Robust and Discriminative Concept Factorization for Image Representation

Yuchen Guo
Tsinghua Univeristy
Beijing, China
yuchen.w.guo@gmail.com

Guiguang Ding
Tsinghua Univeristy
Beijing, China
dinggg@tsinghua.edu.cn

Jile Zhou
Sohu Inc.
Beijing, China
jile.zip@gmail.com

Qiang Liu
Tsinghua Univeristy
Beijing, China
liuqiang@tsinghua.edu.cn

ABSTRACT

Concept Factorization (CF), as a variant of Nonnegative Matrix Factorization (NMF), has been widely used for learning compact representation for images because of its psychological and physiological interpretation of naturally occurring data. And graph regularization has been incorporated into the objective function of CF to exploit the intrinsic low-dimensional manifold structure, leading to better performance. But some shortcomings are shared by existing CF methods. 1) The squared loss used to measure the data reconstruction quality is sensitive to noise in image data. 2) The graph regularization may lead to trivial solution and scale transfer problems for CF such that the learned representation is meaningless. 3) Existing methods mostly ignore the discriminative information in image data. In this paper, we propose a novel method, called **R**obust and **D**iscriminative **C**oncept **F**actorization (RDCF) for image representation. Specifically, RDCF explicitly considers the influence of noise by imposing a sparse error matrix, and exploits the discriminative information by approximate orthogonal constraints which can also lead to nontrivial solution. We propose an iterative multiplicative updating rule for the optimization of RDCF and prove the convergence. Experiments on 5 benchmark image datasets show that RDCF can significantly outperform several state-of-the-art related methods, which validates the effectiveness of RDCF.

Keywords

Concept Factorization, Graph Regularization, Discriminability, Noise, Image Representation

1. INTRODUCTION

Because of the *curse of dimensionality*, applying statistic techniques to image data which is always represented by high-dimensional vector becomes infeasible [8]. Thus learning low-dimensional compact representation lays the fundamental for many real-world applications, like pattern recognition, computer vision and image processing, etc [7, 9, 12, 13, 21]. Among different methods, Nonnegative Matrix Factorization (NMF) [16, 17] which aims to find two *nonnegative* matrices as the basis and the corresponding coefficients whose product can well approximate the original data, has attracted considerable attention because it can learn *parts-based* representation for images which has psychological and physiological interpretation of naturally occurring data [19, 22]. And as a variant of NMF, Concept Factorization (CF) [23], which aims to represent basis by the linear combination of original data, has also shown promising performance for data representation. In addition, previous works on *manifold learning* [1] have been incorporated into CF as the *graph regularization* to exploit the intrinsic low-dimensional manifold structure embedding the high-dimensional data, which leads to better visual analysis performance. One representative and effective work is Locality Consistent Concept Factorization (LCCF) [2], which incorporates graph regularization to the objective function of conventional CF to learn locality consistent data feature.

In spite of its state-of-the-art performance, LCCF can be further improved because it still suffers from the following three shortcomings. First, because of the squared loss adopted in LCCF to measure the quality of data reconstruction, it's very unstable and sensitive to noise in data. Thus the factorization may be dominated by the noise which can degrade the quality of learned representation. Second, the graph regularization may lead to trivial solution and scale transfer problems because it's actually not well-defined and the objective function is not lower-bounded [10]. These problems may result in meaningless representation for image data. Thirdly, though it fully considers the local information of data, it ignores the discriminative information of original data, which is also important for visual analysis tasks like clustering. Furthermore, some recent works claim that guaranteing sparseness of representation for NMF and CF can produce much better performance [5, 14, 18]. But LCCF always results in dense representation for image data.

To address theses issues, in this paper we propose a novel method, referred to as Robust and Discriminative Concept Factorization (RDCF) for image representation. We es-

ICMR '15 June 23 - 26, 2015, Shanghai, China
Copyright © 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00.
http://dx.doi.org/10.1145/2671188.2749317.

Table 1: Comparison between Some Related Works

	CF	LNMF	LCF	LCCF	RDCF
Feature Learning	√	√	√	√	√
Locality			√	√	√
Discriminability					√
Robustness		√			√
Trivial Solution	√	√	√		√
Sparseness	?		√		√

tablish our method based on LCCF such that RDCF has the advantages of CF and can preserve locality naturally. We utilize a *sparse error matrix* to explicitly capture noise which has significant influence upon the data reconstruction. Thus the factorization can capture more intrinsic information from the cleaned data. And we propose to add *approximate orthogonal constraints* to the objective function. With the constraints, RDCF can 1) exploit the discriminative information of data which leads to better representation; 2) avoid trivial solution and scale transfer problems even with strong graph regularization; 3) learn relatively sparse representation for image. Our RDCF is strongly related to but different from several previous works. In table 1, we compare RDCF to several related methods, CF[23] (NMF [16]), LNMF [15], LCCF [2] (GNMF [3]), LCF [18] (NLCF [5]), and NSDR [24]. RDCF can simultaneously perform feature learning, dimension reduction, locality preserving, and discriminative information exploiting. And It's robust to data noise and can avoid trivial solution and scale transfer problems even with graph regularization. The properties mentioned above are all important to achieve superior performance, but previous works always ignores some of them. Furthermore, RCDF can result in sparse representation, which is also very meaningful for image representation. In summary, this paper makes contributions as below,

- We propose a novel method RDCF for image representation. RDCF is robust to data noise, can avoid trivial solution and scale transfer problem and exploit the discriminative information of data, which are the major problems LCCF suffers from. Actually, RDCF satisfies several important properties for image feature learning while previous methods ignore some of them.
- We propose an effective and efficient iterative strategy with multiplicative updating rules for the optimization for RDCF, and we theoretically prove the convergence.
- We carried out extensive experiments on five public image datasets. The experimental results show that RDCF can significantly outperform several state-of-the-art methods, validating the effectiveness of RDCF.

The rest of this paper is organized as follows. Some related works are briefly reviewed in Section 2. We will introduce the proposed RDCF in detail in Section 3 and the corresponding theoretical analysis is given in Section 4. The experimental results on benchmark datasets are presented in Section 5. In the end, conclusions are made in Section 6.

