ROBUST NONNEGATIVE MATRIX FACTORIZATION WITH DISCRIMINABILITY FOR IMAGE REPRESENTATION

Yuchen Guo[1], *Guiguang Ding*[1], *Jile Zhou*[2]

[1]MOE Key Laboratory for Information System Security; TNList; School of Software; Tsinghua Univeristy
[2]Sohu Inc.; {yuchen.w.guo,jile.zip}@gmail.com,dinggg@tsinghua.edu.cn

ABSTRACT

Due to its psychological and physiological interpretation of naturally occurring data, Nonnegative Matrix Factorization (NMF) has attracted considerable attention for learning effective representation for images. And its graph-regularized extensions have shown promising results by exploiting the low dimensional manifold structure of data. Actually, their performance can be further improved because they still suffer from several important problems, i.e., sensitivity to noise in data, trivial solution problem, and ignoring the discriminative information. In this paper, we propose a novel method, referred to as **R**obust **N**onnegative **M**atrix **F**actorization with **D**iscriminability (RNMFD), for image representation, which can effectively and simultaneously cope with problems mentioned above by imposing a sparse noise matrix for data reconstruction and approximate orthogonal constraints. We carried out extensive experiments on five benchmark image datasets and the results demonstrate the superiority of our RNMFD in comparison with several state-of-the-art methods.

1. INTRODUCTION

Images are always represented as vectors of very high dimensionality, which makes it difficult to apply statistical techniques for visual analysis [1]. Thus we hope to find low dimensional representation for images that can capture important information of original data. Nonnegative Matrix Factorization (NMF) [2], which aims to find low-rank nonnegative matrices whose product can approximate the original data matrix, has attracted considerable attention as it can learn parts-based representation for image which has psychological and physiological interpretation of naturally occurring data [3]. It's widely used in image representation [4, 5]. Some variants of NMF have been also proposed, like SparseNMF [6], OTNMF [7], and SNMF [8], etc. And its graph (co-)regularized extensions, such as GNMF [4], DNMF [9], and DRCC [10], have shown promising performance by exploiting intrinsic local manifold structure of data by manifold regularization [11].

THIS RESEARCH WAS SUPPORTED THE NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA (GRANT NO. 61271394), AND THE NATIONAL HEGAOJI KEY PROJECT (NO. 2013ZX01039-002-002).

Table 1. Comparison between Some Related Works

	LNMF	GNMF	IGNMF	NSDR	RNMFD
Feature Learning	√	√	√		√
Locality		√	√	√	√
Discriminability				√	√
Robustness	√				√
Trivial Solution	√		√	√	√

Despite the promising performance, NMF and its graph regularized extensions can be further improved because they still share three major flaws. First, most of existing NMF methods adopt squared loss to measure the data reconstruction quality. However, squared loss is sensitive to noises and outliers which are quite common in visual data, such that a few noisy entries may dominate the factorization because of their large reconstruction error. Second, the graph regularization may lead to trivial solution and scale transfer problem [12], resulting in meaningless image representation. Actually, when the weight of graph regularization is large, the learned representations of all entries tend to be quite similar. Third, existing methods mostly focus on preserving the locality of image data, but ignore the discriminative information which is also a quite important property for effective representation.

To address flaws above, we propose **R**obust **N**onnegative **M**atrix **F**actorization with **D**iscriminability (RNMFD) for image representation. Motivated by robust PCA [13], we impose a sparse error matrix to capture the noises in data for reconstruction. Thus the factorization is expected to capture more intrinsic latent information from the cleaned data, which may lead to better representation for images. On the other hand, we require the learned representation to capture the discriminative structure of data characterized by the scaled group indicator matrix [14]. It can be achieved by incorporating approximate orthogonal constraints into our objective function. In addition, our RNMFD can avoid trivial solution and scale transfer problems effectively because of the constraints. As a result, our RNMFD can perform feature learning, dimension reduction, locality preserving, and discriminative information exploiting simultaneously. It's also robust to data noise, and can avoid trivial solution and scale transfer problem caused by graph regularization. In fact, the properties mentioned above are all important but previous works ignore some of them. Comparison between RNMFD and some related works is shown in Table 1. We can observe that RNMFD can overcome the flaws in previous methods, leading to better result.

