IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 7, 2018, accepted October 31, 2018, date of publication November 9, 2018, date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2879986

Robust Local Learning and Discriminative Concept Factorization for Data Representation

WEI JIANG[1], XIAOTING FENG[1], TINGTING MA[1], LING XING[2], AND KEWEI TANG[1]
[1]School of Mathematics, Liaoning Normal University, Dalian 116029, China
[2]School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

Corresponding author: Wei Jiang (swxxjiang@126.com)

This work was supported in part by the Natural Science Foundation of Liaoning under Grant 2015020070 and in part by the Natural Science Foundations of China under Grant 61771229, Grant 61702243, and Grant 61771185.

ABSTRACT Concept factorization (CF), as a matrix factorization method, has been applied widely in obtaining an optimal data representation and has yielded impressive results. However, some shortcomings exist in the existing CF method. 1) The standard concept factorization uses the squared loss function that is sensitive to outlier points and noises. 2) The graph generated by the original data does not reflect the real geometric structure of the data distribution. 3) The discriminant information is ignored. Herein, we propose a novel method, called robust local learning and discriminative concept factorization (RLLDCF) for data representation. Specifically, RLLDCF adopts the $\ell_{2,1}$-norm-based loss function to improve its robustness against noises and outliers, and exploits the discriminative information by local linear regression constraints. In addition, the method obtains the topology structure of the data distribution during learning rather than known a priori and fixed. A new iterative multiplicative updating rule is derived to solve RLLDCF's objective function. The convergence of the optimization algorithm is proved both theoretically and empirically. Numerous experiments on both synthetic and real-world datasets are conducted, and the results indicate that our proposed method is significantly better than all the comparison methods, thus validating the effectiveness and robustness of RLLDCF.

INDEX TERMS Clustering, data representation, $\ell_{2,1}$-norm, local adaptive learning, local linear regression.

I. INTRODUCTION

With the rapid development of intelligent technology, obtaining an optimal data representation [1]–[8] is a fundamental topic in machine learning, pattern recognition, and bioinformatics. An efficient data representation can reveal the potential geometrical structure of data and further promote performance of clustering and classification. Matrix factorization techniques [9]–[14] have been studied widely for data representation. The early works were primarily based on eigenvalue decomposition or singular value decomposition (SVD). Principle component analysis (PCA) [10] is a classical approach for data representation; it first maps high-dimensional data to a low-dimensional subspace by eigenvalue decomposition. Nonnegative matrix factorization (NMF) [9] is a useful data representation method that decomposes a nonnegative matrix into the product of two nonnegative matrices. NMF leads to the part-based representation of data owing to the nonnegativity constraints imposed on both basis and coefficient matrices. As an extension, concept factorization (CF) [15] was proposed for document clustering. In CF, the original data matrix is factorized into the production of three nonnegative matrices. The primary advantage of CF over NMF is that it can be used in either the original space or reproducing kernel Hilbert space (RKHS). When the data live or are close to the nonlinear low dimensional manifolds embedded in high dimensional ambient space, Euclidean distance can not capture the local structure of the data. Therefore, NMF and CF based on Euclidean distance fail to incorporate the local structure; they consider only the global geometric structure.

In recent years, various local learning models and techniques have been used widely to represent high-dimensional data, such as ISOMAP [16], Laplacian eigenmaps (LE) [17], and local linear embedding (LLE) [18], in order to obtain the optimal data representation. These methods have successfully characterized the intrinsic geometric structure of nonlinear data manifolds. Although they are successful in utilizing the local geometry of data space, these algorithms do not consider the distribution information of data, which is essential for data clustering. It is more reasonable to assign large weights to the data in the large density region [12]. In addition, it has been shown that these algorithms can be reformulated in a graph embedding framework.

Graph-based matrix factorization techniques [8], [19], [20], [20]–[28] have been gaining increasing interest. However, many existing methods establish graphs that are independent of learning processes. Once the graph is constructed, it is unchangeable and does not benefit subsequent learning tasks. The neighbor graph, which is constructed by the sample points in the high-dimensional space, is suboptimal, and cannot reflect the geometrical structure of the sample distribution effectively. To overcome this shortcoming, we attempt to obtain the optimal representation of data by integrating matrix decomposition and graph learning into a joint learning framework. Hence, we herein propose a robust local learning and discriminative concept factorization (RLLDCF) method; it can not only learn the local similarity matrix and discover the local discriminative structure of the data but also address the outlier points and noises simultaneously. Specifically, we used the $\ell_{2,1}$-norm as our error function, such that RLLDCF can accommodate the outliers and noises better than the standard CF using F-norm as the error function. We learned the optimal graph weights jointly with CF rather than fixing the graph in the low-dimensional representation space. Additionally, the local regressive regularization term ensures that data belonging to the same patch have the same representation.

The primary aspects of the proposed approach are listed as follows.

1. Our approach adds $\ell_{2,1}$-norm constraints that can learn data representation; this is the primary issue that is neglected by the previous imperfect method.

2. To discover the local discriminative structure of the data, we add local linear regression constraints into the learned space. By adding the local linear regression constrains, our approach possesses discriminative power than the other improved CFs.

3. The method performs matrix decomposition and structural learning simultaneously. It learns the intrinsic structure of the data adaptively, and can thus achieve a more valuable data representation.

4. We propose an effective iterative strategy with multiplicative updating rules for the optimization, and provide the proof of rigorous convergence and correctness analysis of our model.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the related works. In Section 3, we elaborate our proposed formulation and provide convergence proof. In Section 4, we describe our extensive experiments and analyze their results. Finally, we conclude and discuss future work.

II. RELATED WORKS

In this section, we briefly review the related research on data representation including graph-based matrix factorization techniques, $\ell_{2,p}$-norm ($0 < p \leq 1$)-based matrix factorization techniques, and locality-based approaches.

Graph-based matrix factorization techniques have been studied and used widely for data representation [8], [19], [20], [20]–[28]. By constructing similar graphs with vertices corresponding to data points and edge weights representing degrees of similarity, similar graph terms are introduced into objective functions. Thus, the data representation problem is formulated as a graph-based regularization problem [22]– [29]. Therefore, graphs are crucial in determining their ultimate performances. However, many existing methods establish graphs independent of learning processes. The adaptive graph methods [8], [20], [23]–[31], [31]–[33] perform feature extraction as part of the model construction process, which performs feature extraction and obtains the neighbor graph weight matrix simultaneously.

The graph regularized extensions of CF and NMF have shown good performance by applying an inherent local manifold structure. These methods typically adopt squared loss to measure the data reconstruction quality. Although the performance is encouraging in most cases, it is not always the optimal choice for decomposition of the data matrix. The study indicates that the square error is optimal for Poisson and Gaussian noises [33]. However, it almost always involves data that violate the hypothesis in practical applications. The squared loss is not robust to outlier points and noises that are typical in visual data, because a few large noisy entries may dominate data decomposition. Recently, some improved algorithms have been presented to enhance the robustness of classical NMF and CF. Robust matrix factorization utilized $\ell_{2,p}$-norm ($0 < p \leq 1$) [35], [37]-based objective function, which is not squared; thus, the outlier points and noises do not dominate the objective function.

Local approaches [24], [38]–[48] have been applied widely in many intelligent learning problems such as dimension reduction [18], [39], clustering [40], [41], and classification [24], [42]–[48]. Locally regressive projections (LRPs) [49] are built fundamentally upon the idea of local linear regression and recently applied to ranking [50] and co-clustering [51]. Recently, Li and Tang [21] proposed a novel method called weakly supervised deep matrix factorization (WSDM) to explore local learning for social image understanding, which uncovers the latent image representations and tag representations embedded in the latent subspace by collaboratively exploring the weakly supervised tagging information. Liu *et al.* [41] presented a locality-constrained concept factorization (LCF) method that imposes the locality constraint onto the original CF to explore a faithful intrinsic geometry. These methods have succeeded in discovering the local structure of nonlinear manifolds.

