ab}) \geq F_{v_{ab}}(v)$ is equivalent to prove

$$\frac{(U^TUV + \lambda_2\nabla_v^2 tr(U(S_w^+ + S_{|w|}^- + S_b^+ + S_{|b|}^-)U^T)V)_{ab}}{v^t_{ab}} \geq \frac{1}{2}F''_{v_{ab}}(v^t{}_{ab}) \tag{32}$$

To prove (32), we have

$$(U^TUV)_{ab} \geq v^t_{ab}(U^TU)_{aa}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

$$(\lambda_2\nabla_v^2 tr(U(S_w^+ + S_{|w|}^- + S_b^+ + S_{|b|}^-)U^T)V)_{ab}$$
$$\geq v_{ab}^t(\lambda_2\nabla_v^2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T))_{aa}$$

Thus, (32) holds and $G(v, v^t{}_{ab}) \geq F_{v_{ab}}(v)$. □

Proof of Theorem 1: Putting $G(u, u_{ab}^t)$ in (24) into (23) and putting $G(v, v_{ab}^t)$ in (29) into (23), we obtain:

$$u_{ab}^{(t+1)} = \arg\min_u G(u, u_{ab}^{(t)}) = u_{ab}{}^t \frac{(XV^T + 2\lambda_1 U + \lambda_2 U(S_{|w|}^- + S_b^+))_{ab}}{(UVV^T + 2\lambda_1 UU^T U + \lambda_2 U(S_w^+ + S_{|b|}^-))_{ab}}$$

$$v_{ab}^{(t+1)} = \arg\min_v G(v, v_{ab}^{(t)}) = v_{ab}{}^t \frac{(U^T X + \lambda_2 \nabla_v tr(U(S_{|w|}^- + S_b^+)U^T))_{ab}}{(U^T UV + \lambda_2 \nabla_v tr(U(S_w^+ + S_{|b|}^-)U^T))_{ab}}$$

Since (24) and (29) are auxiliary functions, $F_{u_{ab}}$ and $F_{v_{ab}}$ are nonincreasing under the update rules.

APPENDIX B
PROOF OF THEOREM 2

Let $F(u)$ denote the part of O_{KL} relevant to u. Similar to the proof of Theorem 1, we define the auxiliary function regarding u as follows.

$$F = \sum_{i,j}((UV)_{ij} - x_{ij}\log(UV)_{ij}) + \lambda_1\sum_{i,j}((U^TU)_{ij} - I_{ij}\log(U^TU)_{ij})$$
$$+\lambda_2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T) \quad (33)$$

In order to simplify the proof of Theorem 2, we divide (33) in three parts

$$F_1 = \sum_{i,j}((UV)_{ij} - x_{ij}\log(UV)_{ij}) \quad (34)$$

$$F_2 = \lambda_1\sum_{i,j}((U^TU)_{ij} - I_{ij}\log(U^TU)_{ij}) \quad (35)$$

$$F_3 = \lambda_2 tr(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T) \quad (36)$$

Let $F_1(u)$, $F_2(u)$, and $F_3(u)$ denote the part of F_1, F_2 and F_3 relevant to u, respectively. Similar to the proof of Theorem 1, we define auxiliary functions of $F_1(u)$, $F_2(u)$, and $F_3(u)$ by Lemmas 4-6.

Lemma 4: Function

$$G_1(u,u^t) = \sum_{i,j}(\sum_k u_{ik}v_{kj} - x_{ij}\sum_k(\frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}}\log u_{ik}v_{kj} - \frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}}\log\frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}})) \quad (37)$$

is an auxiliary function of (34) regarding u:

$$F_1(u) = \sum_{i,j}(\sum_k u_{ik}v_{kj} - x_{ij}\log\sum_k u_{ik}v_{kj})$$

Proof: Obviously, $G_1(u,u) = F_1(u)$, we will show that $G_1(u,u^t) \geq F_1(u)$. We have the following inequality

$$-\log\sum_k u_{ik}v_{kj} \leq -\sum_k \alpha_k \log\frac{u_{ik}v_{kj}}{\alpha_k},$$

Setting $\alpha_k = \frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}}$, we obtain

$$-\log\sum_k u_{ik}v_{kj} \leq -\sum_k \frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}}(\log u_{ik}v_{kj} - \log\frac{u^t{}_{ik}v_{kj}}{\sum_k u^t{}_{ik}v_{kj}})$$