2. RELATED WORK

2.1 Preliminaries

Given a set of nonnegative image data represented as $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where n is the number of samples and d is the number of feature dimension. CF aims to find two nonnegative matrices $\mathbf{W} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ where k is the dimension of the new representation, such that the original data can be well approximate. The objective function of the conventional CF can be written as

Table 2: Notations and descriptions in this paper

Notation	Description	Notation	Description
$\mathbf{X}$	input data matrix	n	#images
$\mathbf{W}$	basis coefficients	d	#dimension
$\mathbf{V}$	new representation	k	#basis vectors
$\mathbf{L}$	graph Lap. matrix	p	#NN
$\mathbf{S}$	sparse error matrix	λ	sparse para.
$\mathbf{K}, \hat{\mathbf{K}}$	kernel matrix	α	graph para.
$\mathbf{F}$	scaled indi. matrix	β	orth. para.

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

where $\|\cdot\|_F$ is the *Frobenius norm* of matrix. Generally, Eq. (1) can be optimized by an iterative strategy with multiplicative updating rules for $\mathbf{W}$ and $\mathbf{V}$ as suggested in [23],

$$w_{jl} \leftarrow w_{jl}\frac{(\mathbf{KV})_{jl}}{(\mathbf{KWV}^T\mathbf{V})_{jl}}, v_{jl} \leftarrow v_{jl}\frac{(\mathbf{KW})_{jl}}{(\mathbf{VW}^T\mathbf{KW})_{jl}} \tag{2}$$

where $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ is the kernel matrix. We can first define a p-nearest neighbor matrix $\mathbf{G}$ whose elements are as follows

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

where $\mathcal{N}(\mathbf{x}_i)$ denotes the p-nearest neighbor of $\mathbf{x}_i$. Now we can further define the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{G}$, where $\mathbf{D}$ is the diagonal degree matrix whose diagonal element $D_{ii} = \sum_{j=1}^{n} G_{ij}$. By incorporating this graph regularization, we can obtain the objective function of LCCF as below

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

where α is the graph regularization parameter. The local information can't be fully exploited with a small value for α. However, if α is too large, the graph regularization may dominate the objective function and Eq. (4) reduces to

$$\mathcal{O}_{\text{LCCF}}^{'} = tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) = \sum_{i=1}^{k}\mathbf{v}_{*i}^T\mathbf{L}\mathbf{v}_{*i} \tag{5}$$

where $\mathbf{v}_{*i}$ is the i-th column of $\mathbf{V}$. Actually Eq. (5) can be optimized as k independent subproblems $\mathcal{O}_i = \mathbf{v}_{*i}^T\mathbf{L}\mathbf{v}_{*i}$, leading to the same solutions up to a scale, i.e., $\mathbf{v}_{*1} \propto \mathbf{v}_{*2} \propto ... \propto \mathbf{v}_{*k}$. In fact, the learned representations are meaningless, e.g., the cosine similarity between any images is 1. And this is the so-called *trivial solution* problem in [10].

Furthermore, because the graph regularization isn't well-defined and lower-bounded, suppose we obtain any solution $(\mathbf{W}^*, \mathbf{V}^*)$ to Eq. (4). Given $\forall\gamma > 1$, it's easy to verify that $(\gamma\mathbf{W}^*, \frac{1}{\gamma}\mathbf{V}^*)$ can lead to smaller objective function value. Consequently the ultimate solution will be $\mathbf{W}^* \to \infty$ and $\mathbf{V}^* \to 0$, which is referred to as the *scale transfer* problem.

2.2 Other Work

Recent years, some works have made some effort to address some problems mentioned in Section 1. Though they are actually for NMF, we find they can be applied to CF too. In LNMF [15], $\ell_{2,1}$ norm is utilized to measure the quality of reconstruction, which is more robust than squared loss.

In DRCC [11], they use ℓ_2 normalization in each optimization iteration. In NSDR [24], discriminative information is exploited in the clustering. Here we need to point out that NSDR is a spectral clustering method, but not a feature learning method as others. Hence its applications are limited. In LCF [18], they require the basis to be close to original data points such that $\mathbf{V}$ can be sparse, which is motivated by Local Coordinate Coding [27]. Though the trivial solution and scale transfer problems are avoided in LCF and the locality is preserved, LCF still suffers from the other two main shortcomings of LCCF such thut its best performance is just comparable to LCCF's. In summary, these methods focus on some perspectives for designing effective NMF or CF, but they ignore some others. Motivated by them, we propose a unified method taking all the perspectives into consideration, which will lead to much better performance.

3. THE PROPOSED METHOD

3.1 Objective Function

In spite of the trivial solution and scale transfer problem, graph regularization is still a powerful and effective tool to preserve the locality of image data. Thus we establish our RDCF upon LCCF. Actually, in RDCF, the trivial solution and the scale transfer problems can be effectively addressed.

The first step is to address the problem that squared loss widely utilized in CF is sensitive to noise in data. Two alternatives are available. Using other loss function instead of squared loss as in LNMF, or remove noise from data. In this paper, we choose the latter. Motivated by Robust PCA [4], a data matrix can be $\mathbf{M}$ can be decomposed as the sum of a low-rank component $\mathbf{L}$ and a sparse component $\mathbf{S}$, i.e., $\mathbf{M} = \mathbf{L} + \mathbf{S}$. Actually, we can observe that the term $\mathbf{XWV}^T$ is essentially low-rank. But in CF and LCCF, the sparse component that is always noise is ignored but it indeed has considerable effect. In fact, the sparse component, i.e., noise, often dominates the factorization leading to unsatisfactory representation for data. Thus, we can impose a sparse component into the factorization to capture the noise such that the factorization can capture more intrinsic information from the cleaned data. By incorporating this idea into LCCF, we can obtain the objective function as follows,

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

where $\|\mathbf{S}\|_1 = \sum_{ij} |S_{ij}|$ is the ℓ_1 norm of matrix, which can guarantee the sparseness of the matrix. By imposing this sparse error matrix, the influence of noise is removed and a cleaned data matrix $\mathbf{X} - \mathbf{S}$ can be constructed thus the reconstruction can capture more intrinsic information. Consequently our model is robust to the noise in image data.