Notation: We use boldface capital and lowercase letters, $\mathbf{A}$ and $\mathbf{a}$, to denote matrices and vectors, respectively. $||\mathbf{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ is used to represent the ℓ_p-norm of the vector $\mathbf{x} \in \mathbb{R}^n$. We use $\mathbf{x}^i$ and $\mathbf{x}_j$ to denote the i-th row, and the j-th column of matrix $\mathbf{X}$, respectively. x_{ij} is the element in the i^{th} row and j^{th} column of $\mathbf{X}$; Tr[$\mathbf{X}$] denotes the trace of $\mathbf{X}$ if $\mathbf{X}$ is a square matrix; $\mathbf{X}^T$ denotes the transposed matrix of $\mathbf{X}$. For any two matrices $\mathbf{X}$ and $\mathbf{Y}$, we let $<\mathbf{X}, \mathbf{Y}>= \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$ be the inner product. We define the Frobenius norm of the matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ as

$$||\mathbf{X}||_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^2}. \tag{1}$$

We denote the $\ell_{2,1}$-norm of a matrix as

$$||\mathbf{X}||_{2,1} = \sum_{i=1}^{M} ||\mathbf{x}^i||_2 = \sum_{i=1}^{M}\sqrt{\sum_{j=1}^{N} x_{ij}^2} = \mathrm{Tr}[\mathbf{X}^T\mathbf{D}\mathbf{X}], \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \frac{1}{2||\mathbf{x}^i||_2}$. However, $\|\mathbf{x}^i\|_2$ could approach zero. For this case, we let $D_{ii} = \frac{1}{2\|\mathbf{x}^i\|_2+\epsilon}$, where ϵ is a small constant.

III. REVIEW OF MATRIX FACTORIZATION

In this section, we present a brief review of NMF, CF, and LCCF.

A. NMF

Given a nonnegative matrix $\mathbf{X}$, consisting of N column vectors of dimensionality M, each column of $\mathbf{X}$ is a data point. The idea of NMF is to obtain nonnegative matrices $\mathbf{U}$ and $\mathbf{V}$ to reconstruct the original data matrix $\mathbf{X}$. The corresponding optimization problem is as follows:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2,$$
$$s.t.\ \mathbf{U} \geq 0, \mathbf{V} \geq 0, \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm. $\mathbf{U} \in \mathbb{R}^{M\times K}$ is the basis matrix, and $\mathbf{V} \in \mathbb{R}^{K\times M}$ is the coefficient matrix. Lee and Seung [9] proposed the following iterative multiplicative updating rules to minimize the objective function:

$$u_{jk}^{(t+1)} \leftarrow u_{jk}^{(t)} \frac{(\mathbf{X}\mathbf{V}^T)_{jk}}{(\mathbf{U}\mathbf{V}\mathbf{V}^T)_{jk}},$$
$$v_{ki}^{(t+1)} \leftarrow v_{ki}^{(t)} \frac{(\mathbf{U}^T\mathbf{X})_{ki}}{(\mathbf{U}^T\mathbf{U}\mathbf{V})_{ki}}. \tag{4}$$

Generally, $K < min(M, N)$, and we view $\mathbf{V}$ as the low-dimensional representations.

B. CF

The idea of CF [15] is to characterize every concept as a linear combination of the entire data points, each of which is the linear combination of all the concepts. The objective function of CF is written as follows:

$$\min_{\mathbf{W},\mathbf{V}} \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2$$
$$s.t.\ \mathbf{W} \geq 0, \quad \mathbf{V} \geq 0. \tag{5}$$

Defining the kernel matrix as $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, the matrices $\mathbf{W}$ and $\mathbf{V}$ are obtained by the following multiplicative updating rules:

$$w_{jk} \leftarrow w_{jk} \frac{(\mathbf{K}\mathbf{V})_{jk}}{(\mathbf{K}\mathbf{W}\mathbf{V}^T\mathbf{V})_{jk}},$$
$$v_{ki} \leftarrow v_{ki} \frac{(\mathbf{K}\mathbf{W})_{ki}}{(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W})_{ki}}. \tag{6}$$

C. LOCALLY CONSISTENT CONCEPT FACTORIZATION

Locally consistent concept factorization (LCCF) [23] is presented to obtain concepts by constructing the nearest neighbor graph. The optimization problem of LCCF can be obtained as follows:

$$\min_{\mathbf{W},\mathbf{V}} \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{V}\|_F^2 + \lambda \mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})$$
$$s.t.\ \mathbf{W} \geq 0, \mathbf{V} \geq 0, \tag{7}$$

where λ is the regularization parameter, and $\mathbf{L}$ is the graph Laplacian matrix (see [23] for details). The multiplicative updates rules can be obtained by solving the following problem:

$$w_{jk} \leftarrow w_{jk} \frac{(\mathbf{K}\mathbf{V})_{jk}}{(\mathbf{K}\mathbf{W}\mathbf{V}^T\mathbf{V})_{jk}},$$
$$v_{ki} \leftarrow v_{ki} \frac{(\mathbf{K}\mathbf{W} + \lambda\mathbf{S}\mathbf{V})_{ki}}{(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W} + \lambda\mathbf{D}\mathbf{V})_{ki}}. \tag{8}$$

where $\mathbf{D}$ is a diagonal matrix whose entries are column sums of $\mathbf{S}$, which is called weight matrix.

IV. THE PROPOSED RLLDCF FRAMEWORK

In this section, we present a novel concept decomposition method for data representation, called RLLDCF, which can achieve a more effective and robust data representation.

A. THE OBJECTIVE FUNCTION

Recall that CF attempts to obtain a set of basis vectors that can be used to represent the original data. The coefficient matrix can be regarded as the new representation of each data point in the new basis. In this work, the learning data representation is expected to satisfy the following three properties.

i) The learning data representation is robust to outliers and noises.

ii) The learning data representation can reflect the true geometrical structure of the data distribution.

iii) The learning data representation can discover the intrinsic discriminant structure of the data space.

To satisfy the first property, we use the $\ell_{2,1}$-norm loss function, instead of the square loss function based on the Frobenius norm, to alleviate the influence of noises and outliers. Therefore, our robust concept representation model is represented as

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{V}||_{2,1} + \alpha||\mathbf{V}||_{2,1}, \tag{9}$$

where $\alpha > 0$ is the regularization parameter. The $\ell_{2,1}$-norm-based objective function promotes robustness if the dataset is corrupted by outlier points and noises. Meanwhile, the regularization term $||\mathbf{V}||_{2,1}$ ensures that $\mathbf{V}$ is sparse in rows.

Despite the popularity of exploiting graph regularization for learning an optimal data representation, they do not satisfy properties (i) and (ii). To overcome the shortcomings, we propose a new method to learn the reconstruction coefficients matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, where $\mathbf{S}_{ij}$ represents the i-th data's contribution to reconstruct data point x_j. Because $\mathbf{S}$ is obtained by adaptive learning, our robust local adaptive learning structure can be formulated as follows:

$$\min ||\mathbf{V} - \mathbf{VS}||_{2,1} + \mu \mathrm{Tr}(\mathbf{ES}). \tag{10}$$

where the $\ell_{2,1}$-norm based function is $||\mathbf{V} - \mathbf{VS}||_{2,1} = \sum_i ||v_i - \sum_j S_{ij} v_j||$. Because the $\ell_{2,1}$-norm reduces the components occupied by the large magnitude of error in the loss function, the corrupted samples never dominate the objective function. In this sense, the loss function $||\mathbf{V} - \mathbf{VS}||_{2,1}$ is robust to outlier points and noises. To satisfy the second property, we learn the sparse weight matrix $\mathbf{S}$, where $\mathbf{S}_{ij}$ denotes the i-th data's contribution to reconstruct data point $\mathbf{v}_j$. We optimize the representation matrix $\mathbf{V}$ and sparse weight matrix $\mathbf{S}$ simultaneously in a single objective function. The learned sparse weight matrix $\mathbf{S}$ is more flexible and related to both the original and transformed spaces. Specifically, we repeat learning $\mathbf{V}$ and $\mathbf{S}$ iteratively for improving the model performance. Here, we use the learned representation matrix $\mathbf{V}$ to learn the sparse matrix $\mathbf{S}$, which leads to a better $\mathbf{V}$ in turn. Compared with the sparse similarity, the learned sparse weight matrix can reflect the true geometrical structure of the data distribution. The second term $\mathrm{Tr}(\mathbf{ES})$ is equivalent to $||\mathbf{S}||_1$, with $||\mathbf{S}|| \geq 0$, and $\mathbf{E}$ is a matrix of all ones. It boosts the sparsity of the solution. μ is the regularization parameter.

