IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 20, 2020, accepted March 7, 2020, date of publication March 12, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980394

Exploiting Linear Manifold Features With Parts-Based Representation in Various Scenes

QIAOQIN LI, YONGGUO LIU, AND SHANGMING YANG
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Corresponding author: Shangming Yang (minn003@163.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61572107, in part by the National Key Research and Development Program of China under Grant 2017YFC1703905, in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant 2018ZX10715003, and in part by the Sichuan Science and Technology Program under Grant 2018SZ0065.

ABSTRACT Image recognition in complex scenes is a big challenge in computer vision. Manifold learning has become one of the most popular tools in the application of data dimensionality reduction and image recognition due to its efficiency in retrieving the intrinsic geometric features of image data. In this paper, we propose a new manifold feature extracting model based on the nonnegative matrix factorization (NMF) for image clustering in various scenes. In this model, Pearson distance with multiple manifold regulation constraints are adopted as the objective function to derive NMF based learning algorithms for the feature capturing of high dimensional data. With a variable neighborhood size in the learning, the proposed model can learn the linear features and at the same time learn the local similarity of images in multi-scale neighborhoods of a graph space. For different settings of learning parameters λ_{lx} and λ_{sx}, tests show that the proposed algorithms can efficiently retrieve low dimensional structures of images. Test results on four different image datasets demonstrate that the algorithms can achieve the state of art performance on the clustering of images in different types of scene.

INDEX TERMS NMF, local representation, manifold regularization, multi-scale learning, feature extracting.

I. INTRODUCTION

Neural networks-based feature extracting for images has been widely applied in pattern recognitions. However, due to the complexity of different scenes of sample data, previous networks often have poor solutions [1], [2]. The performance of algorithms was improved by constructing multi-layer or graph neural networks-based learning models [3]–[5]. Dimension reduction of data is an essential step for feature extraction. By exploiting different models and their compounding setting of parameters, a wide range of algorithms were proposed to obtain low dimensional data features. Nonnegative matrix factorization is one of the most popular data dimensional reduction methods for parts-based feature representation [6], [7]. From NMF, many interesting algorithms have been developed for image clustering and classification [8]–[11], including graph regularized NMF algorithms [26], [27] and deep neural network based NMF algorithms, which focus on the extraction of low dimensional features with intrinsic geometric structures in sample data [29]–[32]. For these algorithms, investigations have shown that unsupervised multi-layer and graph regularization techniques can be utilized for image recognitions in various complex scenes, where the introducing of manifold approaches were a significant improvement to the efficiency of the algorithms. In manifold learning, traditional algorithms include Laplacian Eigenmaps algorithm (LE) [17], Locally Linear Embedding (LLE) [18], and Isometric Feature Mapping (ISOMAP) [19], from which many important learning methods, such as Hessian-based locally linear embedding (HLLE) [20], the log Riemannian exponential map expressed in tangent space algorithm (LOGMAP) [21], and other manifold regularized algorithms [22]–[25] have been derived. All these algorithms were motivated by the idea of similarity embedding of graph nodes in a neighborhood. Meanwhile, by constructing similarity graph or similarity matrix for different views of sample data, multi-view learning has been proposed recently to improve the existing manifold learning algorithms, which is becoming a more important type of methods to recognize complex scene images [40]–[43].

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

In general, these models show that incorporating manifold regularization with different constraints can obtain more efficient representation of sample data.

However, test results show that for the LLE, the size of neighborhood determines the topological structures of sample data. If a neighborhood is too small or too big, the manifold features of data cannot be described by the locally linear embedding. The manifold structures of data in a small neighborhood can only be described by the similarity or invariance of nearest neighbors, that is, in this neighborhood, any two nodes will be similar but these nodes may not have the linear combination relation. On the other hand, when using manifold learning for classification or clustering, most current graph regularized NMF algorithms focused on learning the local invariance in a small neighborhood, which cannot completely obtain the intrinsic structures of objects to handle the complexity of images in different scenes. In this paper, we propose a multiscale local manifold constrained NMF (LMNMF) algorithm, which can learn both locally linear representation and local invariance of images in different neighborhood scales to capture the low dimensional geometric architectures of sample data. The main contribution of this paper is summarized as follows:

1) A novel model called multi-scale local manifold regularized NMF is proposed. In this model, NMF based manifold learning algorithms for both locally linear representation and local invariance of image data with different scales of neighborhoods are developed to extract the low dimensional geometric architectures of images in different scenes.

2) The convergent properties of the proposed algorithms are exploited, which show that the best image clustering result is obtained only if in the learning, the objective function is non-increase and its corresponding learning algorithms stably converge. By adjusting the setting of parameters, the convergence of the objective function and the learning algorithms can be controlled efficiently.

3) Experimental results on four image datasets are presented to show the efficiency of the proposed algorithms, which demonstrate that the proposed method can obtain the best or close to the best performance than several other feature extracting algorithms in terms of accuracy (ACC) and normalized mutual information (NMI).

Fig. 1 shows the basic structure of the proposed model. In this model, NMF learning is on the whole domain to obtain the parts-based representation, but locally linear embedding and local invariance feature learning are in different sizes of neighborhoods to obtain their corresponding feature representations. Feature representations are approximated by the linear combination of different basis images. Thus, a multiscale based manifold learning model is constructed.

The rest of this paper is organized as follows: In Section II, the related algorithms are presented. In section III, the Pearson distance based objective function is introduced. In Section IV, the framework of learning algorithms is developed from the proposed objective function. In Section V, the complexity of the proposed algorithms is analyzed. In section VI, experimental results are presented. Finally, in Section VII, the conclusions are provided.