In summary, our contributions are as below. 1) We propose a novel NMF method for image representation, which has several important properties that previous works always ignore some of them. It can address the data noise and also avoid trivial solution problem resulting from graph regularization. It can also exploit the discriminative information of data. 2) We propose an effective optimization strategy for RNMFD and theoretically prove the convergence. This strategy is efficient for high-dimensional visual data and can converge very fast. 3) We conduct extensive experiment on five benchmark image datasets. The results demonstrate the superiority of RNMFD in comparison to state-of-the-art methods.

2. RELATED WORK

2.1. Preliminaries

Given images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$, where d is the dimension of feature and n is the number of images. NMF finds two nonnegative matrices $\mathbf{U} \in \mathbb{R}^{d\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ whose product can well approximate $\mathbf{X}$. And the squared loss is widely adopted for measuring the quality of approximation:

$$\mathcal{O}_{\text{NMF}} = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 \quad \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0 \tag{1}$$

Then $\mathbf{V}$ can be regarded as the new representation for images and used for tasks like clustering. We can adopt an iterative algorithm [15] with multiplicative rules to minimize Eq. (1),

$$u_{il} \leftarrow u_{il}\frac{(\mathbf{X}\mathbf{V})_{il}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{il}}, \quad v_{jl} \leftarrow v_{jl}\frac{(\mathbf{X}^T\mathbf{U})_{jl}}{(\mathbf{V}\mathbf{U}^T\mathbf{U})_{jl}} \tag{2}$$

Now define a p-nearest neighbor matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ as below,

$$W_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{N}(\mathbf{x}_i)$ denotes the p-nearest neighbor of $\mathbf{x}_i$. Then we can obtain a diagonal degree matrix with diagonal element $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Then Graph Regularized NMF (GNMF [4]) is formulated as,

$$\mathcal{O}_{\text{GNMF}} = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) \text{ s.t. } \mathbf{U}, \mathbf{V} \geq 0 \tag{4}$$

where α is the graph regularization parameter. Intuitively, it's not easy to choose proper value for α. Previous work has pointed out that if α is too large, the graph regularization will dominate the objective function with Eq. (4) reduced to below

$$\mathcal{O}'_{\text{GNMF}} = \sum_{i=1}^{k} \mathbf{v}_{*i}^T L \mathbf{v}_{*i} \quad \text{s.t.} \quad \mathbf{V} \geq 0 \tag{5}$$

Eq. (5) can be optimized by solving k independent subproblems: $\mathcal{O}_i = \mathbf{v}_{*i}^T L \mathbf{v}_{*i}$, where all subproblems have the same solutions up to a scale, i.e., $\mathbf{v}_{*1} \propto \mathbf{v}_{*2} \propto ... \propto \mathbf{v}_{*k}$, which is obviously meaningless for image representation. This is the trivial solution problem and some details can be found in [12].

Consider any solution $(\mathbf{U}^*, \mathbf{V}^*)$ to Eq. (4). For any $\rho > 1$, it's easy to verify that $(\rho\mathbf{U}^*, \frac{1}{\rho}\mathbf{V}^*)$ leads to smaller objective function value. Thus the final solution should be $\mathbf{U}^* \to \infty$ and $\mathbf{V}^* \to 0$. This is referred to as the scale transfer problem, which is also undesired for image representation.

2.2. Other Work

Some works have made effort to design robust and effective NMF. In LNMF [16], $\ell_{2,1}$ norm is used to measure the reconstruction error. The data noise and outliers are taken into consideration in RCC [17]. In DRCC [10], they apply ℓ_2 normalization on columns of $\mathbf{U}$ and $\mathbf{V}$ in each iteration of optimization. In IGNMF [12], a Normalized-Cut-like constraint is imposed to avoid trivial solution and scale transfer. In NLCF [5], they require basis to be close to original data points which will lead to sparse representation, which is motivated by Local Coordinate Coding [18]. In NSDR [19], the nonnegative spectral clustering is extended by discriminative regularization so discriminative information can be exploited.

According to their works, we can observe that all of the following properties are important for an effective NMF method, i.e., preserving locality, discriminability, robustness to data noise, avoiding trivial solution and scale transfer. Works mentioned above respectively focus on different perspectives leading to better performance than NMF while ignoring some others. Thus we propose a unified method with all properties above to learn better representation for images.

3. THE PROPOSED METHOD

3.1. Objective Function

We establish our method based on the objective function of GNMF, so our method can naturally preserve locality of data.