To promote the third property, we construct the local predictors [50], and derive a local linear regression function. For each sample point $\mathbf{x}_i$, we define the local region $\mathcal{M}(x_i)$ as the set consisting of $\mathbf{x}_i$ and its $n_i - 1$ nearest neighbor points. We use set $\mathcal{A}_i = \{j|\mathbf{x}_j \in \mathcal{M}(\mathbf{x}_i)\}$ to denote the indices of the samples in $\mathcal{M}(\mathbf{x}_i)$. We use $\mathbf{X}_i \in \mathbb{R}^{M\times n_i}$ to denote the local data matrix containing samples in $\mathcal{M}(\mathbf{x}_i)$. We use matrix $\mathbf{V}_i \in \mathbb{R}^{K\times n_i}$ to denote the new representations of points in $\mathcal{M}(\mathbf{x}_i)$. The primary goal of the local regressive function is to obtain the relation by modeling $\mathbf{X}_i$ to $\mathbf{V}_i$. To fit such relation, we utilize the regularized linear regression function, which minimizes the following objective function:

$$\min_{G_i,\mathbf{b}_i} \frac{1}{n_i} ||\mathbf{V}_i - \mathbf{G}_i^T \mathbf{X}_i - \mathbf{b}_i \mathbf{1}_{n_i}^T||_F^2 + \lambda ||\mathbf{G}_i||_F^2. \tag{11}$$

where λ is a positive regularization parameter. Setting the partial derivatives of the objective function with respect to $\mathbf{W}$ and $\mathbf{b}$ to zero, we obtain

$$G_i^* = (X_i \Pi X_i^T + n_i \lambda I)^{-1} X_i \Pi Y_i^T \tag{12}$$

$$\mathbf{b}_i^* = \frac{1}{n_i}(\mathbf{V}_i - (\mathbf{W}_i^*)^T \mathbf{X}_i)\mathbf{1}_{n_i} \tag{13}$$

where $\mathbf{I}$ is the identity matrix and $\Pi = \mathbf{I} - \frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$ is the centering matrix. Substituting Eq.(12) and Eq.(13) into Eq.(11), the fitting error of the local regressive function is given by

$$\mathcal{L} = \frac{1}{n_i} \mathrm{Tr}(\mathbf{V}_i(\Pi - \Pi \mathbf{X}_i^T(\mathbf{X}_i \Pi \mathbf{X}_i^T + n_i \lambda \mathbf{I})^{-1} \mathbf{X}_i \Pi)\mathbf{V}_i^T). \tag{14}$$

Therefore, the local regressive model can be rewritten as [42]:

$$\mathcal{L} = \mathrm{Tr}(\mathbf{V}_i \mathbf{M}_i \mathbf{V}_i^T) \tag{15}$$

where

$$\mathbf{M}_i = \lambda(n_i \lambda \mathbf{I} + \Pi \mathbf{X}_i^T \mathbf{X}_i \Pi)^{-1} \Pi. \tag{16}$$

Because the local matrix $\mathbf{V}_i$ is part of the global matrix $\mathbf{V}$, we can construct a selection matrix $\mathbf{S}_i \in \mathbb{R}^{M\times n_i}$ for each $\mathbf{V}_i$ such that

$$\mathbf{V}_i = \mathbf{VS}_i, \tag{17}$$

where the selection matrix $\mathbf{S}_i$ is defined as follows: $\mathbf{S}_i = [\mathbf{e}_j]$ for $j \in \mathcal{M}(\mathbf{x}_i)$, where $\mathbf{e}_j$ is an m-dimensional vector whose j-th element is one, and all other elements are zero. After the local matrices are established, we combine them by minimizing the following loss function:

$$\begin{aligned}\min_{\mathbf{V}_i} \sum_{i=1}^{N} \mathrm{Tr}(\mathbf{V}_i \mathbf{M}_i \mathbf{V}_i^T) &= \min_{\mathbf{V}} \sum_{i=1}^{N} \mathrm{Tr}(\mathbf{VS}_i \mathbf{M}_i \mathbf{S}_i^T \mathbf{V}^T) \\ &= \min_{\mathbf{V}} \mathrm{Tr}(\mathbf{VMV}^T),\end{aligned} \tag{18}$$

where

$$\mathbf{M} = \sum_{i=1}^{N} \mathbf{S}_i \mathbf{M_i} \mathbf{S_i^T}. \tag{19}$$

By integrating objective functions (9), (10), and (18), the overall loss function is defined as

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},S\geq 0} &\|\mathbf{X} - \mathbf{XUV}\|_{2,1} + \beta(||\mathbf{V} - \mathbf{VS}||_{2,1} \\ &+ \mu \mathrm{Tr}(\mathbf{ES})) + \gamma \mathrm{Tr}(\mathbf{VMV}^T) + \alpha ||\mathbf{V}||_{2,1},\end{aligned} \tag{20}$$

where β and γ are two trade-off parameters.

We denote $\mathbf{X} - \mathbf{XUV} = [\mathbf{a}^1, \cdots, \mathbf{a}^M]^T$, $\mathbf{V} - \mathbf{VS} = [\mathbf{b}^1, \cdots, \mathbf{b}^K]^T$ and $\mathbf{V} = [\mathbf{v}^1, \cdots, \mathbf{v}^K]^T$. When the nonnegative constraint is considered on $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$, the objective function (20) can be reformulated as

$$\begin{aligned}\mathcal{O} = &\mathrm{Tr}((\mathbf{X} - \mathbf{XUV})^T \mathbf{A}(\mathbf{X} - \mathbf{XUV})) \\ &+ \beta \mathrm{Tr}((\mathbf{V} - \mathbf{VS})^T \mathbf{B}(\mathbf{V} - \mathbf{VS}) + \mu \mathrm{Tr}(\mathbf{ES})) \\ &+ \gamma \mathrm{Tr}(\mathbf{VMV}^T) + \lambda \mathrm{Tr}(\mathbf{V}^T \mathbf{CV}), \\ &s.t. \quad \mathbf{U} \in \mathbb{R}^{M\times K} > 0; \quad \mathbf{V} \in \mathbb{R}^{K\times N} > 0 \\ &\quad \mathbf{S} \in \mathbb{R}^{N\times L_c} > 0\end{aligned} \tag{21}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are three diagonal matrices with their diagonal elements as $\mathbf{A}_{ii} = \frac{1}{2||\mathbf{a}^i||_2}$, $\mathbf{B}_{ii} = \frac{1}{2||\mathbf{b}^i||_2}$, and $\mathbf{C}_{ii} = \frac{1}{2||\mathbf{v}^i||_2}$, respectively.

B. THE UPDATE RULES

The proposed formulation (21) is not convex in $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$ together. Fortunately, we propose an algorithm for optimizing the objective function with respect to any one of them while maintaining the others. Based on the properties of the matrix norm, we obtain the objective function as follows:

$$\begin{aligned}\mathcal{O} = &\operatorname{Tr}(\mathbf{X}\mathbf{X}^T\mathbf{A} + \mathbf{V}^T\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V} - 2\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\\ &+ \beta\operatorname{Tr}(\mathbf{V}\mathbf{V}^T\mathbf{B} + \mathbf{S}^T\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S} - 2\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S})\\ &+ \gamma(\operatorname{Tr}(\mathbf{V}\mathbf{M}^+\mathbf{V}^T) - \operatorname{Tr}(\mathbf{V}\mathbf{M}^-\mathbf{V}^T))\\ &+ \lambda\operatorname{Tr}(\mathbf{V}^T\mathbf{C}\mathbf{V}) + \beta\mu\operatorname{Tr}(\mathbf{E}\mathbf{S})\\ &s.t.\quad \mathbf{U} \in \mathbb{R}^{M\times K} > 0;\quad \mathbf{V} \in \mathbb{R}^{K\times N} > 0\\ &\qquad \mathbf{S} \in \mathbb{R}^{N\times L_c} > 0 \end{aligned} \tag{22}$$

where $\mathbf{M} = \mathbf{M}^+ - \mathbf{M}^-$, with $\mathbf{M}^+_{ij} = \frac{(|\mathbf{M}_{ij}|+\mathbf{M}_{ij})}{2}$ and $\mathbf{M}^-_{ij} = \frac{(|\mathbf{M}_{ij}|-\mathbf{M}_{ij})}{2}$.