II. RELATED ALGORITHMS

NMF is to decompose an $M \times N$ data matrix **Y** into two non-negative matrices **A** and **X** such that the product of **A** and **X** ($\boldsymbol{Y} = \boldsymbol{AX}$) can correctly approximate the original data matrix **Y**, where $\mathbf{A} \in \mathbf{R}^{M\times K}$ is called the basis matrix and $\mathbf{X} \in \mathbf{R}^{K\times N}$ is called the encoding or representation matrix of the original data. Assume that a_{ij} is the element of matrix **A**, $\boldsymbol{a}_i$ is a column vector of **A**, x_{jk} is an element of matrix **X**, and $\boldsymbol{x}_j$ is a column vector of **X**. Frobenius norm is one of the most popular and relatively simple objective functions to obtain the error of decomposition, which was as follows:

$$Div_F\,(\boldsymbol{Y}, \boldsymbol{AX}) = \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{AX}||^2 \tag{1}$$

From the objective function in Eq. (1), Lee and Seung developed the following learning rules to decompose the data matrix into factors **A** and **X** [6].

$$a_{ij} \leftarrow a_{ij}\frac{[\boldsymbol{YX^T}]_{ij}}{[\boldsymbol{AXX^T}]_{\boldsymbol{ij}}}, \quad x_{jk} \leftarrow x_{jk}\frac{[\boldsymbol{A^T Y}]_{jk}}{[\boldsymbol{A^T AX}]_{\boldsymbol{ij}}} \tag{2}$$

where $a_{ij}(a_{ij} \geq 0, i = 1, 2, \ldots, M, j = 1, 2, \ldots, K)$ are the elements of matrix **A** and x_{jk} (x_{jk} $\geq 0, j = 1, 2, \ldots, K, k = 1, 2, \ldots, N$) are the elements of matrix **X**.

By imposing extra term to the Frobenius norm in Eq. (1), graph regularized NMF was developed by Cai *et al.* [30],

$$\begin{aligned} Div_{F+}\,(\boldsymbol{Y}, \boldsymbol{AX}) &= \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{AX}||^2 \\ &\quad+\frac{\lambda}{2}\sum_{ij}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\bar{w}_{ij} \\ &= \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{AX}||^2 + \lambda Tr(\boldsymbol{XLX}^T), \end{aligned} \tag{3}$$

where $Tr(\boldsymbol{XLX}^T) = \frac{1}{2}\sum_{ij}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\bar{w}_{ij}, \boldsymbol{L} = \boldsymbol{D} - \bar{\boldsymbol{W}}$ is called graph Laplacian, **D** is a diagonal matrix with $d_{jj} = \sum_{k=1}^{n}\bar{w}_{jk}$ and $\bar{w}_{jk}$ is from the definition in Eq. (16).

Graph regularized NMF (GNMF) derived from Eq. (3) aims to learn the local invariance of sample data in a neighborhood based on graph theory, which can enhance the efficiency of algorithms by learning the intrinsic geometric features of data. Currently graph regularized NMF has been extended and extensively applied for data representation, including dual embedding regularized NMF, *Lp* smooth NMF, graph-based discriminative NMF, and so on [12]–[16]. This type of algorithms has the problem of redundant solutions and scale transfer problem [33] as graph regularization terms are incorporated into the objective functions for the constraint of local invariance. Based on the GNMF, the original data **Y** can be separated into different views $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$,..., $\mathbf{Y}^{(m_v)}$. Then graph regularized Multiview learning algorithms can be developed. For the original data **Y**, its representation **X** in feature subspace has the same separation $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$,..., $\mathbf{X}^{(m_v)}$. Then, for any view $\mathbf{X}^{(v)}$, it follows that

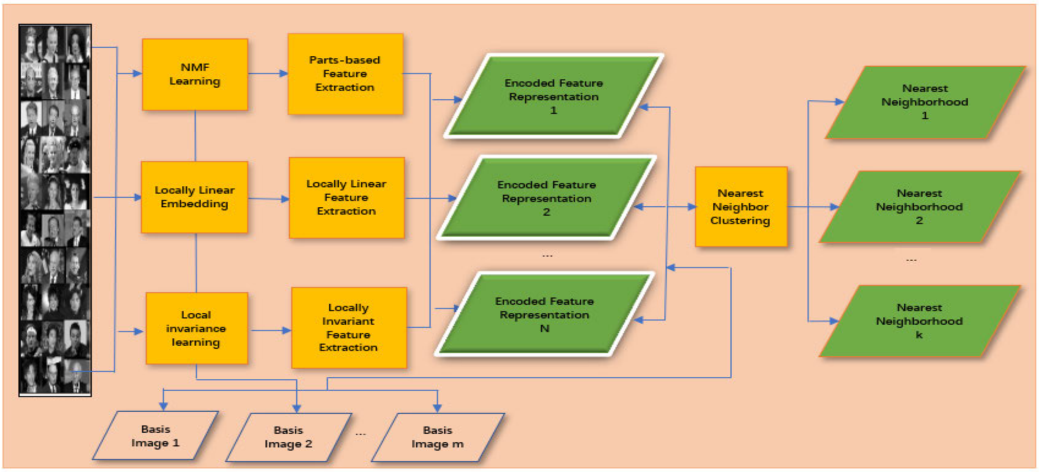

NMF Learning
Parts-based Feature Extraction
Encoded Feature Representation 1
Nearest Neighborhood 1
Locally Linear Embedding
Locally Linear Feature Extraction
Encoded Feature Representation 2
Nearest Neighbor Clustering
Nearest Neighborhood 2
...
...
Local invariance learning
Locally Invariant Feature Extraction
Encoded Feature Representation N
Nearest Neighborhood k
Basis Image 1
Basis Image 2
...
Basis Image m

FIGURE 1. **The block diagram of the proposed learning model.**

$D_{jj}^{(v)} = \sum_k \bar{w}_{jk}^{(v)}$, $\boldsymbol{L}^{(v)} = \boldsymbol{D}^{(v)} - \bar{\boldsymbol{W}}^{(v)}$ will be the Laplacian matrix of the v-*th* view [41], [42]. Since the original dataset is separated into different views according to some of the features of images, this type of algorithms can obtain better performance on image recognition. However, the cost is that labelling data into different views may take lots of time. On the other hand, Graph regularized deep neural network is a significant extension of the original NMF algorithms [27], which incorporates traditional deep auto-encoders (DAEs) with the local variance constraint to generate a graph regularized deep neural network to extract geometric structures of sample data for image clustering. The deep learning model shows the promising performance of graph preserving algorithms [26], [28].