Then we need to make the learned representation $\mathbf{V}$ discriminative, i.e., capture some discriminative information of data. Here we follow the works in [25] and [26]. First we introduce a group indicator matrix $\mathbf{Y} = \{0,1\}^{n\times k}$ where $Y_{ij} = 1$ if the i-th image belongs to the j-th group. The scaled indicator matrix with respect to $\mathbf{Y}$ is defined as below

$$\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} \tag{7}$$

where each column in $\mathbf{F}$, i.e., each image group, is given by

$$\mathbf{F}_{*j} = [0, ..., 0, \underbrace{1, ..., 1}_{n_j}, 0, ...0]^T / \sqrt{n_j} \tag{8}$$

where n_j is the number of samples in the j-th group. We want the learned representation $\mathbf{V}$ to characterize the discriminative structure in $\mathbf{F}$. Intuitively, we just need to force them to be close to each other, i.e., $\|\mathbf{V} - \mathbf{F}\|_F^2 \leq \epsilon$, where ϵ is a small constant. Now we can incorporate this constraint on $\mathbf{V}$ to the objective function defined in Eq. (6) as follows

$$\begin{aligned}\mathcal{O}_2 = &\|\mathbf{X} - \mathbf{S} - \mathbf{XWV}^T\|_F^2 + \lambda\|\mathbf{S}\|_1 \\ &+ \alpha tr(\mathbf{V}^T\mathbf{LV}) \text{ s.t. } \mathbf{W}, \mathbf{V} \geq 0, \ \|\mathbf{V} - \mathbf{F}\|_F^2 \leq \epsilon\end{aligned} \tag{9}$$

Unfortunately, it's impossible to know $\mathbf{F}$ in prior because we have no label information under unsupervised scenario. But in Eq. (7), we can observe that $\mathbf{F}$ is strictly orthogonal

$$\mathbf{F}^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}_k \tag{10}$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix. Look back to Eq. (9), if we set $\epsilon = 0$, then $\mathbf{V}$ is strictly equal to $\mathbf{F}$. Consequently $\mathbf{V}$ should be orthogonal too. However, this constraints may be too heave. Actually, given a small ϵ, the learned $\mathbf{V}$ is close but not totally equal to $\mathbf{F}$. Thus $\mathbf{V}$ isn't strictly but approximately orthogonal, which can be formulated as $\|\mathbf{V}^T\mathbf{V} - \mathbf{I}_k\|_F^2 \leq \epsilon$. By substituting the term containing $\mathbf{F}$ in Eq. (9) with the approximate orthogonal constraints as above, now we can obtain the objective function as follows,

$$\begin{aligned}\mathcal{O}_3 = &\|\mathbf{X} - \mathbf{S} - \mathbf{XWV}^T\|_F^2 + \alpha tr(\mathbf{V}^T\mathbf{LV}) \\ &+ \lambda\|\mathbf{S}\|_1 \quad \text{s.t. } \mathbf{W}, \mathbf{V} \geq 0, \ \|\mathbf{V}^T\mathbf{V} - \mathbf{I}_k\|_F^2 \leq \epsilon\end{aligned} \tag{11}$$

Further we can rewrite Eq. (11) for optimization issues, and obtain the overall objective function of our RDCF as below,

$$\begin{aligned}\mathcal{O} =& \|\mathbf{X} - \mathbf{S} - \mathbf{XWV}^T\|_F^2 + \alpha tr(\mathbf{V}^T\mathbf{LV}) \\ &+ \lambda\|\mathbf{S}\|_1 + \beta\|\mathbf{V}^T\mathbf{V} - \mathbf{I}_k\|_F^2 \quad \text{s.t. } \mathbf{W}, \mathbf{V} \geq 0\end{aligned} \tag{12}$$

where β is the regularization parameter for *approximate orthogonal constraints*. In fact, given any ϵ, a large enough β can make RDCF satisfy the constraint $\|\mathbf{V}^T\mathbf{V} - \mathbf{I}_k\|_F^2 \leq \epsilon$.

Besides capture the discriminative information in data, the approximate orthogonal constraints can also address the trivial solution and scale transfer problems. In fact, given a large β, i.e., $\beta = 10000$, only few elements in each row of $\mathbf{V}$ may have significantly large value because $\mathbf{V}$ is nonnegative and approximately orthogonal, and each column of $\mathbf{V}$ tends to be as different as possible to each other. Thus, even with heavy graph regularization, the solution to Eq. (12) can be nontrivial. Furthermore, if we substitute $\mathbf{V}^*$ by $\frac{1}{\gamma}\mathbf{V}^*$ ($\gamma > 1$), the term $\beta\|\mathbf{V}^T\mathbf{V} - \mathbf{I}_k\|_F^2$ may have larger value leading to larger objective function value with a large β. Consequently the scale transfer problem can be avoided. In addition, a proper β also results in sparse $\mathbf{V}$ for images. As mentioned in [5, 14, 18], the sparse representations can lead to better performance for visual analysis as clustering.

3.2 Optimization Algorithm

The objective function in Eq. (12) is non-convex with $\mathbf{W}, \mathbf{V}$ and $\mathbf{S}$ together. But fortunately, it's convex with respect to any one of them while fixing the others. So we can adopt an iterative strategy for optimizing Eq. (12).