We introduce the Lagrange multiplier matrix $\Psi = [\psi_{jk}]$, $\Phi = [\phi_{ki}]$ and $\Theta = [\theta_{iq}]$ for the nonnegative constraint on $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$. The corresponding Lagrangian function $\mathcal{L}$ is defined as follows:

$$\begin{aligned}\mathcal{L} = &\operatorname{Tr}(\mathbf{X}\mathbf{X}^T\mathbf{A}+\mathbf{V}^T\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}-2\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\\ &+ \beta\operatorname{Tr}(\mathbf{V}\mathbf{V}^T\mathbf{B}+\mathbf{S}^T\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S}-2\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S}) + \beta\mu\operatorname{Tr}(\mathbf{E}\mathbf{S})\\ &+ \gamma(\operatorname{Tr}(\mathbf{V}\mathbf{M}^+\mathbf{V}^T) - \operatorname{Tr}(\mathbf{V}\mathbf{M}^-\mathbf{V}^T)) + \lambda\operatorname{Tr}(\mathbf{V}^T\mathbf{C}\mathbf{V})\\ &- \operatorname{Tr}(\Psi\mathbf{U}^T) - \operatorname{Tr}(\Phi\mathbf{V}^T) - \operatorname{Tr}(\Theta\mathbf{S}^T),\end{aligned} \tag{23}$$

Let the partial derivatives of the optimization function (23) with respect to $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$ be zero. Therefore, we have

$$\Psi = 2\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\mathbf{V}^T - 2\mathbf{X}^T\mathbf{A}^T\mathbf{X}\mathbf{V}^T, \tag{24}$$

$$\begin{aligned}\Phi = &\,2\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V} - 2\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X}\\ &+ 2\beta(\mathbf{B}\mathbf{V} + \mathbf{B}\mathbf{V}\mathbf{S}\mathbf{S}^T - 2\mathbf{B}\mathbf{V}\mathbf{S})\\ &+ 2\gamma(\mathbf{V}\mathbf{M}^+ - \mathbf{V}\mathbf{M}^-) + 2\lambda\mathbf{C}\mathbf{V},\end{aligned} \tag{25}$$

$$\Theta = 2\beta(\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S} - \mathbf{V}^T\mathbf{B}\mathbf{V}) + \beta\mu\mathbf{E}^T. \tag{26}$$

Based on the KKT conditions [48] $\psi_{jk}u_{jk} = 0$, $\phi_{ki}v_{ki} = 0$ and $\theta_{iq}y_{iq} = 0$, we obtain

$$(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\mathbf{V}^T)_{jk}u_{jk} - (\mathbf{X}^T\mathbf{A}^T\mathbf{X}\mathbf{V}^T)_{jk}u_{jk} = 0, \tag{27}$$

$$\begin{aligned}&(\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V})_{ki}v_{ki} - (\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X})_{ki}v_{ki}\\ &\quad + \beta(\mathbf{B}\mathbf{V} + \mathbf{B}\mathbf{V}\mathbf{S}\mathbf{S}^T - 2\mathbf{B}\mathbf{V}\mathbf{S})_{ki}v_{ki}\\ &\quad +\gamma(\mathbf{V}\mathbf{M}^+ - \mathbf{V}\mathbf{M}^-)_{ki}v_{ki} + \lambda(\mathbf{C}\mathbf{V})_{ki}v_{ki} = 0,\end{aligned} \tag{28}$$

$$2\beta(\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S} - \mathbf{V}^T\mathbf{B}\mathbf{V})_{iq}s_{iq} + \beta\mu(\mathbf{E}^T)_{iq}s_{iq} = 0. \tag{29}$$

The corresponding equivalent formulas are as follows:

$$(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\mathbf{V}^T)_{jk}u^2_{jk} - (\mathbf{X}^T\mathbf{A}^T\mathbf{X}\mathbf{V}^T)_{jk}u^2_{jk} = 0, \tag{30}$$

$$\begin{aligned}&(\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V})_{ki}v^2_{ki} - (\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X})_{ki}v^2_{ki}\\ &\quad + \beta(\mathbf{B}\mathbf{V} + \mathbf{B}\mathbf{V}\mathbf{S}\mathbf{S}^T - 2\mathbf{B}\mathbf{V}\mathbf{S})_{ki}v^2_{ki}\\ &\quad +\gamma(\mathbf{V}\mathbf{M}^+ - \mathbf{V}\mathbf{M}^-)_{ki}v^2_{ki} + \lambda(\mathbf{C}\mathbf{V})_{ki}v^2_{ki} = 0,\end{aligned} \tag{31}$$

$$2\beta(\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S} - \mathbf{V}^T\mathbf{B}\mathbf{V})_{iq}s^2_{iq} + \beta\mu(\mathbf{E}^T)_{iq}s^2_{iq} = 0. \tag{32}$$

Solving Eqs. (30), (31), and (32), the multiplicative updates rules can be obtained as follows (33)–(35), as shown at the bottom of this page.

Hence, we provided the solver for the objective function (21).

C. CONVERGENCE ANALYSIS

In this subsection, we demonstrate the convergence of the objective function (22) to the local optimum using the updating rules (33), (34), and (35) after finite iterations. The convergence can be proved by the auxiliary function method [23]. First, the definition of the auxiliary function is introduced; subsequently, the convergence proof of the optimization algorithm is provided.

Definition 1: A function $Z(\mathbf{H}, \mathbf{H}')$ is an auxiliary function of the function $F(\mathbf{H})$ if $Z(\mathbf{H}, \mathbf{H}') \geq F(\mathbf{H})$ and $Z(\mathbf{H}, \mathbf{H}) = F(\mathbf{H})$ for any $\mathbf{H}$ and a constant matrix $\mathbf{H}'$.

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

$$\mathbf{H}^{(t+1)} = \operatorname*{argmin}_{\mathbf{H}} Z(\mathbf{H}, \mathbf{H}^{(t)}).$$

Proof:

$$F(\mathbf{H}^{(t+1)}) \leq Z(\mathbf{H}^{(t+1)}, \mathbf{H}^{(t)})) \leq Z(\mathbf{H}^{(t)}, \mathbf{H}^{(t)}) = F(\mathbf{H}^{(t)}).$$

Lemma 2: For any nonnegative matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times k}$, $\mathbf{S} \in \mathbb{R}^{n\times k}$, $\mathbf{S}' \in \mathbb{R}^{n\times k}$, and where $\mathbf{A}$, $\mathbf{B}$ are symmetric, the following inequality holds

$$\sum_{i=1}^{n}\sum_{s=1}^{k}\frac{(\mathbf{A}\mathbf{S}'\mathbf{B})_{is}\mathbf{S}^2_{is}}{\mathbf{S}'_{is}} \geq \operatorname{Tr}(\mathbf{S}^T\mathbf{A}\mathbf{S}\mathbf{B}).$$

We first introduce the auxiliary function for variable $\mathbf{V}$. It is proved that when $\mathbf{U}$ and $\mathbf{S}$ are fixed, the updating rule for $\mathbf{V}$ in (34) reduces the values of the objective function $\mathcal{O}$.

$$u^{(t+1)}_{jk} \leftarrow u^{(t)}_{jk}\sqrt{\frac{(\mathbf{X}^T\mathbf{A}^T\mathbf{X}\mathbf{V}^T)_{jk}}{(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V}\mathbf{V}^T)_{jk}}}, \tag{33}$$

$$v^{(t+1)}_{ki} \leftarrow v^{(t)}_{ki}\sqrt{\frac{(\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X} + 2\beta\mathbf{B}\mathbf{V}\mathbf{S} + \gamma\mathbf{V}\mathbf{M}^-)_{ki}}{(\mathbf{U}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}\mathbf{V} + \beta\mathbf{B}\mathbf{V} + \mathbf{B}\mathbf{V}\mathbf{S}\mathbf{S}^T + \gamma\mathbf{V}\mathbf{M}^+ + \lambda\mathbf{C}\mathbf{V})_{ki}}}, \tag{34}$$

$$s^{(t+1)}_{iq} \leftarrow s^{(t)}_{iq}\sqrt{\frac{2(\mathbf{V}^T\mathbf{B}\mathbf{V})_{iq}}{(2\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{S} + \mu\mathbf{E}^T)_{iq}}}. \tag{35}$$

We define $\mathcal{O}(\mathbf{V})$ to represent $\mathcal{O}$ that is relevant only to $\mathbf{V}$.

$$\begin{aligned}\mathcal{O}(\mathbf{V}) = &\operatorname{Tr}(\mathbf{V}^T\mathbf{U}^T\mathbf{X}^T\mathbf{AXUV}) - 2\operatorname{Tr}(\mathbf{X}^T\mathbf{AXUV})\\ &+\beta\operatorname{Tr}(\mathbf{VV}^T\mathbf{B}) + \beta\operatorname{Tr}(\mathbf{S}^T\mathbf{V}^T\mathbf{BVS})\\ &-2\beta\operatorname{Tr}(\mathbf{V}^T\mathbf{BVS}) + \gamma\operatorname{Tr}(\mathbf{VM}^+\mathbf{V}^T)\\ &-\gamma\operatorname{Tr}(\mathbf{VM}^-\mathbf{V}^T) + \lambda\operatorname{Tr}(\mathbf{V}^T\mathbf{CV}) \end{aligned}\quad (36)$$