However, current graph regularized algorithms only consider the local invariance of sample data, which has effective test results on datasets such as COIL20, YaleB, MINST, PIE, ORL. All images in these datasets are with simple background. To deal with various real-world data feature extraction with robustness to complex background images, a general divergence measurement defined by Amari (called Amari's α- divergence) [7] can be introduced:

$$Div_\alpha(\boldsymbol{Y},\boldsymbol{AX}) = \sum_{ik} y_{ik}\frac{(y_{ik}/[\boldsymbol{AX}]_{ik})^{\beta-1}-1}{\beta(\beta-1)} + \sum_{ik}\frac{[\boldsymbol{AX}]_{ik}-y_{ik}}{\beta}, \qquad (4)$$

where $\beta = (1+\alpha\ )/2$. The α-divergence was defined in [48]. To simplify the expression, β is introduced to replace α, then it has the result in Eq. (4). The advantage of this divergence is that both the differences of $y_{ik}/[\boldsymbol{AX}]_{ik}$ and $y_{ik} - [\boldsymbol{AX}]_{ik}$ are measured. By adjusting the setting of β, the derived learning algorithms can obtain a trade-off between robustness and accuracy.

III. THE PEARSON DISTANCE WITH MULTIPLE LOCAL MANIFOLD CONSTRAINTS

For the Amari's α-divergences $\mathrm{Div}_\alpha(\boldsymbol{Y},\boldsymbol{AX})$ in Eq. (4), studies have shown that the variation of parameter β in this function will determine the robustness and efficiency of algorithms for feature retrieving of different types of sample data. In the case $\beta = (1+\alpha)/2$, α-divergence is also called Person distance. From this divergence, an interesting objective function can be defined for the development of learning algorithms [7]. We can incorporate the Person distance with some specific constraint terms such as the first or second ordered local similarity to extend the original objective function, from which some improved learning algorithms can be proposed to obtain further sparsity of components. In this paper, we consider the following optimization problem. Minimize a specifically extended α-divergence: local invariance and linear embedding controlled Pearson distance with $\beta = 2$, which is defined as follows.

$$\begin{aligned}
&Div_{\alpha M}(\boldsymbol{Y},\boldsymbol{AX}) \\
&\quad = Div_\alpha(\boldsymbol{Y},\boldsymbol{AX}) + \lambda_X Div_\alpha(\boldsymbol{x_j},\boldsymbol{x_k}) + \lambda_A Div_\alpha(\boldsymbol{a_i},\boldsymbol{a_j}) \\
&\quad = \sum_{ik}\left\{\frac{y_{ik}^2/[\boldsymbol{AX}]_{ik} - 2y_{ik} + [\boldsymbol{AX}]_{ik}}{2}\right\} \\
&\qquad + \lambda_X Div_\alpha(\boldsymbol{x}_j,\boldsymbol{x}_k) + \lambda_A Div_\alpha(\boldsymbol{a}_i,\boldsymbol{a}_j) \\
&\qquad\qquad s.t.\ a_{ij} \geq 0,\quad x_{jk} \geq 0,\ \ \beta \neq 0,\ \ \beta \neq 1, \\
&\qquad\qquad\quad ||a_j||_2 = \sum_{i=1}^{M} a_{ij} = 1,
\end{aligned} \qquad (5)$$

where the parameter λ_X and λ_A will determine the effect of regularization on different scenes or noises of sample data. For the extra terms in Eq. (5), we constrain $Div_\alpha(\boldsymbol{x}_j, \boldsymbol{x}_k)$ and $Div_\alpha(\boldsymbol{a}_i, \boldsymbol{a}_j)$ to enforce the local features of the decomposed factors. The detail definitions of these terms will be given in Section IV.

Applying the gradient descent approach to Eq. (5), we have the following results:

$$x_{jk} \leftarrow x_{jk} - \eta_{jk}\frac{\partial Div_{\alpha M}\left(\boldsymbol{Y},\boldsymbol{AX}\right)}{\partial x_{jk}} \tag{6}$$

$$a_{ij} \leftarrow a_{ij} - \delta_{ij}\frac{\partial Div_{\alpha M}\left(\boldsymbol{Y},\boldsymbol{AX}\right)}{\partial a_{ij}}, \tag{7}$$

where η_{kj} and δ_{ij} are called the learning rates or step size parameters. The partial derivatives of elements in (6) and (7) can be computed as the follows.

$$\frac{\partial Div_{\alpha M}\left(\boldsymbol{Y},\boldsymbol{AX}\right)}{\partial x_{jk}} = x_{jk}\sqrt{\sum_{i=1}^{M} a_{ij}\frac{y_{ik}^2}{[\boldsymbol{AX}]_{ik}^2} + \frac{\lambda_X}{2}\frac{\partial Div_\alpha(\boldsymbol{x}_j,\boldsymbol{x}_k)}{\partial x_{jk}}}, \tag{8}$$

$$\frac{\partial Div_{\alpha M}\left(\boldsymbol{Y},\boldsymbol{AX}\right)}{\partial a_{ij}} = a_{ij}\sqrt{\sum_{k=1}^{N} x_{jk}\frac{y_{ik}^2}{[\boldsymbol{AX}]_{ik}^2} + \frac{\lambda_A}{2}\frac{\partial Div_\alpha(\boldsymbol{a}_i,\boldsymbol{a}_j)}{\partial a_{ij}}}. \tag{9}$$

Substituting (8) and (9) into (6), (7) respectively, and setting the step size parameters, the learning rules will be obtained.

IV. THE PROPOSED ALGORITHMS

In Eq. (5), how to define the extra terms $Div_\alpha(\boldsymbol{x}_j, \boldsymbol{x}_k)$ and $Div_\alpha(\boldsymbol{a}_i, \boldsymbol{a}_j)$ for more efficient manifold learning is an important research topic. LLE is one the most interesting manifold learning algorithms for dimensionality reduction of linear data, which was extensively applied to image classification and clustering, text recognition, and multi-dimensional data visualization. LLE is constructed with a simple geometric intuition, that is, in an s-vertex graph, nodes (or data points) are sampled from some underlying manifold. Assume that each node and its neighbors are always closely on a locally linear patch of the manifold [18], then the learned representation data points of these nodes in the feature subspace are also locally linear related with the corresponding coefficients in the sample data space. Manifold learning focuses on retrieving the geometric structures of images for feature recognition. To apply learning rules developed from Eq. (8) and Eq. (9) for manifold learning, we can define the regulation functions $Div_\alpha(\boldsymbol{x}_{\mathrm{j}}, \boldsymbol{x}_{\mathrm{k}})$ and $Div_\alpha(\boldsymbol{a}_{\mathrm{i}}, \boldsymbol{a}_{\mathrm{j}})$ with manifold constraints and impose them to the objective function in Eq. (5). From the manifold assumption in [18], [20], for the column vector $\boldsymbol{y}_j(j = 1, 2, \ldots, N)$ in matrix $\mathbf{Y}$, if they are locally linear related in some neighborhood, then their corresponding low dimensional representation vectors $\boldsymbol{x}_{\mathrm{j}} = [x_{\mathrm{j}1}, \ldots, x_{\mathrm{jn}}]^{\mathrm{T}}$ will also be in some neighborhood and linearly related each other with the same coefficients w_{jk} in the high dimensional space. The cost function of the LLE is defined in a neighborhood with s nearest neighbors, which assumes that the node x_{j} is linearly related with all the nodes in the neighborhood. If $\mathbf{W}$ is the weight matrix and w_{jk} is the *jk-th* element of this matrix, then the following linear reconstructing error was proposed to be the objective function [18] for LLE:

$$LE\left(\boldsymbol{W}\right) = \sum_{j=1}^{N} ||\boldsymbol{y}_j - \sum_{k=1}^{s} w_{jk}\boldsymbol{y}_k||_2^2, \tag{10}$$

$$s.t. \sum_{k=1}^{N} w_{jk} = 1. \tag{11}$$

Objective function (10) satisfying condition (11) has the following optimal solution:

$$\boldsymbol{w}_i = \frac{\boldsymbol{Z}_i^{-1}\boldsymbol{I}_k}{\boldsymbol{I}_k^T\boldsymbol{Z}_i^{-1}\boldsymbol{I}_k}, \tag{12}$$

where $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ik})^{\mathrm{T}}$, $\boldsymbol{I}_k = (1, 1, \ldots, 1)^{\mathrm{T}}$, and $\boldsymbol{Z}_i = (\boldsymbol{y}_j - \boldsymbol{y}_k)(\boldsymbol{y}_j - \boldsymbol{y}_k)^{\mathrm{T}}$. To reduce the redundant solutions [33], [34], we assume that the node connections are only in their corresponding neighborhoods, and the sample space are separated into l_1 neighborhoods, then according to Eq. (10) and (11), the locally linear regularizing term $LE\left(\mathbf{W}\right) = d_{lm}(\boldsymbol{x}_j, \boldsymbol{x}_k)$ in the low dimensional feature space can be defined as follows.

$$\begin{aligned} LE\left(\boldsymbol{W}\right) &= \sum_{j=1}^{N-1}\left(||\boldsymbol{x}_j - \sum_{k=1}^{s}\boldsymbol{x}_k w_{jk}||_2^2\right), \\ &= \sum_{s=1}^{l_1}\sum_{j=1}^{s_1}\left(||\boldsymbol{x}_j - \sum_{k=1}^{s_1-1}\boldsymbol{x}_k w_{jk}||_2^2\right) \\ s.t. &\sum_{k=1}^{N} w_{jk} = 1. \end{aligned} \tag{13}$$

where l_1 is the number of neighborhoods, s_1 is the number of elements in a neighborhood. If we only have two elements in a neighborhood, since $\sum_{k=1}^{n} w_{jk} = 1$, then the linear relationship of elements in the neighborhood will have

$$\begin{aligned} d_{lm}\left(\boldsymbol{x}_j, \boldsymbol{x}_k\right) &= \sum_{j=1}^{N-1}\left(||\boldsymbol{x}_j - \sum_{k=1}^{s}\boldsymbol{x}_k||_2^2 w_{jk}\right) \\ &= \sum_{j=1}^{N-1}||\boldsymbol{x}_j - \boldsymbol{x}_{j+1}||_2^2. \end{aligned} \tag{14}$$

Thus, learning the local invariance of elements in a neighborhood is the special case of learning the locally linear representation of elements in a neighborhood.

However, as we have mentioned above, if the size of a neighborhood is too small or too big, the locally linear embedding of elements may not exit in this neighborhood. Therefore, we cannot define the similarity measurement and the locally linear relation in the same neighborhood. The locally linear embedding and the local invariance should be considered in two different scales of neighborhood.

From the divergence in Eq. (13), a robust locally linear regularization term can be defined as the following to develop the new learning algorithms.

$$\begin{aligned} &d_{\alpha\_lm}(\boldsymbol{x}_j,\boldsymbol{x}_k) \\ &= \sum_{s=1}^{l_1}\sum_{j=1}^{s_1}\left(||\frac{\boldsymbol{x}_j^2}{\sum_{k=1}^{s_1}\boldsymbol{x}_k w_{jk}}-2x_j+\sum_{k=1}^{s_1-1}\boldsymbol{x}_k w_{jk}||_2^2\right) \\ &\quad s.t. \sum_{k=1}^{N} w_{jk}=1. \end{aligned} \tag{15}$$

Spectral graph theory and manifold learning theory show that nearest neighbor graph modelling can be applied to the extracting of local geometric structures on a high dimensional sample dataset. For a graph with s vertices, we assume that each vertex represents a data point. For each data point $\boldsymbol{y}_j$, we define its p nearest neighbors and connect them with edges. $\bar{w}_{jk}$ is defined to be the weight of the jk-th edge. For any two points $\boldsymbol{y}_j$ and $\boldsymbol{y}_k$ on a nearest neighbor graph, three different choices can be used to define the weight matrix $\bar{\boldsymbol{W}}$ on the graph [30], including:

1. 0-1 weighting,

$$\bar{w}_{jk} = \begin{cases} 1, & \textit{if } \boldsymbol{y}_j \textit{ and } \boldsymbol{y}_k \textit{ connected} \\ 0, & \textit{otherwise}, \end{cases}$$

2. Heat kernel weighting,

$$\bar{w}_{jk} = \begin{cases} e^{\frac{-||y_j-y_k||^2}{\sigma}}, & \textit{if } \boldsymbol{y}_j \textit{ and } \boldsymbol{y}_k \textit{ connected} \\ 0, & \textit{otherwis}, \end{cases}$$

and

3. Dot-product weighting,

$$\bar{w}_{jk} = \begin{cases} \boldsymbol{y}_j^T\boldsymbol{y}_k, & \textit{if } \boldsymbol{y}_j \textit{ and } \boldsymbol{y}_k \textit{ connected} \\ 0, & \textit{otherwis}. \end{cases}$$