In real world, a data matrix is always the superposition of a low-rank component which captures the intrinsic information of data and a sparse component which is the noise in data [20]. In the objective function of NMF where squared loss is widely used, the noise data (sparse component) is always ignored but it indeed has important effect on the factorization. So we need to take this sparse component into consideration first. As suggested in [13], a given data matrix $\mathbf{M}$ can be decomposed as $\mathbf{M} = \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ is a low-rank matrix and $\mathbf{S}$ is a sparse matrix. We can notice that in NMF the data matrix $\mathbf{X}$ is approximated by a low-rank matrix $\mathbf{U}\mathbf{V}^T$ but it ignores the sparse component containing noises. Intuitively, we can impose a sparse error matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ to the objective function such that the low-rank reconstruction can be cleaner which can better capture the intrinsic information of data. Thus we can rewrite the objective function of GNMF as

$$\begin{aligned} \mathcal{O}_1 = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1 + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0 \end{aligned} \tag{6}$$

where $\|\mathbf{S}\|_1 = \sum_{ij}|S_{ij}|$ is the ℓ_1 norm which can guarantee the sparseness of $\mathbf{S}$ and λ is the regularization parameter to control the weight of sparse component. In the derivation later, we can see this sparse component can markedly alleviate the effect of elements with large error in factorization that is also noise which otherwise will dominate the factorization. With $\mathbf{S}$, our method shows robustness to noise in image data.

Furthermore, the learned representation should characterize the discriminative information of data. To address this issue, we follow the work in [14, 21] where scaled indicator matrix is introduced. First we can denote a group indicator matrix $\mathbf{Y} \in \{0,1\}^{n\times k}$, where $Y_{ij} = 1$ if the i-th image belongs to the j-th group. And the scaled indicator matrix can be defined as $\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, where every column of $\mathbf{F}$ is

$$\mathbf{F}_{*j} = [0, ..., 0, \underbrace{1, ..., 1}_{n_j}, 0, ...0]^T/\sqrt{n_j} \tag{7}$$

where n_j is the number of images in j-th group. So if the learned representation can capture this group information, it can be discriminative. Thus, it's reasonable to require the learned representation to be close to $\mathbf{F}$. Now by incorporating this idea into Eq. (6), we can obtain the objective function as,

$$\begin{aligned}\mathcal{O}_2 = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1 + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0, \|\mathbf{V} - \mathbf{F}\|_F^2 \leq \epsilon\end{aligned} \tag{8}$$

However, since there is no supervision information available, we indeed have no knowledge about $\mathbf{F}$. But fortunately, we could observe from the definition that $\mathbf{F}$ is strictly orthogonal:

$$\mathbf{F}^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I} \tag{9}$$

where $\mathbf{I}$ is a $k \times k$ identity matrix. Furthermore, if ϵ is small enough, $\mathbf{V}$ is very close to $\mathbf{F}$. Because of the orthogonality of $\mathbf{F}$, $\mathbf{V}$ should be *approximatively* orthogonal. Under the ultimate situation where $\epsilon = 0$, $\mathbf{V}$ is exactly orthogonal because it's equal to $\mathbf{F}$. Thus we can substitute the constraints containing $\mathbf{F}$ with *approximate orthogonal constraints* below,

$$\begin{aligned}\mathcal{O}_3 = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1 + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0, \|\mathbf{V}^T\mathbf{V} - \mathbf{I}\|_F^2 \leq \epsilon\end{aligned} \tag{10}$$

For the convenience for optimization, we rewrite Eq. (10) as below. Then we can obtain the objective function of RNMFD,

$$\begin{aligned}\mathcal{O} = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1 + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ + \beta\|\mathbf{V}^T\mathbf{V} - \mathbf{I}\|_F^2 \quad \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0\end{aligned} \tag{11}$$

where parameter β controls the orthogonality of $\mathbf{V}$ and a properly large β can make $\|\mathbf{V}^T\mathbf{V} - \mathbf{I}\|_F^2 \leq \epsilon$ in Eq. (10) satisfied for any ϵ. By incorporating the approximate orthogonal constraints into the objective function, $\mathbf{V}$ learned from Eq. (11) can characterize the discriminative structure of data.