Theorem 1: The following function

$$\begin{aligned}&Z(\mathbf{V},\mathbf{V}')\\ &= \sum_{ki}\frac{(\mathbf{U}^T\mathbf{X}^T\mathbf{AXUV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}} + \beta\sum_{ki}\frac{(\mathbf{BV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}\\ &\quad+\beta\sum_{ki}\frac{(\mathbf{BV}'\mathbf{SS}^T)_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}} + \gamma\sum_{ki}\frac{(\mathbf{V}'\mathbf{M}^+)_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}\\ &\quad+\beta\sum_{ki}\frac{(\mathbf{CV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}} - 2\sum_{ki}(\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X})_{ki}\mathbf{V}'_{ki}\left(1+\log\frac{\mathbf{V}_{ki}}{\mathbf{V}'_{ki}}\right)\\ &\quad-2\beta\sum_{kipq}\mathbf{B}_{kp}\mathbf{S}_{qi}\mathbf{V}'_{ki}\mathbf{V}'_{pq}\left(1+\log\frac{\mathbf{V}_{ki}\mathbf{V}_{pq}}{\mathbf{V}'_{ki}\mathbf{V}'_{pq}}\right)\\ &\quad-\gamma\sum_{kiq}\mathbf{M}_{iq}^-\mathbf{V}'_{ki}\mathbf{V}'_{kq}\left(1+\log\frac{\mathbf{V}_{ki}\mathbf{V}_{kq}}{\mathbf{V}'_{ki}\mathbf{V}'_{kq}}\right)\end{aligned}\quad (37)$$

is an auxiliary function for $\mathcal{O}(\mathbf{V})$.

Proof: On the one hand, $Z(\mathbf{V},\mathbf{V}) = \mathcal{O}(\mathbf{V})$ is obvious. On the other hand, we must prove that $Z(\mathbf{V},\mathbf{V}') \geq \mathcal{O}(\mathbf{V})$. Hence, we compare Eq.(36) with Eq.(37) to obtain that $Z(\mathbf{V},\mathbf{V}') \geq \mathcal{O}(\mathbf{V})$.

By applying Lemma 3, we obtain the upper bounds of every five positive terms, and the following inequality is established:

$$\operatorname{Tr}(\mathbf{V}^T\mathbf{U}^T\mathbf{X}^T\mathbf{AXUV}) \leq \sum_{ki}\frac{(\mathbf{U}^T\mathbf{X}^T\mathbf{AXUV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}, \quad (38)$$

$$\beta\operatorname{Tr}(\mathbf{V}^T\mathbf{BV}) \leq \beta\sum_{ki}\frac{(\mathbf{BV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}. \quad (39)$$

$$\beta\operatorname{Tr}(\mathbf{S}^T\mathbf{V}^T\mathbf{BVS}) \leq \beta\sum_{ki}\frac{(\mathbf{BV}'\mathbf{SS}^T)_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}. \quad (40)$$

$$\gamma\operatorname{Tr}(\mathbf{VM}^+\mathbf{V}^T) \leq \gamma\sum_{ki}\frac{(\mathbf{V}'\mathbf{M}^+)_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}. \quad (41)$$

$$\lambda\operatorname{Tr}(\mathbf{V}^T\mathbf{CV}) \leq \lambda\sum_{ki}\frac{(\mathbf{CV}')_{ki}\mathbf{V}_{ki}^2}{\mathbf{V}'_{ki}}. \quad (42)$$

To obtain the lower bounds of the residual terms, we use inequality $z \geq 1 + logz, \forall z$; subsequently,

$$2\operatorname{Tr}(\mathbf{X}^T\mathbf{AXUV}) \geq 2\sum_{ki}(\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X})_{ki}\mathbf{V}'_{ki}\left(1+\log\frac{\mathbf{V}_{ki}}{\mathbf{V}'_{ki}}\right), \quad (43)$$

$$\begin{aligned}2\beta\operatorname{Tr}(\mathbf{V}^T\mathbf{BVS}) &= 2\beta\sum_{kipq}\mathbf{V}_{ki}\mathbf{B}_{kp}\mathbf{V}_{pq}\mathbf{S}_{qi}\\ &\geq 2\beta\sum_{kipq}\mathbf{B}_{kp}\mathbf{S}_{qi}\mathbf{V}'_{ki}\mathbf{V}'_{pq}\left(1+\log\frac{\mathbf{V}_{ki}\mathbf{V}_{pq}}{\mathbf{V}'_{ki}\mathbf{V}'_{pq}}\right)\end{aligned}\quad (44)$$

$$\gamma\operatorname{Tr}(\mathbf{VM}^-\mathbf{V}^T) \geq \gamma\sum_{kiq}\mathbf{M}_{iq}^-\mathbf{V}'_{ki}\mathbf{V}'_{kq}\left(1+\log\frac{\mathbf{V}_{ki}\mathbf{V}_{kq}}{\mathbf{V}'_{ki}\mathbf{V}'_{kq}}\right). \quad (45)$$

Summing all inequalities, we can obtain $Z(\mathbf{V},\mathbf{V}') \geq \mathcal{O}(\mathbf{V})$, satisfying $Z(\mathbf{V},\mathbf{V}') \geq \mathcal{O}(\mathbf{V})$. Therefore, $Z(\mathbf{V},\mathbf{V}')$ is an auxiliary function of $\mathcal{O}(\mathbf{V})$.

Theorem 2: The updating rule (34) can be obtained by minimizing the auxiliary function $Z(\mathbf{V},\mathbf{V}')$.

Proof: To obtain the minimum of $Z(\mathbf{V},\mathbf{V}')$, we set the derivative $\frac{\partial Z(\mathbf{V},\mathbf{V}')}{\partial \mathbf{V}_{ki}} = 0$, and have

$$\begin{aligned}\frac{\partial Z(\mathbf{V},\mathbf{V}')}{\partial \mathbf{V}_{ki}} = &\frac{2(\mathbf{U}^T\mathbf{X}^T\mathbf{AXUV}')_{ki}\mathbf{V}_{ki}}{\mathbf{V}'_{ki}} + 2\beta\frac{(\mathbf{BV}')_{ki}\mathbf{V}_{ki}}{\mathbf{V}'_{ki}}\\ &+2\beta\frac{(\mathbf{BV}'\mathbf{SS}^T)_{ki}\mathbf{V}_{ki}}{\mathbf{V}'_{ki}} + 2\gamma\frac{(\mathbf{V}'\mathbf{M}^+)_{ki}\mathbf{V}_{ki}}{\mathbf{V}'_{ki}}\\ &+2\lambda\frac{(\mathbf{CV}')_{ki}\mathbf{V}_{ki}}{\mathbf{V}'_{ki}} - \frac{2(\mathbf{U}^T\mathbf{X}^T\mathbf{A}^T\mathbf{X})_{ki}\mathbf{V}'_{ki}}{\mathbf{V}_{ki}}\\ &-4\beta\frac{(\mathbf{BV}'\mathbf{S})_{ki}\mathbf{V}'_{ki}}{\mathbf{V}_{ki}} - \gamma\frac{2(\mathbf{VM}^-)_{ki}\mathbf{V}'_{ki}}{\mathbf{V}_{ki}}.\end{aligned}\quad (46)$$

Thus, by simple algebra formulation, we can obtain the iterative updating rule for $\mathbf{V}$ as Eq.(34). According to Lemma 1, the objective function (22) decreases monotonically with the updating of v_{ki}.

The convergence proves that updating u_{jk} and s_{iq} using (33) and (35) are similar to the above.

D. COMPLEXITY ANALYSIS OF RLLDCF

In this subsection, we discuss the computational complexity of our proposed algorithm, and use the capital O notation to express the complexity. The computation cost of the pairwise distances between the samples is $O(N^2M)$; the computation cost of finding k-nearest neighbors of all sample points is $O(N\log N)$; $O(N(Mk^2+k^3))$ is the computation cost of Eq.(14). The computation cost of the multiplicative updating in (33), (34), and (35) is $O(NK^2)$. If the updating procedure stops after t iterations, the overall cost of the multiplicative updating is $O(tNK^2)$. Because $M >> K$ and $M >> k$, the total cost of RLLDCF is $O(N^2M + N\log N + N(Mk^2 + k^3) + tNK^2)$.