On the other hand, the local invariance in the manifold space assumes that in a sample data space, if two data points $\boldsymbol{y}_j$ and $\boldsymbol{y}_k$ are close in the intrinsic geometry of the data distribution, the representations of these two points $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ in the feature representation subspace are also close to each other. These points will be defined in a neighborhood and connected with edges each other [30]. To build the general local feature extracting model with different scales, we define a smaller neighborhood in the linear representation subspace to capture the local similarity of nearest neighbors in the neighborhood. The similarity measurement d_{sm} in the low dimensional feature space is defined as the following:

$$\begin{aligned} d_{sm}\left(\boldsymbol{x}_j,\boldsymbol{x}_k\right) &= \sum_{j,k=1}^{N}\left(||\boldsymbol{x}_j-\boldsymbol{x}_k||_2^2\bar{w}_{jk}\right) \\ &= \sum_{s=1}^{l_2}\sum_{j,k=1}^{s_2}\left(||\boldsymbol{x}_j-\boldsymbol{x}_k||_2^2\bar{w}_{jk}\right), \end{aligned} \tag{16}$$

where l_2 is the number of neighborhoods in the subspace, $\bar{w}_{jk}$ is the edge of linking two nodes in a graph, which is defined with s_2 nearest neighbors on the graph. The simplest approach to define the matrix $\bar{\boldsymbol{W}}$ is the 0-1 weighting. Similar to the expression in Eq. (15), the robust similarity measurement can be defined as follows to develop our learning algorithms.

$$d_{\alpha\_sm}(\boldsymbol{x}_j,\boldsymbol{x}_k) = \sum_{s=1}^{l_2}\sum_{j=1}^{s_2}\left(||\frac{\boldsymbol{x}_j^2}{|\boldsymbol{x}_k|}-2\boldsymbol{x}_j+\boldsymbol{x}_k||_2^2\bar{w}_{jk}\right) \tag{17}$$

Combining Eq. (15) and Eq. (17) with Eq. (5), the multiscale representation objective function can be defined as:

$$\begin{aligned} Div_{\alpha M}\left(\boldsymbol{Y},\boldsymbol{AX}\right) = Div_{\alpha}\left(\boldsymbol{Y},\boldsymbol{AX}\right) &+\lambda_{lx}d_{\boldsymbol{\alpha}\_lm}\left(\boldsymbol{x}_j,\boldsymbol{x}_k\right) \\ &+\lambda_{sx}d_{\alpha\_sm}\left(\boldsymbol{x}_j,\boldsymbol{x}_k\right)+\lambda_A d_{\alpha\_sm}\left(\boldsymbol{a}_i,\boldsymbol{a}_j\right), \end{aligned} \tag{18}$$

where the definition of divergence $d_{\alpha\_sm}(\boldsymbol{a}_\mathrm{i},\boldsymbol{a}_\mathrm{j})$ is similar to $d_{\alpha\_sm}(\boldsymbol{x}_\mathrm{j},\boldsymbol{x}_\mathrm{k})$. In Eq. (18), the locally linear embedding and local invariance measurements are defined in different scales of neighborhoods. Thus, the local feature extracting algorithms for multi-scale representations can be derived as follows:

$$\begin{aligned} x_{jk} \leftarrow x_{jk}\sqrt{\sum_{i=1}^{M}a_{ij}\frac{y_{ik}^2}{[\boldsymbol{AX}]_{ik}^2}+\lambda_{lx}\frac{\partial d_{\alpha\_lm}(\boldsymbol{x}_j,\boldsymbol{x}_k)}{\partial x_{jk}}} \\ +\lambda_{sx}\frac{\partial d_{\alpha\_sm}(\boldsymbol{x}_j,\boldsymbol{x}_k)}{\partial x_{jk}}, \end{aligned} \tag{19}$$

$$a_{ij} \leftarrow a_{ij}\sqrt{\sum_{k=1}^{N}x_{jk}\frac{y_{ik}^2}{[\boldsymbol{AX}]_{ik}^2}+\lambda_A\frac{\partial d_{\alpha\_sm}(\boldsymbol{a}_i,\boldsymbol{a}_j)}{\partial a_{ij}}}. \tag{20}$$

With the variations of neighborhood size p/s and the learning parameters λ_{lx} and λ_{sx}, the algorithms in (19) and (20) learn different scales of manifold features, which is possible to capture more complex geometric structures of images. In the following section, we will show the application of the proposed algorithms to dimensionality reduction and clustering of image data with different scenes.

V. COMPLEXITY ANALYSIS

For the proposed algorithms, to simplify the computing for complexity, we separate the measurements of local invariance and locally linear embedding into two parts, one is their difference, another one is their division. In general, they need $O((K+1)(p+q)N^2)$ to construct the k-nearest neighbor graphs, where p is the neighborhood size of local invariance and q is the neighborhood size of locally linear embedding. In the learning, the complexity of non-negative matrix factorization is $O(MNK^2)$ since the computing of $[\boldsymbol{AX}]_{\mathrm{jk}}$ needs K extra times, Assume that t is the iteration numbers, then the overall complexity for the algorithms is $O(t(K+1)(p+q)N^2+tMNK^2)$. As p and q are very small numbers, the proposed algorithm has almost the same cost to the GNMF algorithm.

VI. EXPERIMENTAL RESULTS

A. COMPARED ALGORITHMS AND DATASETS

To show the efficiency of the proposed algorithms, in the experiments we compare our algorithms in learning rules (19)

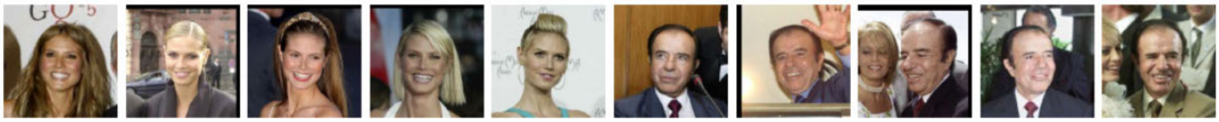


FIGURE 2. The original photos selected from dataset: Labeled Faces in the Wild. Each person has different poses. The hair styles of Heidi are various, and there are different people in the background of Carlos, which may degrade the recognition results in the learning.

FIGURE 3. The original photos selected from the Caltech 101 image Objects dataset. The images are with complex background and the size of target objects are various, which leads the difficulties for the image clustering in this dataset.

FIGURE 4. The original photos selected from the UFI large dataset that contains images extracted from real photographs acquired by reporters. Facial images are with different scenes. The size and the pose of target images are quite different.