From this inequality, we obtain that $G_1(u,u^t) \geq F_1(u)$. □

Lemma 5: Function

$$G_2(u,u^t) = \sum_{i,j}(\sum_k u_{ki}u_{ik} -$$
$$I_{ij}\sum_k(\frac{u^t{}_{ki}u_{ik}}{\sum_k u^t{}_{ki}u_{ik}}\log u_{ik}u_{ki} - \frac{u^t{}_{ki}u_{ik}}{\sum_k u^t{}_{ki}u_{ik}}\log\frac{u^t{}_{ki}u_{ik}}{\sum_k u^t{}_{ki}u_{ik}})) \quad (38)$$

is an auxiliary function of (35) regarding u:

$$F_2(u) = \sum_{i,j}(\sum_k u_{ki}u_{ik} - I_{ij}\log\sum_k u_{ki}u_{ik}).$$

Proof: The proof of Lemma 5 all the same as the proof of Lemma 4, so we omit it here. □

Lemma 6: Function

$$G_3(u,u^t) = tr\sum_{i,k}(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T)_{ik}$$
$$+tr\sum_{i,k}(U(S_w^+ + S_{|w|}^- + S_b^+ + S_{|b|}^-)U^T)_{ik}(u-u^t)^2 \quad (39)$$

is an auxiliary function of (36) regarding u:

$$F_3(u) = tr\sum_{i,k}(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T)_{ik}.$$

Proof: Obviously, $G_3(u,u) = F_3(u)$, and $G_3(u,u^t) \geq F_3(u)$. □

From Lemma 4 to Lemma 6, we can show that the update rule of (31) and (32) are exactly the updates of (19). Let $F(v)$ denote the part of O_{KL} relevant to v. We define the auxiliary function regarding v as follows:

$$F(v) = F_1(v) + F_2(v).$$

Then we have Lemmas 7-8 which give the auxiliary functions of $F_1(v)$ and $F_2(v)$.

Lemma 7: Function

$$G_1(v,v^t) = \sum_{i,j}(\sum_k u_{ik}v_{kj} -$$
$$x_{ij}\sum_k(\frac{u_{ik}v^t{}_{kj}}{\sum_k u_{ik}v^t{}_{kj}}\log u_{ik}v_{kj} - \frac{u_{ik}v^t{}_{kj}}{\sum_k u_{ik}v^t{}_{kj}}\log\frac{u_{ik}v^t{}_{kj}}{\sum_k u_{ik}v^t{}_{kj}})) \quad (40)$$

is an auxiliary function of (34) regarding v:

$$F_1(v) = \sum_{i,j}(\sum_k u_{ik}v_{kj} - x_{ij}\log\sum_k u_{ik}v_{kj})$$

The proof of Lemma 7 is essentially similar to the proof of Lemma 4, we omit it here due to space limitation.

Lemma 8: Function

$$G_2(v,v^t) = tr\sum_{i,k}(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T)_{ik}$$
$$+tr\sum_{i,k}(U(S_w^+ + S_{|w|}^- + S_b^+ + S_{|b|}^-)U^T)_{ik}(v-v^t)^2$$

is an auxiliary function of (36) regarding v:

$$F_2(v) = tr\sum_{i,k}(U((S_w^+ - S_{|w|}^-) - (S_b^+ - S_{|b|}^-))U^T)_{ik}.$$

The proof of Lemma 8 is essentially similar to the proof of Lemma 6, we omit it here due to space limitation.

Proof of Theorem 2: Setting the gradient of (34-36) and (40) to zero:

$$\frac{\partial G_1(u,u^t)}{\partial u_{ij}} + \frac{\partial G_2(u,u^t)}{\partial u_{ij}} + \frac{\partial G_3(u,u^t)}{\partial u_{ij}} = 0$$

and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

$$\frac{\partial G(v,v^t)}{\partial v_{ij}}=0$$

We can obtain

$$u_{ij}=u_{ij}\frac{\sum_b v_{jb}x_{ib}/\sum_k u_{ik}v_{kb}+2\lambda_1\sum_k u_{ik}/\sum_p u_{pi}u_{ip}+2\lambda_2(U(S_{|w|}^{-}+S_b^{+}))_{ij}}{\sum_b v_{jb}+2\lambda_1\sum_k u_{ik}+2\lambda_2(U(S_w^{+}+S_{|b|}^{-}))_{ij}}$$

$$v_{ij}=v_{ij}\frac{\sum_p u_{pi}x_{pi}/\sum_k u_{pk}v_{kj}+(\lambda_2\nabla_v tr(U(S_{|w|}^{-}+S_b^{+})U^T))_{ij}}{\sum_p u_{pi}+(\lambda_2\nabla_v tr(U(S_w^{+}+S_{|b|}^{-})U^T))_{ij}}.$$

Theorem 2 guarantees that the update rules in (21) and (22) converge and the final solution will be a local optimum. □

REFERENCES

[1] D. Cai, X. He, X. Wu, and J. Han, "Nonnegative matrix factorization on manifold," in Proc. 8th IEEE International Conference on Data Mining, Pisa, Italy, pp. 63–72, 2008.