Besides leading to discriminative representation, the approximate orthogonal constraints can also address the trivial solution and scale transfer problems. As the columns of $\mathbf{V}$ are approximatively orthogonal and $\mathbf{V}$ is nonnegative, just few elements in each row may have significantly large value and each column should be as different as possible to each other. Hence solution is obviously nontrivial. Also, by substituting any solution $\mathbf{V}^*$ with $\frac{1}{\rho}\mathbf{V}^*$ ($\rho > 1$), the forth term will have larger value which may lead to larger objective function value if we set β large enough (e.g., $\beta > 10^2$). Consequently, the scale transfer problem can be addressed at the same time, too.

3.2. Optimization Algorithm

The objective function $\mathcal{O}$ in Eq. (11) isn't convex in $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{S}$ together. Fortunately, we can minimize $\mathcal{O}$ iteratively by updating one while fixing the others. Denote $\hat{\mathbf{X}} = \mathbf{X} - \mathbf{S}$.

When $\mathbf{V}$ and $\mathbf{S}$ are fixed, the objective function w.r.t. $\mathbf{U}$ is equivalent to original NMF. Thus the updating rule for $\mathbf{U}$ is

$$u_{il} \leftarrow u_{il}\frac{(\hat{\mathbf{X}}\mathbf{V})_{il}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{il}} \tag{12}$$

Now we need to derive the updating rule for $\mathbf{V}$ by fixing $\mathbf{U}$ and $\mathbf{S}$. Noticing that $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}\mathbf{A}^T)$ and $tr(\mathbf{A}\mathbf{B}) = tr(\mathbf{B}\mathbf{A})$, so we can rewrite the objective function as follows,

$$\begin{aligned}\mathcal{O} = tr(\hat{\mathbf{X}}\hat{\mathbf{X}}^T - 2\hat{\mathbf{X}}\mathbf{V}\mathbf{U}^T + \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T) + \lambda\|\mathbf{S}\|_1 \\ + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) + \beta tr(\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V} - 2\mathbf{V}^T\mathbf{V} + \mathbf{I})\end{aligned} \tag{13}$$

Let ϕ_{jl} be the Lagrange multiplier for constraint $v_{jl} \geq 0$, and denote $\mathbf{\Phi} = [\phi_{jl}]$. The Lagrange $\mathcal{L}$ can be written as follows,

$$\mathcal{L} = \mathcal{O} + tr(\mathbf{\Phi}\mathbf{V}^T) \tag{14}$$

We can compute the partial derivative of $\mathcal{L}$ w.r.t. $\mathbf{V}$ as below,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{V}} = -2\hat{\mathbf{X}}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} + 2\alpha\mathbf{L}\mathbf{V} \\ + 4\beta\mathbf{V}\mathbf{V}^T\mathbf{V} - 4\beta\mathbf{V} + \mathbf{\Phi}\end{aligned} \tag{15}$$

Now by using the Karush-Kuhn-Tucker (KKT) conditions, that is, $\phi_{jl}v_{jl} = 0$, we obtain the following equation for v_{jl},

$$\begin{aligned}-(\hat{\mathbf{X}}^T\mathbf{U})_{jl}v_{jl} + (\mathbf{V}\mathbf{U}^T\mathbf{U})_{jl}v_{jl} + \alpha(\mathbf{L}\mathbf{V})_{jl}v_{jl} \\ + 2\beta(\mathbf{V}\mathbf{V}^T\mathbf{V})_{jl}v_{jl} - 2\beta(\mathbf{V})_{jl}v_{jl} = 0\end{aligned} \tag{16}$$

Then Eq. (16) results in the updating rule for $\mathbf{V}$ as follows,

$$v_{jl} \leftarrow v_{jl}\frac{(\hat{\mathbf{X}}^T\mathbf{U} + \alpha\mathbf{W}\mathbf{V} + 2\beta\mathbf{V})_{jl}}{(\mathbf{V}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^T\mathbf{V})_{jl}} \tag{17}$$

Now we can update $\mathbf{S}$ by fixing $\mathbf{U}$ and $\mathbf{V}$. Actually, the optimizing problem with respect to $\mathbf{S}$ is element-wise decoupled and the unique solution to each subproblem has a closed form with soft-thresholding operator. Now denote $\mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{V}^T = [e_{ij}]$. The subproblem for s_{ij} is as below,

$$\mathcal{O}_{ij} = (e_{ij} - s_{ij})^2 + \lambda|s_{ij}| \tag{18}$$

It's not difficult to derive the solution for Eq. (18) as follows,

$$s_{ij} = \begin{cases} e_{ij} - \frac{\lambda}{2}\text{sign}(e_{ij}), & \text{if } |e_{ij}| > \frac{\lambda}{2} \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