V. EXPERIMENTS

We evaluated the clustering performance of our presented RLLDCF algorithm systematically with the popular matrix factorization algorithms including robust nonnegative matrix factorization using $\ell_{2,1}$-norm (RNMF) [35], CF [15], LCF [41], robust semi-supervised concept factorization (RSCF) [34], and concept factorization with adaptive neighbors (CFANs) [32]. We set the number of clusters to be the same as the true number of categories. We varied α and μ in the range of $\{1, 10, 20, 30, 40, 50, 60, 70, 80\}$. The regularization parameters β and γ were searched over the grid $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4\}$. The proposed RLLDCF algorithm was tuned to achieve the best

TABLE 1. Performance comparison with illumination variations.

k	Acc(%)					
	RNMF	CF	LCF	RSCF	CFANs	RLLDCF
3	90.96 ± 13.23	78.85 ± 15.54	90.29 ± 11.18	76.82 ± 12.39	93.22 ± 12.37	95.83 ± 10.93
5	85.73 ± 11.43	70.07 ± 10.36	83.33 ± 8.14	69.74 ± 7.35	88.39 ± 9.22	93.55 ± 8.46
7	82.46 ± 6.61	63.91 ± 6.48	78.09 ± 6.46	68.13 ± 5.71	86.42 ± 5.78	91.31 ± 7.25
9	83.87 ± 4.27	58.45 ± 3.72	76.42 ± 4.16	65.66 ± 4.38	83.56 ± 4.52	88.43 ± 5.37
11	77.26 ± 3.01	55.95 ± 3.26	74.41 ± 3.63	64.22 ± 4.33	76.23 ± 3.19	86.72 ± 4.54
13	73.25 ± 2.81	53.44 ± 3.11	72.15 ± 3.37	63.61 ± 3.84	74.96 ± 4.53	85.75 ± 3.83
15	72.48 ± 2.47	52.50 ± 3.42	69.52 ± 3.74	62.58 ± 2.08	72.73 ± 2.39	84.18 ± 3.38
17	68.67 ± 2.17	52.07 ± 2.97	68.83 ± 3.12	60.89 ± 2.22	70.02 ± 2.54	82.83 ± 2.44
19	67.66 ± 2.01	52.60 ± 2.24	66.53 ± 3.37	59.40 ± 2.93	68.62 ± 2.37	80.26 ± 2.34
Avg	78.04	59.76	75.71	65.67	79.35	87.65
k	NMI(%)					
	RNMF	CF	LCF	RSCF	CFANs	RLLDCF
3	92.97 ± 15.72	85.78 ± 16.23	91.71 ± 13.03	80.81 ± 14.54	95.11 ± 15.13	98.21 ± 12.28
5	88.08 ± 13.01	73.35 ± 12.43	86.92 ± 9.76	74.11 ± 9.36	92.18 ± 12.82	97.19 ± 10.74
7	87.23 ± 8.55	65.45 ± 8.63	84.09 ± 7.94	73.40 ± 7.75	90.78 ± 8.27	96.65 ± 9.24
9	88.44 ± 6.46	63.01 ± 4.81	83.72 ± 6.91	72.88 ± 6.41	88.42 ± 7.78	96.39 ± 6.33
11	85.62 ± 4.97	62.08 ± 4.17	81.47 ± 5.79	71.75 ± 6.25	86.07 ± 5.56	95.59 ± 5.87
13	83.45 ± 4.53	61.50 ± 4.15	80.44 ± 4.36	70.12 ± 4.34	84.99 ± 6.21	94.96 ± 4.99
15	81.89 ± 3.26	59.19 ± 3.24	78.90 ± 4.45	69.48 ± 3.06	83.93 ± 3.44	94.16 ± 4.77
17	80.68 ± 3.17	57.99 ± 3.11	77.87 ± 4.12	67.14 ± 3.42	81.11 ± 3.82	92.32 ± 4.89
19	77.42 ± 3.02	56.26 ± 2.97	77.70 ± 4.57	65.42 ± 3.97	80.32 ± 3.23	90.85 ± 4.21
Avg	85.09	64.96	82.31	71.68	86.99	95.37

performance by selecting the best parameters, and the comparison methods' parameter settings followed the authors' suggestions. In the experiment, we employed two standard face datasets to evaluate the clustering performance of the proposed method. **The extended YaleB dataset** is composed of 38 subjects with 2414 frontal-face images. In this dataset, the size of each face image is 192 × 168, and each subject is acquired with 64 illuminate conditions and 9 different poses. **The AR dataset** contains over 4,000 frontal-face images of 126 subjects (70 men and 56 women) with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). There are 26 images per subject in two separated sessions.

A. EVALUATION METRICS

In this work, we used two standard clustering metrics to measure the clustering performance, i.e., the accuracy (Acc) and the normalized mutual information (NMI). We evaluated the algorithms by comparing the obtained cluster labels of each data point with its ground truth label provided by the dataset. The Acc metric is defined as follows:

$$\mathrm{Acc} = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{47}$$

where n is the total number of images; r_i and l_i denote the cluster label and the label by dataset, respectively. $\delta(x, y)$ is the delta function that equals one if $x = y$, and equals to zero otherwise. Further, $map(r_i)$ is the mapping function that maps the obtained label r_i to the equivalent label from the dataset. The best mapping function can be determined using the Kuhn–Munkres algorithm [50]. The second metric is the NMI that could measure the similarity between the clustering results and the true classes. Assume that $\mathcal{C}$ and $\mathcal{C}'$ are the sets of clustering result and the true class, respectively. The NMI is defined as follows:

$$\mathrm{NMI}(\mathcal{C}, \mathcal{C}') = \frac{\mathrm{MI}(\mathcal{C}, \mathcal{C}')}{\max(H(\mathcal{C}), H(\mathcal{C}'))}, \tag{48}$$

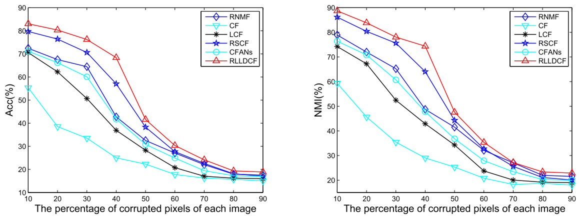
where $H(\mathcal{C})$ and $H(\mathcal{C}')$ are the entropies of the cluster sets $\mathcal{C}$ and $\mathcal{C}'$, respectively. It is noteworthy that $\mathrm{NMI}(\mathcal{C}, \mathcal{C}')$ ranges from 0 to 1. NMI $=$ 1 if the two sets of clusters are identical, and NMI $=$ 0 if the two sets are independent.

B. CLUSTERING WITH ILLUMINATION VARIATIONS

In this experiment, we tested the impact of illumination changes on the clustering performance in the extended YaleB dataset. More than half of the face images are damaged by a large area of "shadow." Therefore, the dataset can be regarded as a serious damage. To reduce the cost of computing, we resized each image to 32×32 pixels and rearranged it to a 1,024-dimensional vector. Each vector was normalized to a unit length. We established the data matrix $\mathbf{X}$, which is composed of various numbers of subjects $k \in \{3, 5, 7, \cdots, 19\}$ in the extended YaleB dataset. k classes were randomly selected from the dataset. All of these data were mixed as matrix $\mathbf{X}$ for clustering. We set the clustering number k equal to the dimension of the new representation space. For a fair comparison, we used K-means to cluster the samples based on the new data representations. Because the results of K-means are related to the initialization, we conducted 20 trials with different initializations, and the means and variances of Acc and NMI were reported as the final result. Table 1 shows the detailed clustering results on different clustering numbers. The final row shows the average clustering accuracy (normalized mutual information) over k. Compared with the second optimal method, the accuracy of clustering using our method is increased by 8.30%. For mutual information, it exhibits an 8.38% improvement over the second optimal algorithms.

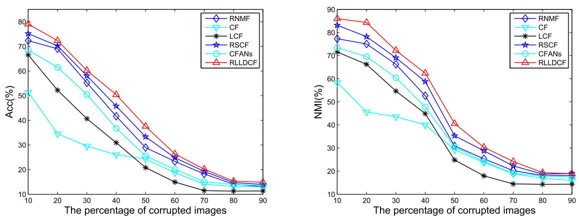
C. CLUSTERING WITH SYNTHETIC CORRUPTIONS

We investigated the image to be corrupted by white Gaussian noise and random noise using the extended YaleB dataset, where the former is additive and the latter is nonadditive. For the white Gaussian noise case, each image $\mathbf{x}$ was corrupted by adding the pixel value from a standard normal distribution, that is, $\tilde{\mathbf{x}} = \mathbf{x} + \alpha\mathbf{n}$, where α is the corruption ratio from 10% to 90% with an interval of 10%, and $\mathbf{n}$ is the noise. For the random pixel corruption, we replaced the values of the pixel with a uniform distribution over $[0, 255]$. We varied the percentage



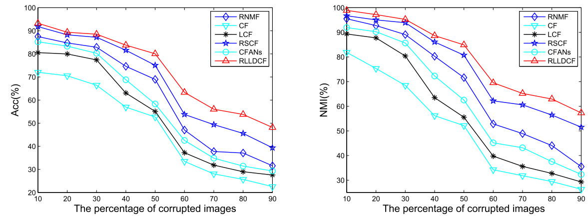
(a) (b)

FIGURE 1. Performance comparison curves with varying percentage of the white Gaussian noise corruption on the extended YaleB dataset. (a) The Acc results. (b) The NMI results.