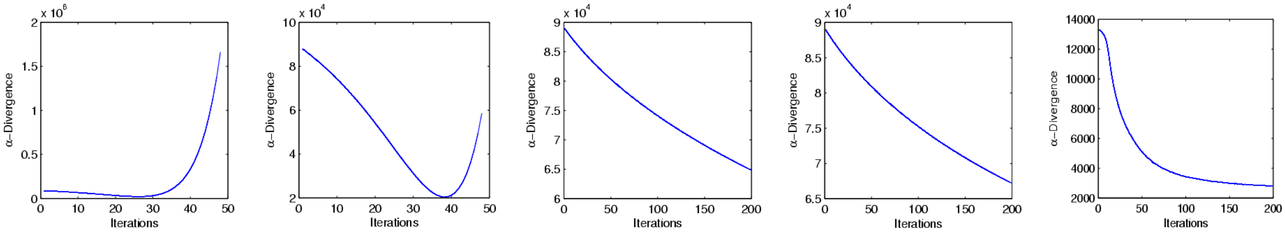


FIGURE 5. The variation curves of $Div_{\alpha}(Y, AX)$ in the learning of clustering the UFI large images with $\lambda_{lx} = 95$ and $\lambda_{sx} = 15, 10, 0.1, 0.001$, and 0.00001 (From left to right respectively).

and (20) with the following algorithms: Some early developed algorithms including Normalized Cut (NCut) [35] and Lee and Seung's NMF [6]. Some recently developed methods including GNMF, a Graph regularized Nonnegative Matrix Factorization method, LGNMF [36], an algorithm employed local centroid structured constraint to achieve sparse representation $\mathbf{X}$, RSNMF [37], a semi-supervised NMF which was introduced to obtain the robust discriminative representation, and MPMNMF [43], a multi-view clustering based NMF algorithm aimed to seek the manifold measurements in the decomposed factors. Deep WSF [26], a multi-layer algorithm to learn a hierarchy of hidden representations so that the final lower-dimensional representation of the data can be extracted with higher quality. The drawback of this model is that the datasets must be with mixed attribute knowledge such as attributes pose, expression, and identity. GR-DNN [27], a deep neural network with traditional auto-encoding to obtain the ability of local geometric structure retrieving of images. In this model, only the local invariance feature of images is learnt. The images that we have selected for the tests include the following four datasets, and each image is resized to 32 × 32 gray scale for the neural network training and clustering. COIL20 database, which contains 20 different sample objects. In this set, each object has 72 images, which were taken 5 degrees apart with the object rotating on a turntable [44]. The images in this dataset have only the target object in a picture. Labeled Faces in the Wild (LFW), a database of face photographs designed for studying the unconstrained face recognition. The data set is with more than 13,000 images of faces collected from the web. 1680 of the people pictured have two or more distinct photos in the data set [45]. We select this dataset for image clustering to show the effectiveness of the proposed algorithms in a complex scene since facial images in this dataset have different occlusions in the front or different persons and scenes on the background. Unconstrained Facial Images (UFI) large database. We only select the training dataset for the tests. The total number of the subjects in this dataset is 530 and an average number/person of training images is 8.2. The original size of images is 384 × 384 pixels. The images in this set have

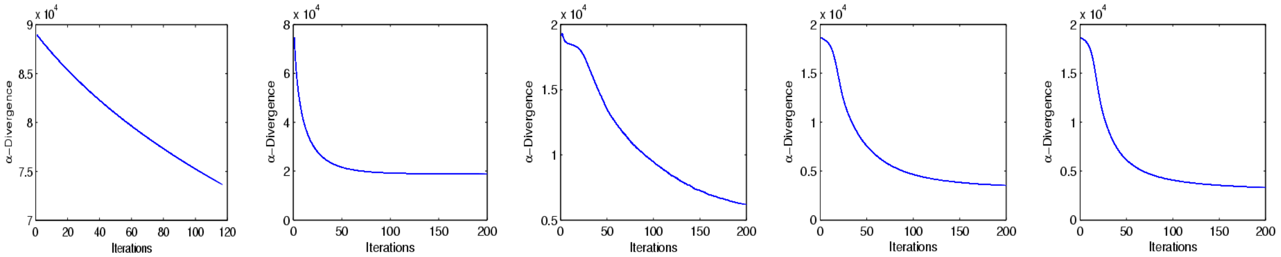


FIGURE 6. The variation curves of $Div_{\alpha}(Y, AX)$ in the learning of clustering the UFI large images with $\lambda_{lx} = 0.00001$ and $\lambda_{sx} = 10, 1, 0.01, 0.001$, and 0.0001 (From left to right respectively).

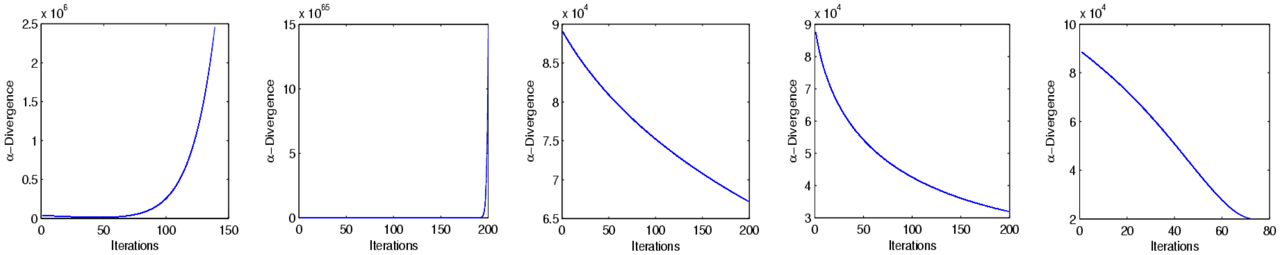


FIGURE 7. The variation curves of $Div_{\alpha}(Y, AX)$ in the learning of clustering the UFI large images with $\lambda_{sx} = 0.01$ and $\lambda_{lx} = 10, 0.1, 0.01, 0.001$, and 0.0001 (From left to right respectively).

complex background, the face size also significantly differs and the faces are not localized [46]. This dataset is selected to show the effectiveness of the proposed algorithms in a more complex scene.

The cropped UFI images is selected from the UFI large dataset. This dataset contains images of 605 people with an average of 7.1 images per person in the training set. The images are cropped to an original size of 128 × 128 pixels with only the face part of these selected persons. Similar to the images in the COIL20 dataset, the images in this dataset also have only the target face in a picture but the pose of facial images may vary arbitrarily.