We can see that for elements with large reconstruction error which is always caused by noise, the influence of large error can be alleviated by $\mathbf{S}$. Otherwise, they will dominate the factorization and lead to unsatisfactory result. And for elements with small error which is common, $\mathbf{S}$ has no effect on them. In fact, by substituting Eq. (19) into Eq. (18), we may obtain

$$\mathcal{O}_{ij} = \begin{cases} e_{ij}^2, & \text{if } |e_{ij}| \leq \frac{\lambda}{2} \\ \lambda|e_{ij}| - (\frac{\lambda}{2})^2, & \text{otherwise} \end{cases} \tag{20}$$

The result reveal that RNMFD can self-adaptively apply ℓ_2 loss to small-error entries for accurate reconstruction, and ℓ_1 loss to large-error entries to alleviate the influence from noise.

By applying Eq. (12), Eq. (17) and Eq. (19) iteratively, the objective function can converge to a local minima, which is theoretically guaranteed by Theorem 1 introduced as below.

3.3. Proof of Convergence

Theorem 1 *The function value in Eq. (11) is nonincreasing under the updating rules in Eq. (12), Eq. (17) and Eq. (19).*

The updating rule for $\mathbf{U}$ is the same as in the original NMF. Thus $\mathcal{O}$ in Eq. (11) is nonincreasing under Eq. (12), whose proof can be found in [15]. Furthermore, the derivation of updating rule for $\mathbf{S}$ is also quite simple and $\mathcal{O}$ is obviously nonincreasing under Eq. (19). Thus we just need to prove that $\mathcal{O}$ is nonincreasing under updating rule for $\mathbf{V}$ in Eq. (17). The proof uses the auxiliary function [22] defined as follows.

Definition 1 *$G(v, v')$ is an auxiliary function for $F(v)$ if*

$$G(v, v') \geq F(v), \quad G(v, v) = F(v)$$

is satisfied.

Lemma 1 *If G is an auxiliary function of F, then F is nonincreasing under the updating rule*

$$v^{(t+1)} = \arg\min_{v} G(v, v^{(t)}) \tag{21}$$

Proof 1 (for Lemma 1)

$$F(v^{(t+1)}) \leq G(v^{(t+1)}, v^{(t)}) \leq G(v^{(t)}, v^{(t)}) = F(v^{(t)})$$

We just need to show the updating rule for $\mathbf{V}$ in Eq. (17) is exactly the rule in Eq. (21) with a proper auxiliary function. Denote F_{ab} as the part of $\mathcal{O}$ which is only relevant to v_{ab}. One auxiliary function for F_{ab} can be defined in the lemma below

Lemma 2 *The function*

$$\begin{aligned} G(v, v_{ab}^{(t)}) &= F_{ab}(v_{ab}^{(t)}) + F'_{ab}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{(\mathbf{V}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^T\mathbf{V})_{ab}}{v_{ab}^{(t)}}(v - v_{ab}^{(t)})^2 \end{aligned} \tag{22}$$

is an auxiliary function for F_{ab}.

Table 2. Description of benchmark datasets

Dataset	#Example	#Features	#Classes
ORL	400	1024	40
YALE	165	1024	15
UMIST	398	644	20
MNIST	1000	784	10
Semeion	1593	256	10

Proof 2 (for Lemma 2) It's obvious that $G(v, v) = F_{ab}(v)$. By comparing $G(v, v_{ab}^t)$ to Taylor expansion of $F_{ab}(v)$, we have $G(v, v_{ab}^t) \geq F_{ab}(v)$. Similar proof can be found in [4].