3.2.1 Update U and V

For the convenience for derivation, firstly we can make the following denotations, $\hat{\mathbf{X}} = \mathbf{X} - \mathbf{S}$, $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, and $\hat{\mathbf{K}} = \hat{\mathbf{X}}^T\mathbf{X}$. By applying the properties of matrix norm $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}^T\mathbf{A})$, $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and $tr(\mathbf{A}) = tr(\mathbf{A}^T)$, we can obtain the objective function from Eq. (12) as follows,

$$\begin{aligned}\mathcal{O} = & tr(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W}\mathbf{V}^T) - 2tr(\hat{\mathbf{K}}\mathbf{W}\mathbf{V}^T) + \lambda\|\mathbf{S}\|_1 \\ & + \alpha tr(\mathbf{V}^T\mathbf{L}\mathbf{V}) + \beta tr(\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}) \\ & - 2\beta tr(\mathbf{V}^T\mathbf{V}) + \beta + tr(\hat{\mathbf{X}}^T\hat{\mathbf{X}}) \quad \text{s.t. } \mathbf{W}, \mathbf{V} \geq 0\end{aligned} \tag{13}$$

Now let ψ_{jl} and ϕ_{jl} be the Lagrange multiplier for constraints $w_{jl} \geq 0$ and $v_{jl} \geq 0$ respectively, and denote $\mathbf{\Psi} = [\psi_{jl}]$ and $\mathbf{\Phi} = [\phi_{jl}]$. Then we can write the Lagrange

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

The partial derivatives of $\mathcal{L}$ with respect to $\mathbf{W}$ and $\mathbf{V}$ are as

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}} = 2\mathbf{K}\mathbf{W}\mathbf{V}^T\mathbf{V} - 2\hat{\mathbf{K}}\mathbf{V} + \mathbf{\Psi} \tag{15}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{V}} = & 2\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W} - 2\hat{\mathbf{K}}\mathbf{W} + 2\alpha\mathbf{L}\mathbf{V} \\ & + 4\beta\mathbf{V}\mathbf{V}^T\mathbf{V} - 4\beta\mathbf{V} + \mathbf{\Phi}\end{aligned} \tag{16}$$

By using the Karush-Kuhn-Tucker conditions, i.e., $\psi_{jl}w_{jl} = 0$ and $\phi_{jl}v_{jl} = 0$, we get the following equations,

$$(\mathbf{K}\mathbf{W}\mathbf{V}^T\mathbf{V})_{jl}w_{jl} - (\hat{\mathbf{K}}\mathbf{V})_{jl}w_{jl} = 0 \tag{17}$$

$$\begin{aligned}&(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W})_{jl}v_{jl} - (\hat{\mathbf{K}}\mathbf{W})_{jl}v_{jl} + \alpha(\mathbf{L}\mathbf{V})_{jl}v_{jl} \\ &\quad + 2\beta(\mathbf{V}\mathbf{V}^T\mathbf{V})_{jl}v_{jl} - 2\beta(\mathbf{V})_{jl}v_{jl} = 0\end{aligned} \tag{18}$$

Then we obtain the following multiplicative updating rules:

$$w_{jl} \leftarrow w_{jl}\frac{(\hat{\mathbf{K}}\mathbf{V})_{jl}}{(\mathbf{K}\mathbf{W}\mathbf{V}^T\mathbf{V})_{jl}} \tag{19}$$

$$v_{jl} \leftarrow v_{jl}\frac{(\hat{\mathbf{K}}\mathbf{W} + \alpha\mathbf{G}\mathbf{V} + 2\beta\mathbf{V})_{jl}}{(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^T\mathbf{V})_{jl}} \tag{20}$$

3.2.2 Update S

We can observe that the optimization problem with respect to $\mathbf{S}$ is element-wise decoupled, i.e., we can optimize every element independently. Denote $\mathbf{E} = \mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{V}^T = [e_{ij}]$. Then each subproblem with respect to s_{ij} can be written as,

$$\mathcal{O}_{ij} = (e_{ij} - s_{ij})^2 + \lambda|s_{ij}| \tag{21}$$

After some simple derivation, we could obtain the solution to Eq. (21) as follows, which is also the updating rule for $\mathbf{S}$

$$s_{ij} = \begin{cases} 0, & \text{if } |e_{ij}| \leq \frac{\lambda}{2} \\ e_{ij} - \frac{\lambda}{2}\text{sign}(e_{ij}), & \text{otherwise} \end{cases} \tag{22}$$

In addition, by substituting Eq. (22) into Eq. (21), we have

$$\mathcal{O}_{ij} = \begin{cases} e_{ij}^2, & \text{if } |e_{ij}| \leq \frac{\lambda}{2} \\ \lambda|e_{ij}| - (\frac{\lambda}{2})^2, & \text{otherwise} \end{cases} \tag{23}$$

This is an interesting and important result. Intuitively, the large reconstruction error is often caused by data noise. If ℓ_2 norm (squared loss) is applied to all entries, the factorization will be dominated by the large-error entries, i.e., noise data. But from Eq. (23), we can observe that the factorization is self-adaptive to the reconstruction error. With the presence of $\mathbf{S}$, the factorization uses ℓ_1 norm as measure for large-error entries to alleviate the influence of noise, while ℓ_2 norm is applied to small error entries for more accurate factorization to capture the intrinsic information. Thus the factorization in RDCF can be robust to noise in image data.

4. THEORETICAL ANALYSIS

4.1 Proof of Convergence

We can use Eq. (19), Eq. (20) and Eq. (22) iteratively to update $\mathbf{W}$, $\mathbf{V}$ and $\mathbf{S}$ respectively, and the value of objective function in Eq. (12) will finally converge to a local minimum, which is theoretically guaranteed by Theorem 1.

THEOREM 1. *Objective function $\mathcal{O}$ in Eq. (12) is nonincreasing under rules in Eq. (19), Eq. (20) and Eq. (22).*

In fact, it's obvious that $\mathcal{O}_{ij}$ reaches the minimum when s_{ij} is computed as Eq. (22). Thus $\mathcal{O}$ is definitely nonincreasing under the updating rule for $\mathbf{S}$ in Eq. (22). Now we need to prove that $\mathcal{O}$ is nonincreasing under Eq. (19) and Eq. (20). Because of the limitation of space, we just show the the proof with respect to Eq. (20). Actually, the proof with respect to Eq. (19) is analogous. Our proof takes advantage of the *auxiliary function* [6] which can be defined as follows

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

$$G(v, v') \geq F(v), \quad G(v, v) = F(v) \tag{24}$$

are satisfied.

Then we need to make use of an important lemma as below,

LEMMA 1. If $G(v, v')$ is an auxiliary function of $F(v)$, then $F(v)$ is nonincreasing under the following updating rule

$$v^{(t+1)} = \arg\min_v G(v, v^{(t)}) \tag{25}$$

PROOF (PROOF OF LEMMA 1).