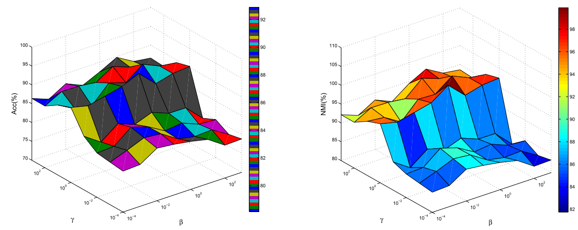
(a) (b)

FIGURE 2. Performance comparison curves with varying percentage of the random noise corruption on the extended YaleB dataset. (a) The Acc results. (b) The NMI results.



(a) (b)

FIGURE 3. Clustering Acc and NMI curves of the compared methods on varying percent block occlusion for the extended YaleB dataset. (a) The Acc results. (b) The NMI results.



(a) (b)

FIGURE 4. Clustering accuracy of the proposed method with respect to the parameters β and γ on the extended YaleB dataset. (a) The Acc results. (b) The NMI results.

of corruption from 10% to 90%. The location of the corrupted pixels was chosen randomly for each image. We conducted the evaluations 20 times with different corruption percentages and computed the average recognition accuracies of Acc and NMI. Figures 1 and 2 display the recognition accuracies over different levels of corruption. The recognition accuracies of these methods decrease rapidly as the level of corruption increases. From Figures 1 and 2 that depict the recognition accuracies, we can observe that our approach is better than the other methods. All compared algorithms perform worse with the increase in noise and obtain better performance in white noise corruption than random pixel corruption.

D. CLUSTERING WITH SYNTHETIC OCCLUSION

We used the extended YaleB dataset to test the robustness of our algorithm to synthetic occlusion. We added various levels of contiguous occlusions in each image using an unrelated image of size $p \times p$ with $p \in \{10, 20, \cdots, 90\}$. The evaluations were conducted 20 times at each occlusion level, and the average Acc and NMI curves were reported. Figure 3 plots the clustering of Acc and NMI results of the compared methods under different occlusion levels. Although the clustering accuracy of each method degrades with the increment in the occlusion level, RLLDCF consistently exceeds other methods. When the occlusion size increases to 50×50, the occluding part will dominate the image and decrease the clustering performance rapidly.

E. CLUSTERING WITH REAL OCCLUSIONS

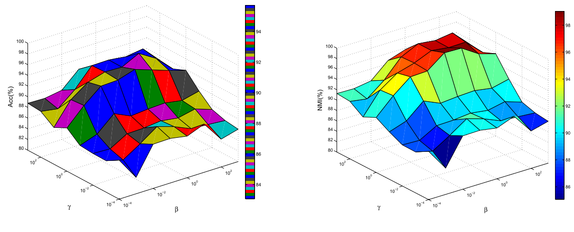
We tested the robustness to real malicious occlusions of the investigated approaches over the AR dataset. The AR dataset has 2600 frontal-face images from 100 individuals comprising 50 males and 50 females, captured in two separate sessions. In each session, each subject provides seven face images with different facial variations, three face images with an occlusion of sunglasses, and three face images with a scarf occlusion. We performed two experiments that correspond to the sunglasses and scarf occlusions for each session, respectively. For the sunglasses occlusion, we selected all the normal face images and three face images with sunglasses. For the scarf occlusion, we selected all the normal face images and three face images with scarf. All images were cropped to 32×32-pixel gray-scale images and shaped into a 1024-dimensional vector. Table 2 tabulates the comparison results on the AR dataset. It is noteworthy that the performance gain of the proposed algorithm is obvious in comparison with the other five algorithms.

F. PARAMETER SENSITIVITY

Our proposed RLLDCF required four parameters: α, β, γ, and μ to be tuned in advance. First, we focused on discussing the parameters β and γ. We conducted experiments on extended YaleB and the first session of AR datasets to observe the effects on clustering performance with different values of β and γ. Two parameters were tuned from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$. We plotted the Acc and NMI of RLLDCF with respect to β and γ, by setting $\alpha = 10$ and $\mu = 10$. Figures 4 and 5 show the 3D results of RLLDCF clearly. The horizontal axes are the parameters β and γ, whereas the vertical axis represents the clustering accuracy of RLLDCF. As shown in Figures 4 and 5, the clustering performance of RLLDCF is

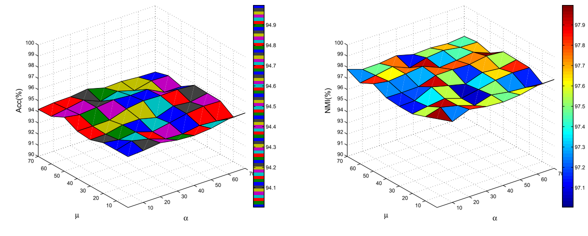
TABLE 2. Clustering performance on the AR dataset.

Methods	Acc(%)				NMI(%)			
	Session 1		Session 2		Session 1		Session 2	
	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
RNMF	82.39 ± 3.83	80.81 ± 3.48	84.08 ± 4.31	81.24 ± 2.51	84.93 ± 3.83	82.21 ± 3.06	87.32 ± 3.03	85.69 ± 3.07
CF	51.92 ± 3.48	50.87 ± 3.52	52.42 ± 3.37	52.08 ± 3.27	61.88 ± 2.71	60.16 ± 3.77	65.92 ± 3.41	64.35 ± 3.73
LCF	74.68 ± 3.29	69.69 ± 2.55	75.57 ± 3.09	71.32 ± 3.94	84.36 ± 3.82	79.77 ± 2.92	84.72 ± 2.24	81.31 ± 3.63
RSCF	77.29 ± 3.75	72.81 ± 2.73	78.74 ± 2.93	80.26 ± 2.94	86.37 ± 2.63	82.84 ± 2.42	86.73 ± 2.64	89.58 ± 2.85
CFANs	79.39 ± 2.74	75.61 ± 3.63	84.58 ± 2.63	80.64 ± 2.92	83.76 ± 3.64	81.84 ± 3.97	88.59 ± 2.69	85.95 ± 2.76
RLLDCF	89.94 ± 2.67	83.79 ± 3.49	90.24 ± 2.35	85.78 ± 2.15	95.66 ± 3.86	92.49 ± 4.19	96.74 ± 2.05	93.34 ± 2.45



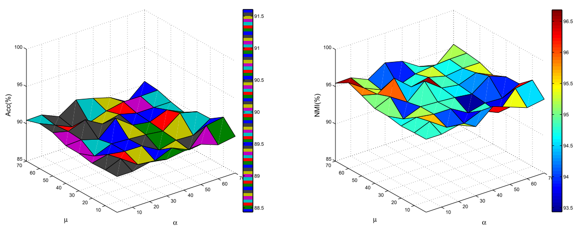
(a) (b)

FIGURE 5. Clustering accuracy of the proposed method with respect to the parameters β and γ on the AR dataset. (a) The Acc results. (b) The NMI results.



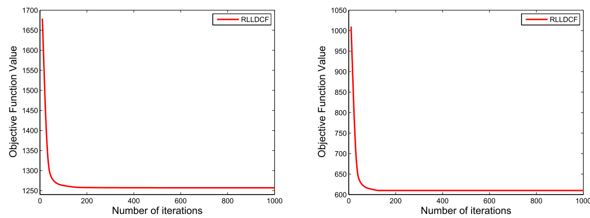
(a) (b)

FIGURE 7. Clustering accuracy of the proposed method with respect to the parameters α and μ on the AR dataset. (a) The Acc results. (b) The NMI results.



(a) (b)

FIGURE 6. Clustering accuracy of the proposed method with respect to the parameters α and μ on the extended YaleB dataset. (a) The Acc results. (b) The NMI results.



(a) (b)

FIGURE 8. Convergence curves of RLLDCF. (a) The Extended YaleB result. (b) The AR result.

relatively insensitive when parameters β and γ are selected at an appropriate range. This renders RLLDCF easy to apply without much effort for parameter tuning.

Next, the parameter sensitivities of α and μ were tested. We selected extended YaleB and the first session of the AR datasets as test examples. We analyzed the sensitivity of the parameters by Acc and NMI under different values of α and μ. We set $\beta = 100$ and $\gamma = 100$, and varied α and μ in the range of $\{1, 10, 20, 30, 40, 50, 60, 70, 80\}$. We depicted the clustering performances with different values of α and μ, and the results are shown in Figures 6 and 7. As shown, the clustering performance tends to change little when the values of α and μ are different, thus indicating that RLLDCF is relatively insensitive to the selection of parameters α and μ.