Proof 3 (for Theorem 1) By substituting $G(v, v_{ab}^{(t)})$ in Eq. (21) with Eq. (22), we can obtain the updating rule as below,

$$\begin{aligned} v_{ab}^{(t+1)} &= v_{ab}^{(t)} - v_{ab}^{(t)} \frac{F'_{ab}(v_{ab}^{(t)})}{2(\mathbf{V}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^T\mathbf{V})_{ab}} \\ &= v_{ab}^{(t)} \frac{(\hat{\mathbf{X}}^T\mathbf{U} + \alpha\mathbf{W}\mathbf{V} + 2\beta\mathbf{V})_{ab}}{(\mathbf{V}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^T\mathbf{V})_{ab}} \end{aligned} \tag{23}$$

which is identical to Eq. (17). Because $G(v, v_{ab})$ is the auxiliary function of F_{ab}, F_{ab} is nonincreasing under this updating rule. Therefore $\mathcal{O}$ in Eq. (11) is nonincreasing under Eq. (17).

3.4. Computational Complexity

The time complexity for constructing nearest neighbor graph is $\mathcal{O}(n^2d)$. And the updating complexity in each iteration is $\mathcal{O}(ndk+(n+d)k^2+npk+n^2(k+1)+nk(p+1)+n^2(k+1)+n(k+1)+(d+n)k)$. Thus the overall complexity of RNMFD is $\mathcal{O}(tndk + n^2d)$ when $d > n$, where t is the number of iterations. It's linear to the number of images and dimension when high-dimensional visual data is given, which is the same as GNMF. Thus it is quite efficient for high-dimensional data.

4. EXPERIMENT AND DISCUSSION

4.1. Datasets, Metrics and Details

To validate the effectiveness of RNMFD for image representation, we conduct image clustering as in [4, 5] on five benchmark image datasets, ORL, YALE, UMIST, MNIST and Semeion. Table 2 summarizes the statistics of these five datasets.

We adopt two standard metrics which are widely used for clustering as the evaluation metrics, i.e., Clustering Accuracy and Normalized Mutual Information, whose definition can be found in [12]. Actually, since the clustering task and the corresponding metrics are widely used to evaluate the effectiveness of image representation, therefore we follow the settings.

The following representation learning and clustering methods are compared to RNMFD. Kmeans is chosen as the base method. The representation learning methods include NMF, LNMF [16], GNMF [4], DNMF [9], NLCF [5], and IGNMF [12]. These methods can learn a new representation for images. The clustering methods are RCC [17] and NSDR [19]. These methods directly perform clustering on original image features. Actually, our RNMFD belongs to the former.

Table 3. Clustering Accuracy (%)

Dataset	Kmeans	NMF	LNMF	GNMF	DNMF	NLCF	RCC	IGNMF	NSDR	RNMFD
ORL	41.00	52.75	52.50	54.50	54.50	53.75	57.00	56.00	57.75	**65.25**
YALE	32.73	35.15	36.15	39.39	39.84	41.21	42.32	40.57	39.39	**46.67**
UMIST	46.48	46.23	47.34	56.28	52.01	53.37	60.11	58.84	64.08	**70.10**
MNIST	47.50	47.90	47.10	50.70	49.90	48.70	52.40	52.10	55.80	**59.20**
Semeion	55.56	45.49	46.42	58.43	60.14	56.41	62.37	60.11	63.34	**66.79**
Average	44.65	45.50	45.90	51.86	51.29	50.69	54.85	53.52	56.07	**61.60**

Table 4. Normalized Mutual Information (%)

Dataset	Kmeans	NMF	LNMF	GNMF	DNMF	NLCF	RCC	IGNMF	NSDR	RNMFD
ORL	67.01	74.76	73.85	75.81	75.47	74.90	75.81	75.93	75.78	**82.35**
YALE	40.32	44.97	44.84	46.37	45.02	48.80	49.31	47.01	48.29	**53.01**
UMIST	63.81	64.74	65.16	75.85	71.64	72.13	76.43	76.31	76.03	**80.07**
MNIST	47.16	44.93	44.84	48.19	47.90	50.09	57.20	52.11	58.11	**62.11**
Semeion	50.94	41.04	42.27	55.88	57.33	54.21	59.94	56.72	61.90	**63.54**
Average	53.85	54.09	54.19	60.42	59.47	60.03	63.74	61.61	64.02	**68.22**

For meaningful comparison, we perform grid search in parameter spaces and the *best results* of baselines are reported. When constructing p-NN graph for methods like GNMF, p is chosen from $\{1, 2, ..., 9\}$. For graph (co-)regularization, its weight parameter is chosen from $\{0.01, 0.05, ..., 1000\}$. And for NSDR and NLCF, their respectively corresponding model parameters are selected from $\{10^{-2}, 10^{-1}, 1, ..., 10^{6}\}$.