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

□

Now we need to show the updating rule for $\mathbf{V}$ in Eq. (20) is exactly the update in Eq. (25) with a proper auxiliary function. Let F_{ab} denote the the part of $\mathcal{O}$ that is only relevant to v_{ab}. The second-order partial derivative of F_{ab} is as follows

$$\begin{aligned}F_{ab}^{''} = & 2(\mathbf{W}^T\mathbf{K}\mathbf{W})_{bb} + 2\alpha\mathbf{L}_{aa} - 4\beta\mathbf{I}_{ab}^{ab} \\ & + 4\beta(\mathbf{I}^{ab}\mathbf{V}^T\mathbf{V} + \mathbf{V}\mathbf{I}^{ab^T}\mathbf{V} + \mathbf{V}\mathbf{V}^T\mathbf{I}^{ab})_{ab}\end{aligned} \tag{26}$$

where $\mathbf{I}^{ab}$ is a $n \times k$ matrix with 1 at (a, b) and 0 at all others. It's easy to validate the following three inequalities,

$$(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W})_{ab} \geq v_{ab}^{(t)}(\mathbf{W}^T\mathbf{K}\mathbf{W})_{bb} \tag{27}$$

$$(\mathbf{D}\mathbf{V})_{ab} = \sum_{i=1}^{n} D_{aj}v_{jb}^{(t)} \geq (\mathbf{D} - \mathbf{W})_{aa}v_{ab}^{(t)} \tag{28}$$

$$\begin{aligned}(\mathbf{V}\mathbf{V}^T\mathbf{V})_{ab} = & \sum_{j=1}^{n}\sum_{i=1}^{k} v_{ai}v_{ji}v_{jb} \\ \geq & (\mathbf{I}^{ab}\mathbf{V}^T\mathbf{V} + \mathbf{V}\mathbf{I}^{ab^T}\mathbf{V} + \mathbf{V}\mathbf{V}^T\mathbf{I}^{ab} - 1)_{ab}v_{ab}\end{aligned} \tag{29}$$

Table 3: 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

Here we need to point out that the inequality in (29) does not mathematically and strictly hold. In fact, during the derivation of it, we need to prove that $\sum_{j\neq a}^{n}\sum_{i\neq b}^{k} v_{ai}v_{ji}v_{jb} \geq 2v_{ab}^{3} - v_{ab}$. In fact, just by setting $\mathbf{V} = \mathbf{I}^{ab}$, this inequality is wrong. However, in real application, this inequality is always true because of the following two reasons. First, if the right part of the equality is nonpositive, e.g., $v_{ab} \in [0, \frac{1}{\sqrt{2}}]$, the left part which is nonnegative is definitely greater than the nonpositive right part. In fact, if $\mathbf{V}$ can capture the structure of $\mathbf{F}$, the largest element in $\mathbf{V}$ is close to $\frac{1}{\sqrt{n_j}}$ where $n_j \gg 2$ in real world. Thus the right part is always nonpositive. And this can be guaranteed by setting a large value for β. Second, even though the right part is positive, we can observe that the left part contains $(n-1)(k-1)$ terms, whose sum is always very large. Thus we can expect the sum of these terms could generate a larger value than the right part.

Lemma 2. The function

$$\begin{aligned} G(v, v_{ab}^{(t)}) &= F_{ab}(v_{ab}^{(t)}) + F_{ab}^{'}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^{T}\mathbf{V})_{ab}}{v_{ab}^{(t)}}(v - v_{ab}^{(t)})^{2} \end{aligned} \tag{30}$$

is an auxiliary function for $F_{ab}(v)$.

Proof (proof to Lemma 2). It's obvious that $G(v, v) = F_{ab}(v)$. Now we need to show $G(v, v_{ab}^{(t)}) \geq F_{ab}(v)$. Here we compare the Taylor series expansion of $F_{ab}(v)$ at $v_{ab}^{(t)}$ as

$$\begin{aligned} F_{ab}(v) =& F_{ab}(v_{ab}^{(t)}) + F_{ab}^{'}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{1}{2}F_{ab}^{''}(v_{ab}^{(t)})(v - v_{ab}^{(t)})^{2} \end{aligned} \tag{31}$$

Based on the definition in Eq. (26) and Eq. (30), and three important inequalities mentioned in Eq. (27), Eq. (28) and Eq. (29), it's very easy to validate that $G(v, v_{ab}^{(t)}) \geq F_{ab}(v)$. □

Proof (proof of Theorem 1). We can replace $G(v, v_{ab}^{(t)})$ in Eq. (25) by Eq. (30), which result in the following rule,

$$\begin{aligned} v_{ab}^{(t+1)} =& v_{ab}^{(t)} - v_{ab}^{(t)}\frac{F_{ab}^{'}(v_{ab}^{(t)})}{2(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^{T}\mathbf{V})_{ab}} \\ =& v_{ab}^{(t)}\frac{(\hat{\mathbf{K}}\mathbf{W} + \alpha\mathbf{G}\mathbf{V} + 2\beta\mathbf{V})_{ab}}{(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W} + \alpha\mathbf{D}\mathbf{V} + 2\beta\mathbf{V}\mathbf{V}^{T}\mathbf{V})_{ab}} \end{aligned} \tag{32}$$

which is identical to Eq. (20). Because $G(v, v_{ab}^{(t)})$ is an auxiliary function of F_{ab}, F_{ab} is nonincreasing under this updating rule. Therefore $\mathcal{O}$ is nonincreasing under Eq. (20). □

4.2 Complexity Analysis

The time complexity for p-nearest graph construction is $\mathcal{O}(n^{2}d)$. In each iteration, the complexity for updating is $\mathcal{O}(n^{2}d + n^{2}k)$. Suppose the optimization terminate at iteration t, thus the overall complexity is $\mathcal{O}(n^{2}((t+1)d + tk))$, which is linear to n^{2}. Consequently RDCF, LCCF and LCF will have similar computational complexity with a large n.

5. EXPERIMENT AND DISCUSSION

5.1 Datasets, Metrics and Baseline Methods

To demonstrate the effectiveness of RDCF for image representation, we carried out extensive experiment on five public image dataset, ORL[1], YALE[2], UMIST[3], MNIST[4] and Semeion[5]. The details of them are presented in Table 3.