Caltech 101 dataset. Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. The original size of each image is roughly 300 × 200 pixels [47].

We select these four datasets to test the efficiency of our algorithms since they have various scenes. Fig. 2, Fig. 3, and Fig. 4 show the original images selected from the LFW, Caltech101, and the UFI large dataset respectively, which clearly show the complexity of the image data in different scenes. In Fig. 2, the hair styles of Heidi are various, and there are different people or objects in the background of Carlos. In Fig. 3, we can see that the images are with complex background and the sizes of target objects are various. Some target images are so blurred even our eyes cannot identify them clearly. In Fig. 4, each person's facial images are with different scenes. The size and the pose of target images are also quite different.

In the tests, we set the model with three scales training. The NMF model is used to learn the global feature of sample data for parts-based representation. The locally linear feature and the local invariant feature can be learned by set different scales of neighborhood size. Tests results show that the proposed algorithms can learn the best performance with some specific settings.

B. CONVERGENCE STUDY

The convergence of objective function and the learning rules will significantly determine the accuracy of proposed algorithms. Thus, it is very necessary to study the convergent properties of the new learning algorithms when they are utilized to cluster data. Since a_{ij} is normalized in the learning, we always have $0 \leq a_{ij} \leq 1$. Thus, element a_{ij} will be non-divergent. We only discuss the convergent properties of the model for objective function and components $x_{jk} (j = 1, 2, \ldots, K, k = 1, 2, \ldots, N)$ in the learning.

Fig. 5, Fig. 6 and Fig. 7 show the variations of objective function in the clustering of UFI large dataset. From Fig. 5, we can see that in the cases $\lambda_{sx} = 15, 10, 0.1, 0.001$, and 0.00001 from left to right respectively, the objective function $Div_{\alpha M}(\boldsymbol{Y}, \boldsymbol{AX})$ becomes convergent from divergent, where λ_{lx} is fixed with $\lambda_{lx} = 95$. These figures show that, the proposed objective function diverges at the cases of $\lambda_{lx} = 95, \lambda_{sx} = 15$, 10. Obviously, when λ_{sx} is too big, the objective function will be divergent. Thus, the component x_{jk} converges only if we have smaller λ_{sx}. Fig. 6 shows that in thex cases of $\lambda_{lx} = 0.00001$, $\lambda_{sx} = 10, 1, 0.01, 0.001$, and 0.0001, all the curves of $Div_{\alpha M}(\mathbf{Y}, \mathbf{AX})$ are convergent. Meanwhile, the convergence will become faster when we have smaller λ values. From Fig. 7 we can see that when $\lambda_{sx} = 0.01$,

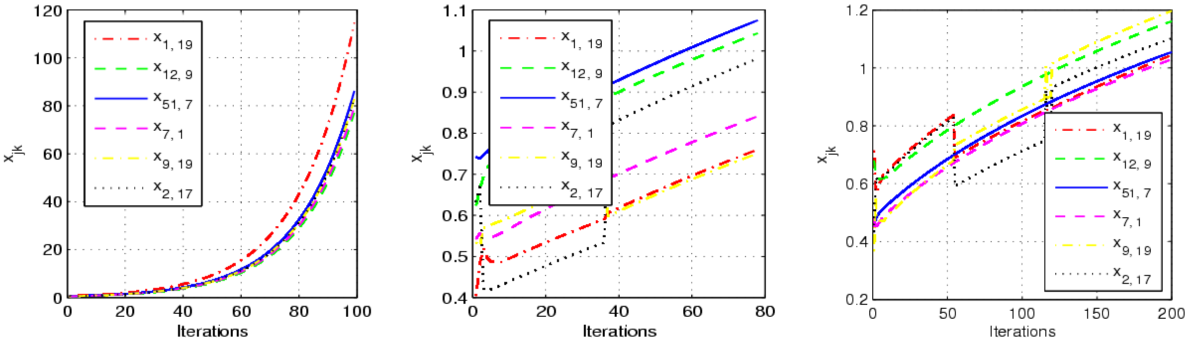


FIGURE 8. The variation curves of x_{jk} in the learning of clustering the UFI-large images with $\lambda_{lx} = 95$, $\lambda_{sx} = 10, 0.1, 0.00001$ (from left to right respectively), 6 elements are selected from the encoding matrix to present the convergence of the algorithm.

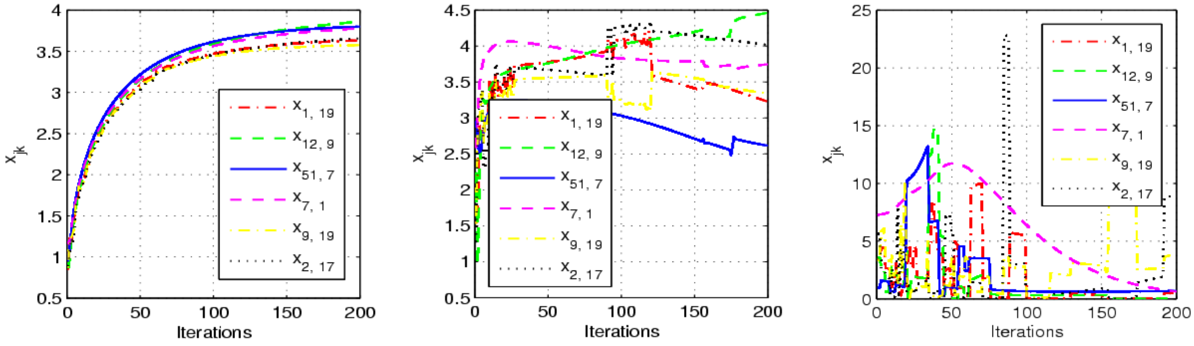


FIGURE 9. The variation curves of x_{jk} in the learning of clustering the UFI large images with $\lambda_{lx} = 1, 0.1, 0.00001$ (from left to right respectively), $\lambda_{sx} = 0.00001$. 6 elements are selected from the encoding matrix to present the convergence of the algorithm.