RNMFD also has some parameters, i.e., λ, p, α and β. We can set λ empirically according to the properties of the specific dataset. Therefore λ can be adaptive to the dataset. When comparing RNMFD to the baselines, we set $p = 5$, $\alpha = 100$ and $\beta = 100$ consistently for RNMFD in all experiments. In the coming section, we conduct empirical analysis on parameter sensitivity, which verifies that RNMFD can achieve superior and stable performance under a wide range of parameter values. And for all methods in our experiment, we set k to be the number of true classes following previous works [12, 17].

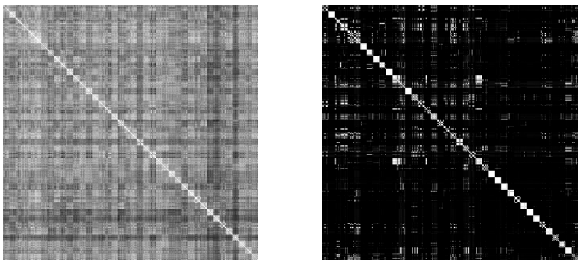
To remove the influence of random initialization, all the results reported are the average values over 10 repeated runs.

4.2. Results and Discussion

We first compare the image clustering performance of all methods. The clustering results are shown in Table 3 and Table 4. We can observe that RNMFD can significantly outperform all baseline methods regardless of the datasets. In addition, the experiment results reveal some important points.

First, methods considering the locality, like GNMF and RNMFD, achieves better performance than NMF, highlighting the importance of preserving locality, as suggested in [4].

Second, by explicitly considering the unexpected noise in data, RCC and RNMFD can outperform GNMF and DNMF who are affected by the noise, which implies that the real-world dataset may contain noises to some extent. Removing the noises from data can indeed result in better performance.



(a) Similarity of GNMF (b) Similarity of RNMFD

Fig. 1. Comparison Between GNMF and RNMFD

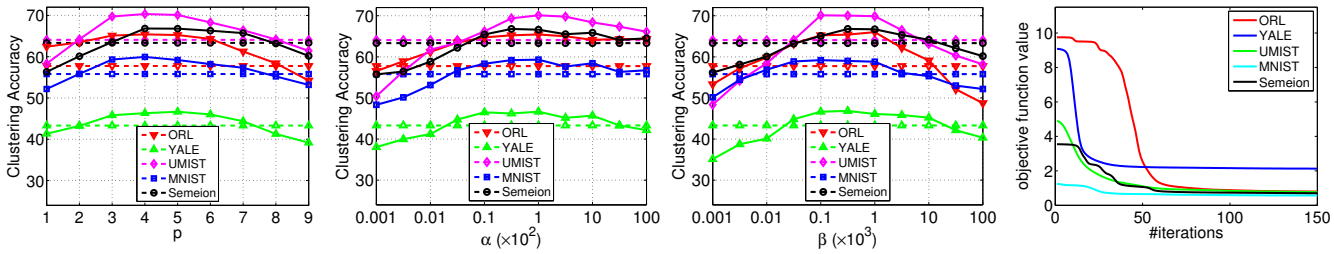
Third, compared to GNMF and DNMF, IGNMF and RNMFD can avoid trivial solution and scale transfer, which leads to better result. And we noticed that IGNMF and RNMFD can achieve stable performance under a wide range of value for α while GNMF shows unstable performance when changing α.

Fourth, among all methods, NSDR and RNMFD achieve best performance. In fact, they are the only two methods exploiting the discriminative information. This phenomenon indicates that besides the locality of data, the discriminative information is also essential for effective image representation.

In a summary, the experiments validates our observation mentioned before, i.e., all the following properties, locality preserving, discriminability, robustness to noise and the ability to avoid trivial solution, are important for image representation. But previous works on NMF ignore some of them. RNMFD satisfies all properties hence achieves best performance.

Furthermore, we compare the similarity between samples represented by new features learned by GNMF and RNMFD on ORL dataset, which is shown in Figure 1. Brighter color indicates more similarity. We can see that the inter-class similarity of GNMF is very large. But for RNMFD, intra-class similarity is large while inter-class similarity is very small, which is a desired property for effective image representation.