Following previous works [3, 5, 18], we utilize clustering performance to evaluate the effectiveness of image representation. And Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) are adopted as the evaluation metrics for clustering, whose definitions are as follows

$$\text{ACC} = \frac{\sum_{i=1}^{n}\delta(s_i, map(r_i))}{n} \tag{33}$$

$$\text{NMI}(C, C') = \frac{\text{MI}(C, C')}{\sqrt{H(C)H(C')}} \tag{34}$$

$\delta(x, y)$ is the indicator function that equals one if $x = y$ and zero otherwise. $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data corpus. And the best mapping can be found by the Kuhn-Munkres algorithm [20]. $H(C)$ and $H(C')$ are the entropies of C and C' respectively. And $\text{MI}(C, C')$ is the mutual information between C and C' defined as below

$$\text{MI}(C, C') = \sum_{c_i\in C, c_j^{'}\in C'} p(c_i, c_j^{'}) \cdot \log_2 \frac{p(c_i, c_j^{'})}{p(c_i)\cdot p(c_j^{'})} \tag{35}$$

where $p(c_i)$ and $p(c_j^{'})$ represents the probabilities that an image arbitrarily selected from the data corpus belongs to clusters c_i and $c_j^{'}$ respectively. Analogously, $p(c_i, c_j^{'})$ stands for the joint probability that the any arbitrarily selected image belongs to the clusters c_i and $c_j^{'}$ at the same time.

We compare RDCF to the following NMF and CF methods. Kmeans is chosen as the base algorithm. Then are NMF [16], CF [23] and LNMF [15]. Graph regularized methods, GNMF [3] and LCCF [2]. NLCF [5] and LCF [18] based on Local Coordinate Coding. We also compare RDCF to NSDR [24] which is a spectral clustering method taking discriminative information into consideration. For all baseline methods above except NSDR, the cluster label can be generated from $\mathbf{V}$ in two ways. One is applying Kmeans to the learned representation and the other is setting the cluster label of image i as $c = \text{argmax}_j \mathbf{v}_{ij}$. Both ways are applied and the best performance of each baseline method is presented. And RDCF just utilizes the latter one.

5.2 Implementation Details

There are several important tunable parameters for baseline methods. For meaningful comparison, we perform grid

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive

[2] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[3] http://www.sheffield.ac.uk/eee/research/iel/research/face

[4] http://yann.lecun.com/exdb/mnist/

[5] https://archive.ics.uci.edu/ml/datasets

Table 4: Clustering Accuracy (%)

Dataset	Kmeans	NMF	LNMF	GNMF	NLCF	NSDR	CF	LCF	LCCF	RDCF
ORL	41.00	52.75	52.50	54.50	53.75	57.75	53.25	54.00	53.50	**63.00**
YALE	32.73	35.15	36.15	39.39	41.21	39.39	36.00	40.14	39.15	**47.23**
UMIST	46.48	46.23	47.34	56.28	53.37	64.08	47.11	54.13	56.28	**68.59**
MNIST	47.50	47.90	47.10	50.70	48.70	55.80	47.40	51.00	52.00	**63.50**
Semeion	55.56	45.49	46.42	58.43	56.41	63.34	45.21	57.21	55.49	**73.57**
Average	44.65	45.50	45.90	51.86	50.69	56.07	45.79	51.30	51.29	**63.19**

Table 5: Normalized Mutual Information (%)

Dataset	Kmeans	NMF	LNMF	GNMF	NLCF	NSDR	CF	LCF	LCCF	RDCF
ORL	67.01	74.76	73.85	75.81	74.90	75.78	75.04	75.22	76.36	**80.67**
YALE	40.32	44.97	44.84	46.37	48.80	48.29	44.82	47.17	46.58	**53.43**
UMIST	63.81	64.74	65.16	75.85	72.13	76.03	65.02	74.31	76.18	**82.37**
MNIST	47.16	44.93	44.84	48.19	50.09	58.11	44.61	51.05	54.02	**65.17**
Semeion	50.94	41.04	42.27	55.88	54.21	61.90	41.84	53.57	51.60	**65.87**
Average	53.85	54.09	54.19	60.42	60.03	64.02	54.27	60.26	60.95	**69.50**

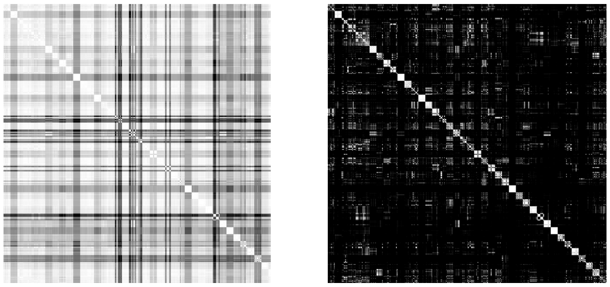

(a) Similarity of LCCF (b) Similarity of RDCF

Figure 1: Similarity between Images

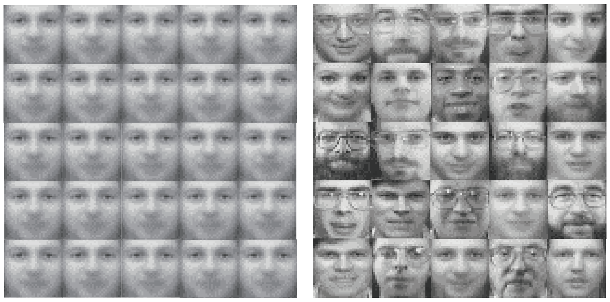

(a) LCCF (b) RDCF

Figure 2: Basis Images

search in the parameter space for each method and the best results are reported. Specifically, we search p, the number of nearest neighbors for constructing NN-graph, in $\{1, 2, ..., 9\}$. And the graph regularization parameter α for GNMF and LCCF is chosen from $\{0.1, 0.5, 1, 5, ..., 10^5\}$. The regularization parameter for NLCF and LCF is selected from $\{10^{-2}, ..., 10^4\}$. And for all factorization-based methods including RDCF, we set k, the dimension of latent space, to the number of true classes of dataset, as in [3, 18].