G. CONVERGENCE ANALYSIS

We proved the convergence of our update rules in the previous section. An experiment was performed to validate its convergence, and the speed of convergence on the extended YaleB and AR datasets was studied. The convergence curves of the objective value are shown in Figure 8. The horizontal axis represents the number of iterations, and the vertical axis represents the value of the objective function. We observe that the proposed optimization method for RLLDCF converges within 50 iterations for two datasets, demonstrating that our algorithm is effective and converges quickly.

VI. CONCLUSION

We herein proposed a novel learning method named RLLDCF that extended robust CF by exploring the local discriminative and intrinsic structure explicitly. We formulated the problem into a unified optimization framework and designed an efficient optimization algorithm. The optimization algorithm and an analysis of its convergence were also presented. Further work is required to extend RLLDCF to multiview settings.

ACKNOWLEDGMENT

The authors would like to thank all anonymous reviewers for their helpful comments.

REFERENCES

[1] Z. Yang, Y. Xiang, K. Xie, and Y. Lai, "Adaptive method for nonsmooth nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 948–960, Apr. 2017.

[2] H. Qi, K. Li, Y. Shen, and W. Qu, "Object-based image retrieval with kernel on adjacency matrix and local combined features," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 4, Nov. 2012, Art. no. 54.

[3] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2018.

[4] J. Wei, L. Min, and Z. Yongqing, "Neighborhood preserving convex nonnegative matrix factorization," *Math. Problems Eng.*, vol. 2014, no. 2, pp. 1–8, Jan. 2014.

[5] J. Ye and Z. Jin, "Graph-regularized local coordinate concept factorization for image representation," *Neural Process. Lett.*, vol. 46, no. 2, pp. 427–449, Feb. 2017.

[6] J. Wang, F. Tian, C. H. Liu, H. Yu, X. Wang, and X. Tang, "Robust nonnegative matrix factorization with ordered structure constraints," in *Proc. Int. Joint Conf. Neural Netw.*, Anchorage, AK, USA, May 2017, pp. 478–485.

[7] Y. Cao *et al.*, "Binary hashing for approximate nearest neighbor search on big data: A survey," *IEEE Access*, vol. 6, pp. 2039–2054, 2018.

[8] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, "Multilabel prediction via cross-view search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4324–4338, Sep. 2018.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[10] P. N. Belhumeur, J. P. 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.

[11] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 284–295, Jan. 2018.

[12] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, and Q. Dai, "Effective uyghur language text detection in complex background images for traffic prompt identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 220–229, Jan. 2018.

[13] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2042–2054, Apr. 2017.

[14] W. Zhang, S. Hu, K. Liu, and J. Yao, "Motion-free exposure fusion based on inter-consistency and intra-consistency," *Inf. Sci.*, vol. 376, pp. 190–201, Jan. 2016.

[15] W. Xu and Y. Gong, "Document clustering by concept factorization," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Sheffield, U.K., Jul. 2004, pp. 202–209.

[16] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[19] X. Li, X. Shen, Z. Shu, Q. Ye, and C. Zhao, "Graph regularized multilayer concept factorization for data representation," *Neurocomputing*, vol. 238, pp. 139–151, May 2017.

[20] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, and H. T. Shen, "Semi-Paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4275–4288, Dec. 2017.

[21] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 276–288, Jan. 2017.

[22] D. Cai, X. He, 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.

[23] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 902–913, Jun. 2011.

[24] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[25] W. Zhang, C. Qu, L. Ma, J. Guan, and R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognit.*, vol. 59, pp. 176–187, Nov. 2016.

[26] W. Zhang, K. Liu, W. Zhang, Y. Zhang, and J. Gu, "Deep neural networks for wireless localization in indoor and outdoor environments," *Neurocomputing*, vol. 194, pp. 279–287, Jun. 2016.

[27] Q. Gu, C. Ding, and J. Han, "On trivial solution and scale transfer problems in graph regularized NMF," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, Jul. 2011, pp. 16–22.

[28] J. Ye and Z. Jin, "Feature selection for adaptive dual-graph regularized concept factorization for data representation," *Neural Process. Lett.*, vol. 45, no. 2, pp. 667–688, Jun. 2017.

[29] F. Shang, L. C. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognit.*, vol. 45, no. 6, pp. 2237–2250, 2012.

[30] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, pp. 977–986.

[31] X. Guo, "Robust subspace segmentation by simultaneously learning data representations and their affinity matrix," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 3547–3553.

[32] X. Pei, C. Chen, and W. Gong, "Concept factorization with adaptive neighbors for document clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 343–352, Feb. 2018.

[33] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csiszár's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.* Berlin, Germany: Springer, 2006, pp. 32–39.

[34] W. Yan, B. Zhang, and S. Ma, "Robust semi-supervised concept factorization," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 1011–1017.

[35] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using ℓ_2, 1-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl.*, Glasgow, U.K., 2011, pp. 24–28.

[36] H. Zhang, Z.-J. Zha, S. Yan, M. Wang, and T.-S. Chua, "Robust non-negative graph embedding: Towards noisy data, unreliable graphs, and noisy labels," in *Proc. CVPR*, Jun. 2012, pp. 2464–2471.

[37] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[38] W. Zhang, W. Zhang, K. Liu, and J. Gu, "A feature descriptor based on local normalized difference for real-world texture classification," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 880–888, Apr. 2018.

[39] W. Jiang, J. Liu, H. Qi, and Q. Dai, "Robust subspace segmentation via nonconvex low rank representation," *Inf. Sci.*, vols. 340–341, pp. 144–158, May 2016.

[40] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 969–979, Mar. 2013.

[41] H. Liu, Z. Yang, and Z. Wu, "Locality-constrained concept factorization," in *Proc. Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, Jul. 2011, pp. 1378–1383.

[42] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2223–2231.

[43] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

[44] C.-P. Wei, Y.-W. Chao, Y.-R. Yeh, and Y.-C. F. Wang, "Locality-sensitive dictionary learning for sparse representation based classification," *Pattern Recognit.*, vol. 46, no. 5, pp. 1277–1287, May 2013.

[45] Y.-W. Chao, Y.-R. Yeh, Y.-W. Chen, Y.-J. Lee, and Y.-C. F. Wang, "Locality-constrained group sparse representation for robust face recognition," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 761–764.

[46] W. Zhang and W.-K. Cham, "Gradient-directed composition of multi-exposure images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 530–536.

[47] C. Yan *et al.*, "Efficient parallel framework for HEVC motion estimation on many-core processors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2077–2089, Dec. 2014.

[48] C. Yan *et al.*, "A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 573–576, May 2014.

[49] L. Zhang, "Locally regressive projections," *Int. J. Softw. Inform.*, vol. 7, no. 3, pp. 435–451, 2013.

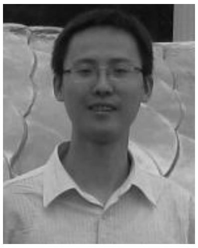
[50] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, 2009, 175–184.

[51] L. Zhang, C. Chen, J. Bu, Z. Chen, D. Cai, and J. Han, "Locally discriminative coclustering," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1025–1035, Jun. 2012.

WEI JIANG received the B.S. and Ph.D. degrees from the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, in 2004 and 2011, respectively. He became a Faculty Member with the School of Mathematics, Liaoning Normal University, Dalian, China, in 2004, where he is currently a Professor. His research interests include computer vision and machine learning.

XIAOTING FENG received the B.S. degree from Eastern Liaoning University, Shenyang, China. She is currently pursuing the Ph.D. degree with the School of Mathematics, Liaoning Normal University, Dalian, China. Her research interests include computer vision and machine learning.

LING XING received the Ph.D. degree in communication and information system from the Beijing Institute of Technology in 2008. She is currently a Professor with the School of Information Engineering, Henan University of Science and Technology, China. Her research interests include information intelligent management, multimedia semantic concept discovery, and big data mining.



KEWEI TANG received the B.S. degree in mathematics from Liaoning Normal University in 2008 and the M.Sc. and Ph.D. degrees in mathematics from the Dalian University of Technology, Dalian, China, in 2011 and 2015, respectively. He was a Visiting Scholar with the Department of Statistical Science, Duke University, from 2013 to 2014. He is currently a Lecturer with the School of Mathematics, Liaoning Normal University. His research interests include computer vision and machine learning.