We further conduct empirical analysis on parameter sen-



(a) # nearest neighbor p (b) graph regularization α (c) orthogonality regularization β (d) convergence

Fig. 2. Parameter Sensitivity and Convergence Analysis

sitivity and the results are shown in Figure 2(a) to Figure 2(c). The dashed lines are the best result from all baseline methods.

In Figure 2(a), we show the performance of RNMFD w.r.t. p, the number of nearest neighbors. If p is too small, the local structure can't be fully utilized. On the contrary, if p is too large, the nearest neighbor graph may connect two dissimilar samples. We can observe that our RNMFD can outperform the best baseline method on all five datasets when $p \in [3, 7]$.

The effect of α on RNMFD is shown in Figure 2(b). A small α may cause a weak regularization such that the locality of data can't be effectively preserved while a too large α may lead to trivial solution for graph-regularized methods like GNMF. However, because we have approximate orthogonal constraints which can make RNMFD avoid trivial solution, RNMFD can achieve stable performance even with quite large value, e.g., $\alpha = 5000$, which is similar to IGNMF. RNMFD shows superior and stable performance when $\alpha \in [10, 10^4]$ that is much wider than methods such as GNMF and DNMF.

We plot the performance of RNMFD with respect to β in Figure 2(c). Theoretically, when β is too small, RNMFD may be ill-defined and prone to trivial solution and the discriminative information can't be exploited. On the other hand, when β is too large, the orthogonal constraints can be so heavy that the learned representation may be too sparse, which is also undesired in real world. We can observe that our RNMFD is able to achieve satisfactory performance when $\beta \in [50, 10^4]$.

We have proven the proposed multiplicative updating rules are convergent, now we investigate how fast they can converge. We plot the objective function value (averaged by the number of samples) with respect to iterations on all datasets in Figure 2(d). It is observe that the updating rules for RNMFD converge very fast, usually within 100 iterations.

5. CONCLUSION

In this paper, we propose a novel RNMFD as an extension of NMF, for learning image representation. RNMFD can simultaneously perform feature learning, dimension reduction, locality preserving and discriminative information exploiting. And it's also robust to data noise and can avoid trivial solution and scale transfer problem caused by graph regularization. We propose an iterative strategy for RNMFD and prove the convergence. We conduct extensive experiments on five image datasets. Results validate the effectiveness of RNMFD.

6. REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," in *Wiley-Interscience, Hoboken, NJ, 2nd edition*, 2000. 1

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," in *Nature*, 1999. 1

[3] 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. 1

[4] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *TPAMI*, 2011. 1, 2, 4, 5

[5] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," *TIP*, 2013. 1, 2, 4

[6] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *JMLR*, vol. 5, pp. 1457 – 1469, 2004. 1

[7] C. H. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *ACM SIGKDD*, 2006. 1

[8] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE TPAMI*, vol. 99, no. 1, 2008. 1

[9] F. Shang, L. C. Jiao, and F. Wang, "Graph dual regularization nonnegative matrix factorization for co-clustering," *PR*, 2012. 1, 4

[10] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proc. of ACM SIGKDD*, 2009. 1, 2

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *JMLR*, vol. 7, pp. 2399–2434, 2006. 1

[12] Q. Gu, C. Ding, and J. Han, "On trivial solution and scale transfer problems in graph regularized nmf," in *IJCAI*, 2011. 1, 2, 4, 5

[13] E. Cands, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *JACM*, 2011. 1, 2

[14] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, , and Yueting Zhuang, "Image clustering using local discriminant models and global integration," *IEEE TIP*. 1, 3

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," in *NIPS*, 2001. 2, 4

[16] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proc. of CIKM*, 2011. 2, 4

[17] L. Du and Y. Shen, "Towards robust co-clustering," in *Proc. of 23nd IJCAI*, 2013. 2, 4, 5

[18] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *NIPS*, 2009. 2

[19] Y. Yang, H. Shen, R. Ji F. Nie, and X. Zhou, "Nonnegative spectral clustering with discriminative regularization," in *AAAI*, 2011. 2, 4

[20] Guangyu Zhu, Shuicheng Yan, and Yi Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *ACM MM*, 2010. 2

[21] J. Ye, Z. Zhao, and M. Wu, "Discriminative k-means for clustering," in *Advances in Neural Information Processing Systems*, 2007. 3

[22] A. P. Dempster, N. M. Laird, , and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal RSSS*, 1977. 4