[2] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear Sparse Principal Component Analysis," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 10, pp. 1942-1950, Oct. 2014.

[3] M. Jian, K. Lam, J. Dong, and L. Shen, "Visual-Patch-Attention-Aware Saliency Detection," IEEE Trans. Cybern., vol. 45, no. 8, pp. 1575–1586, Aug. 2015.

[4] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-Rank Preserving Projections," IEEE Trans. Cybern., Doi: 10.1109/TCYB.2015.2457611.

[5] P. O. Hoyer, "Nonnegative matrix factorization with sparseness constraints," J. Mach. Learn. Res., vol. 5, no. 37, pp. 1457–1469, 2004.

[6] I. Jolliffe, Principle Component Analysis . New York: Springer-Verlag, 1986.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711-720, Jul. 1997.

[8] D. Cai, X. He, X. Wu, J. Han and T. S. Huang, "Graph Regularized Nonnegative Matrix factorization for data representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1548-1560, Aug. 2011.

[9] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. Kluwer Acad. Press, 1992.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. NJ: Wiley-Interscience, Hoboken, 2000.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, vol.401, pp. 788–791, 1999.

[12] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," Annu. Rev. Neurosci. , vol. 19, pp. 577–621, Mar. 1996.

[13] S. E. Palmer, "Hierarchical structure in perceptual representation," Cognit. Psychol., vol. 9, no. 4, pp. 441–474, 1977.

[14] M. W. O. E. Wachsmuth and D. I. Perrett, "Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque," Cerebral Cortex, vol. 4, no. 5, pp. 509–522, 1994.

[15] M. Heiler and C. Schnörr, "Learning sparse representations by nonnegative matrix factorization and sequential cone programming," J. Mach. Learn. Res., vol. 7, pp. 1385–1407, Jul. 2006.

[16] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in Proc. IEEE Computer Vision and Pattern Recognition, pp. 207–212, 2001.

[17] T. Hazan, S. Polak, and A. Shashua, "Sparse image coding using a 3d nonnegative tensor factorization," in Proc. Int. Conf. Computer Vision, vol. 1, pp. 50–57, 2005.

[18] Y. Wang, Y. Jiar, C. Hu, and M. Turk, "Fisher nonnegative matrix factorization for learning local features," in Proc. Asian Conf. Computer Vision, pp. 1–8, 2004.

[19] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting Discriminant Information in Nonnegative Matrix Factorization with Application to Frontal Face Verification," IEEE Trans. Neural Netw., vol. 17, no. 3, pp. 683-695, May 2006.

[20] I. Kotsia, S. Zafeiriou, and I. Pitas, "A Novel Discriminant Nonnegative Matrix Factorization Algorithm with Applications to Facial Image Characterization Problems," IEEE Transaction Information Forensics and Security, vol. 2, no. 3, pp. 588-595, Sept. 2007.

[21] Vilamala A, Lisboa P J G, Ortega-Martorell S, Vellido A, "Discriminant Convex Non-negative Matrix Factorization for the classification of human brain tumours," Pattern Recognition Letters, vol. 34, pp. 1734-1747, 2013.

[22] H. F. Liu, Z. H. Wu, D. Cai, and T. S. Huang, "Constrained Nonnegative Matrix Factorization for Image Representation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 16, pp. 1299–1311, 2012.

[23] X. B Liu, S. C. Yan, and H. Jin, "Projective Nonnegative Graph Embedding", IEEE Transactions on Image Processing, vol. 19, no. 5, pp. 1126-1137, May 2010.

[24] A. Heger and L. Holm, "Sensitive pattern discovery with 'fuzzy alignments of distantly related proteins," Bioinformatics, vol. 19, no. 1, pp. 130–137, 2003.

[25] P. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," Genome Res., vol. 13, pp. 1706–1718, 2003.

[26] N. Guan, D. Tao, Z. Luo, B. Yuan, "Non-negative patch alignment framework," IEEE Transactions on Neural Networks, vol. 22, pp. 1218-1230, 2011.