There are also some important parameters for RDCF, i.e., p for constructing nearest neighbor graph, λ for noise regularization, α for graph regularization and β for approximate orthogonal constraints. Specifically, we set $\lambda = 2\max(\text{median}_{ij}(e_{ij}), \text{mean}_{ij}(e_{ij}))$. Therefore it can be self-adaptive to the dataset and can change automatically in each iteration. And we set $p = 5$ for MNIST and $p = 3$ for the others. We set $\alpha = 1,000$ for UMIST and Semeion and $\alpha = 500$ for the others. And we set $\beta = 1,000$ for all datasets. Actually, compared to LCCF and LCF, RDCF is more robust to parameter change. In the coming section, we conduct empirical analysis on parameter sensitivity, and the results show that RDCF can achieve superior and stable performance under a wide range of value for both α and β.

5.3 Clustering Performance

The clustering results of all methods on five datasets measured by ACC and NMI are shown in Table 4 and Table 5 respectively. We can observe that RDCF can significantly outperform all baseline methods regardless of datasets because RDCF are robust to data noise and can preserve locality and exploit discriminative information simultaneously. Besides validating the effectiveness and superiority of RDCF, the experiment also reveals some important points.

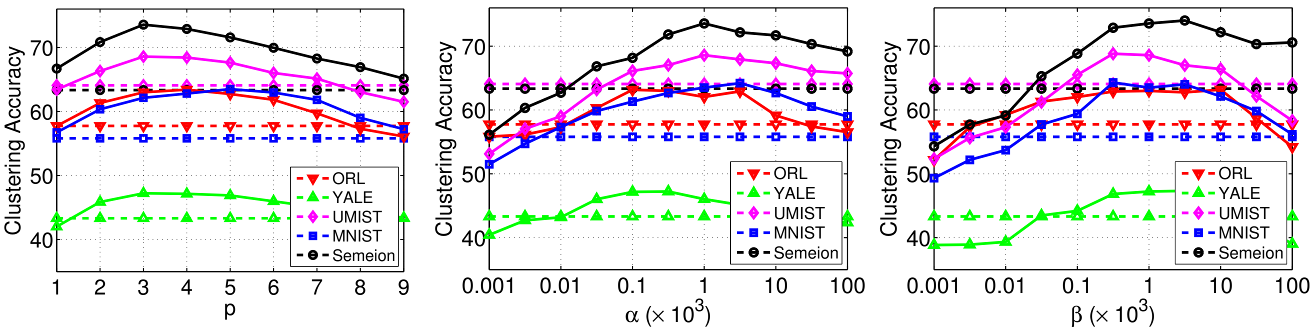
First, NMF and CF can outperform Kmeans though slightly. This phenomenon verifies the effectiveness of NMF and CF as feature learning methods, as discussed in [2, 3].

Second, we can observe that methods considering the local geometry structure of data, such as GNMF, LCCF, LCF and RDCF, significantly outperform NMF and CF, which also validates the importance of preserving locality of data.

Third, among all methods, NSDR and RDCF are the only two methods which explicitly exploit the discriminative information of data. And they are also the best two methods highlighting the power of discriminative information. But NSDR is affected by the noise and outlier in data, thus significantly degrading the performance compared with RDCF.

At last, RDCF is the only method satisfying all the following properties, i.e., being robust to data noise, preserving locality, exploiting discriminative information. Therefore it can achieve the best performance. Actually, satisfying any one of these properties can lead to better performance, and combining them all can fully exploit their power and they can as well promote each other for superior result.

Now we directly compare RDCF to LCCF on ORL dataset. The cosine similarity between images represented by the learned features (i.e., $\mathbf{V}$) is shown in Figure 1, where brighter color indicates larger similarity. We can observe RDCF achieves large inter-class similarity and small inter-class similarity, which is an ideal result, while LCCF achieves large



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

Figure 3: Parameter Sensitivity Analysis

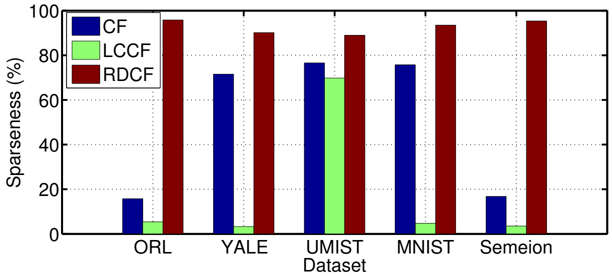


Figure 4: Sparseness Comparison.

inter and intra class similarity simultaneously. In addition, we show 25 basis images learned by LCCF and RDCF in Figure 2. Obviously, RDCF can capture more intrinsically discriminative information in images, while the basis learned by LCCF are almost indistinguishable. Although the representation learned by LCCF may achieve state-of-the-art performance for image clustering, this representation can't capture the discriminative information in image data, and is actually meaningless to some extent. But our RDCF is able to overcome the shortcomings of LCCF. RDCF can not only achieve much better clustering performance than state-of-the-art methods, but also learn discriminative and meaningful representation for images. Hence RDCF is more practical in real-world scenarios than LCCF.

5.4 Parameter Sensitivity Analysis

We further conduct empirical analysis on parameter sensitivity on all datasets. The results are shown in Figure 3. The dashed lines are the best results of all baseline methods.

The performance of RDCF with respect to p is shown in Figure 3(a). p controls the complexity of the graph. If it's too small, the graph may be too simple to fully exploit the local information. If it's too large, two images with different label may be connected such that decrease the quality of graph. Both will degrade the performance of graph-regularized methods, like GNMF, LCCF and RDCF. RDCF outperforms best baselines on each dataset when $p \in [2, 7]$.

The influence of α is shown in Figure 3(b). α controls the weight of graph regularization. A small α will lead to weak regularization thus it can't affect the objective function such that the locality can't be preserved. While a too large α may cause trivial solution problem to graph-regularized methods, like GNMF and LCCF. Though RDCF has graph regularization, we also incorporate approximate orthogonal constraints such that the trivial solution problem can be effectively avoided. Thus RDCF can achieve superior performance even with a large α. In fact, GNMF and LCCF are sensitive to α to some extent. But RDCF can achieve superior and stable performance under a very wide range of parameter value, i.e., $\alpha \in [10, 10^5]$ and it markedly outperforms all baselines on all datasets when $\alpha \in [10^2, 5 \times 10^4]$.