[27] N. Guan, D. Tao, Z. Luo, B. Yuan, "NeNMF: an optimal gradient method for nonnegative matrix factorization," IEEE Transactions on Signal Processing, vol. 60, pp. 2882-2898, 2012.

[28] Y. X. Wang, Y. J. Zhang, "Nonnegative Matrix Factorization: A Comprehensive Review," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1336-1353, June 2013.

[29] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 755-761, Apr. 2009.

[30] T.-K. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 3, pp. 318-327, Mar. 2005.

[31] M. Ordowski and G. G. L. Meyer, "Geometric linear discriminant analysis for pattern recognition," Pattern Recognition, vol. 37, pp. 421-428, 2004.

[32] Q. Gu and J. Zhou, "Two dimensional Maximum Margin Criterion," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1621–1624, 2009.

[33] H. Li, T. Jiang, K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," IEEE Trans. Neural Netw., vol. 17, no. 1, pp. 157-165, Jan. 2006.

[34] N. Guan, D. C. Tao, Z. G. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent", IEEE Transactions on Image Processing, vol. 20, no. 7, July 2011.

[35] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Joural the Royal Statistics Society, vol. 39, no. 1, pp. 1-38, 1977.

[36] J. Kivinen and M. Warmuth, "Additive versus exponentiated gradient updates for linear prediction," in Proc. 27th Annu. ACM Symp. Theory Comput., pp. 209-218, 1995.

[37] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Adv. Neural INf. Process. Syst., pp. 556-562, 2001.

[38] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in Proc. 2nd IEEE Int. Workshop Appl. Comput. Vis Sarasota, FL, Dec., pp. 138-142, 1994.

[39] Available: http://www.itl.nist.gov/iad/humanid/feret/feret_master.html.

[40] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1615-1618, Dec. 2003.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient Based Learning Applied to Document Recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[42] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539779, IEEE Transactions on Circuits and Systems for Video Technology

>

Yuwu Lu received the B.S., M.S. and Ph.D. degrees in 2008, 2011 and 2015, respectively. He is a Post-Doctoral Fellow with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. He is the author of more than 12 scientific papers in pattern recognition and computer vision. His current research interests include pattern recognition and machine learning.

Zhihui Lai received the B.S degree in mathematics from South China Normal University, M.S degree from Jinan University, and the PhD degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a research associate, postdoctoral fellow and research fellow at The Hong Kong Polytechnic University since 2010. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He serves as an associate editor on International Journal of Machine Learning and Cybernetics. For more information, the readers are referred to the website (http://www.scholat.com/laizhihui).

Yong Xu (M'06-SM'15) was born in Sichuan, China, in 1972. He received the B.S. and M.S. degrees in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005. Currently, he is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, biometrics, machine learning, image processing, and video analysis.

Xuelong Li (M'02-SM'07-F'12) is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.

David Zhang (F'08) received the bachelor's degree in computer science from Peking University, Beijing, China, the M.Sc. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, and then an Associate Professor at the Academia Sinica, Beijing. He is currently a Chair Professor at the Hong Kong Polytechnic University, Hong Kong, where he is also the Founding Director of the Biometrics Technology Centre (UGC/CRC), supported by the Hong Kong SAR Government in 1998. He also serves as the Visiting Chair Professor with Tsinghua University, and an Adjunct Professor with Shanghai Jiao Tong University, Shanghai, China, Peking University, Harbin Institute of Technology, and the University of Waterloo. He has authored more than ten books and 200 journal papers.

Dr. Zhang is the Founder and Editor-in-Chief of the International Journal of Image and Graphics, the Book Editor for Springer International Series on Biometrics, and an Associate Editor of more than ten international journals including the IEEE Transactions and Pattern Recognition. He was also the organizer of the 1st International Conference on Biometrics Authentication, and is the Technical Committee Chair of IEEE CIS. He is a Croucher Senior Research Fellow, Distinguished Speaker of the IEEE Computer Society, and a fellow of IAPR.

Chun Yuan is currently an Associate Professor in the Division of Information Science and Technology in Graduate school at Shen Zhen, Tsinghua University. He received the M.S. and Ph.D. degrees from the Department of Computer Science and technology, Tsinghua University, Beijing, China, in 1999 and 2002, respectively. He once worked at the INRIA-Rocquencourt, Paris, France, as a Post-doc research fellow from 2003 to 2004. In 2002, he worked at Microsoft Research Asia, Beijing, China, as an intern. His research interests include computer vision, machine learning, video coding and processing, cryptography and digital rights management.