We plot the performance of RDCF with respect to different values of β in Figure 3(c). The parameter β controls the orthogonality of learned representations. Theoretically, if β is too small, the orthogonal constraint will be too weak and RDCF will be ill-defined and prone to trivial solution like GNMF and LCCF. On the contrary, if β is too large, the constraint may dominate the objective function of RDCF and it's so heave that the learned representation can be extremely sparse (under the ultimate situation, i.e., $\beta = \infty$, there is only one non-zero element in each row of $\mathbf{V}$) which is also unexpected in real-world scenarios. Fortunately, we can observe from the results that RDCF consistently outperforms best baseline methods on five datasets when $\beta \in [10^2, 10^4]$.

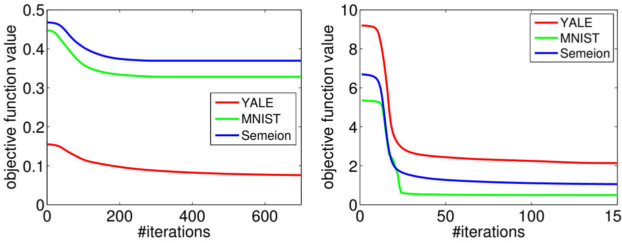
5.5 Other Issues

Previous works [5, 14, 18] mentioned that sparseness is also an important property of good image representation. Here we compare the sparseness of $\mathbf{V}$ learned by CF, LCCF and RDCF. As in [5], the sparseness is measured as follows,

$$\mathrm{SP}(\mathbf{V}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{k}-(\sum_{j=1}^{k}|v_{ij}|/\sqrt{\sum_{j=1}^{k}v_{ij}^2})}{\sqrt{k}-1} \qquad (36)$$

where $\mathrm{SP} \in [0, 1]$ and larger value indicates more sparse representation. The sparseness of them on five datasets are shown in Figure 4. We can observe that RDCF can indeed learn sparse representation while LCCF always learns dense one, and the sparseness of CF depends on datasets. Actually, larger β can result in more sparse representation. But this is unexpected in real world. With a proper β, the sparseness is about 90%, which leads to effective image representation.

It has been proven theoretically in Section 4 that the objective function will converge under the proposed multiplicative updating rules. Now we want to show how fast it can converge. The objective function value (averaged by the number of samples) with respect to the number of iterations is shown in Figure 5(b). We can observe that the objective function value is decreasing steadily with more iterations and can converge very fast, usually within 100 iterations. In addition, we compare the convergency property of RDCF to conventional CF whose result is shown in Figure 5(a). We can see RDCF can converge faster and more stablely than CF which requires more than 200 iterations to convergency.



(a) CF (b) RDCF

Figure 5: Convergency Study

6. CONCLUSION

In this paper, we propose a novel method called RDCF, which can simultaneously preserve local geometry structure and exploit discriminative information. It's also robust to data noise while previous works with squared loss can't address this problem. Furthermore, RDCF can avoid trivial solution and scale transfer problems even with graph regularization. All properties above make RDCF an effective method for image representation. We also propose an iterative strategy with multiplicative updating rules for the optimization of RDCF and prove the convergence theoretically. Extensive experiments on five public image datasets are carried out and the results demonstrate that RDCF can significantly outperform several state-of-the-art related methods. Analysis about parameter sensitivity validates that RDCF can achieve superior and stable performance under wide range of parameter values, even if it's regularized by graph.

7. ACKNOWLEDGEMENT

This research was supported the National Natural Science Foundation of China (Grant No. 61271394), and the National HeGaoJi Key Project (No. 2013ZX01039-002-002). And the authors would like to thank the anonymous reviewers for their valuable comments.

8. REFERENCES

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006.

[2] D. Cai, X. He, and J. Han. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2011.

[3] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transaction Pattern Anal. Mach. Intell.*, 2011.

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

[5] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He. Nonnegative local coordinate factorization for image representation. *IEEE Transactions on Image Processing*, 2013.

[6] A. P. Dempster, N. M. Laird, , and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] G. Ding, Y. Guo, and J. Zhou. collective matrix factorization hashing for multimodal data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014.

[8] R. O. Duda, P. Hart, and D. Stork. Pattern classification. In *Wiley-Interscience, Hoboken, NJ, 2nd edition*, 2000.

[9] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *TIP*, 2008.

[10] Q. Gu, C. Ding, and J. Han. On trivial solution and scale transfer problems in graph regularized nmf. In *Proc. of 22nd International Joint Conference on Artificial Intelligence*, 2011.

[11] Q. Gu and J. Zhou. Co-clustering on manifolds. In *Proc. of ACM SIGKDD*, 2009.

[12] Y. Guo, G. Ding, X. Jin, and J. Wang. Learning predictable and discriminative attributes for visual recognition. In *AAAI Conference on Artificial Intelligence*, 2015.

[13] X. He. Laplacian regularized d-optimal design for active learning and its application to image retrieval. *TIP*, 2010.

[14] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[15] D. Kong, C. Ding, and H. Huang. Robust nonnegative matrix factorization using l21-norm. In *CIKM*, 2011.

[16] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. In *Nature*, 1999.

[17] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

[18] H. Liu, Z. Yang, J. Yang, Z. Wu, and X. Li. Local coordinate concept factorization for image representation. *IEEE Trans. Neural Netw. Learning Syst.*, 2014.

[19] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19, 1996.

[20] L. Lovasz and M. Plummer. *Matching Theory*, 1986.

[21] D. Tao, X. Li, X. Wu, and S. J. Maybank. Geometric mean for subspace selection. *TPAMI*, 2009.

[22] 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*, 4(5):509–522, 1994.

[23] W. Xu and Y. Gong. Document clustering by concept factorization. In *Proc. of ACM RDIR*, 2004.

[24] Y. Yang, H. Shen, F. Nie, R. Ji, and X. Zhou. Nonnegative spectral clustering with discriminative regularization. In *Proc. of 25th AAAI*, 2011.

[25] Y. Yang, D. Xu, F. Nie, S. Yan, , and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 2010.

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

[27] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.