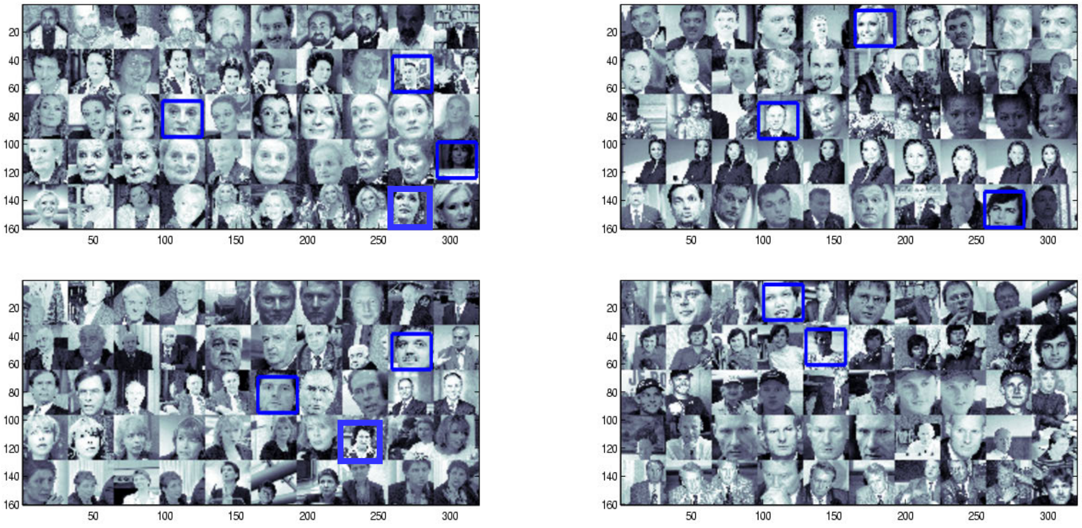
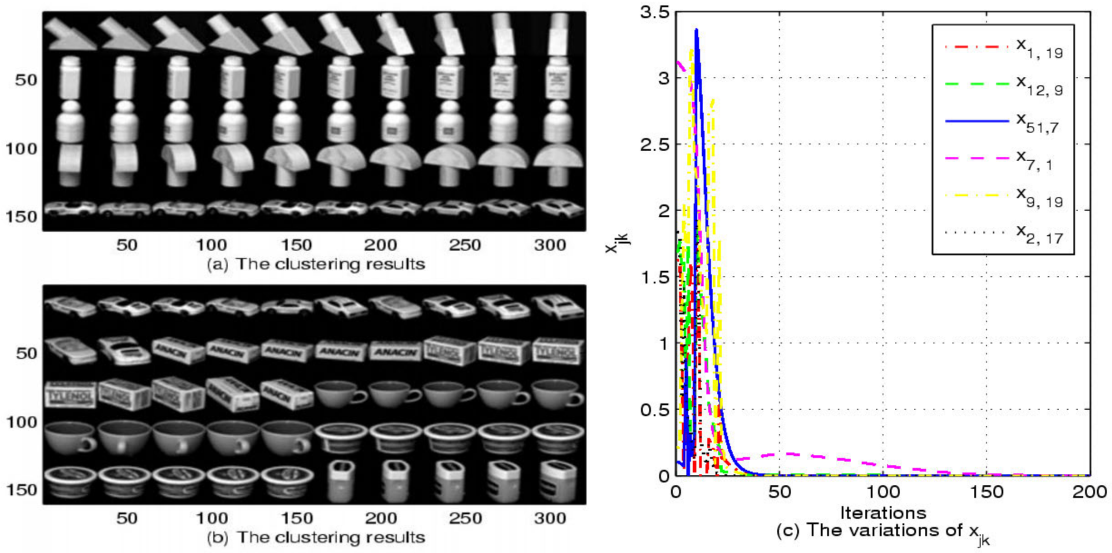


FIGURE 10. The clustered images by the proposed method on UFI large dataset with $\lambda_{lx} = 95$, $\lambda_{sx} = 0.00001$.

$\lambda_{lx} = 10, 0.1$, the objective function is divergent. With the decrease of λ_{lx} the objective function begins to converge. When λ_{sx} is fixed, the smaller λ_{lx} will lead better convergence of objective function. Test results in Fig. 7 show when $\lambda_{lx} \geq 0.01$, the objective function may diverge. Decreasing λ_{sx}, then the learning becomes convergent. In general, the convergence of objective function and learning updates may not guarantee the learning to obtain the best clustering result, but if the objective function and/or learning updates diverge, the clustering certainly cannot obtain the best clustering result. Fig. 10 shows that the best clustering result on UFI set is obtained only when $\lambda_{lx} = 95$, $\lambda_{sx} = 0.00001$. In this



(a) The clustering results

(b) The clustering results

(c) The variations of x_{jk}

FIGURE 11. The clustered images and the corresponding variations of x_{jk} in the learning of the proposed method on COIL20 dataset with $\lambda_{lx} = \lambda_{sx} = 0.00001$.

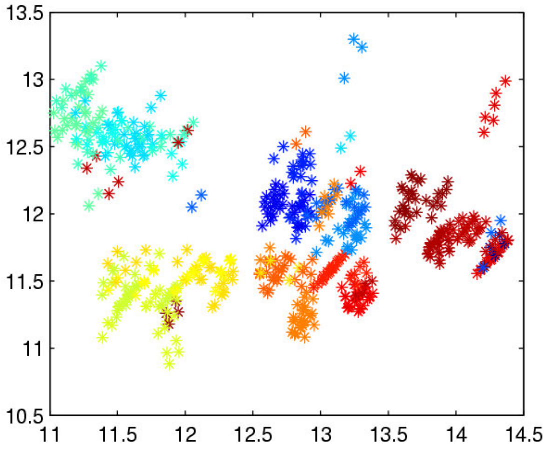


FIGURE 12. The clustered results when we map the feature data into two-dimensional data space in the learning of the proposed method on UFI dataset with $\lambda_{lx} = 95$, $\lambda_{sx} = 0.00001$.

case both the objective function and the learning algorithm are convergent. Thus, the setting of parameters is important for the clustering.

Fig. 8 and Fig. 9 show the variations of different elements x_{jk} in the clustering of image data, which indicate that the proposed algorithms have the case of divergence since the curves in the first sub-figure of Fig. 8 are always going up with the increase of the iterations. To study the convergent properties of this algorithm, all the x_{jk} in $\mathbf{X}$ are selected arbitrarily to present. This figure also shows that with the decrease of λ_{sx}, the learning becomes convergent gradually. Fig. 9 shows the non-divergence of the algorithms when we have relatively small λ_{lx} and λ_{sx} settings. The left sub-figure in Fig. 9 shows that the learning converges at about the iteration of 100, but for the right two sub-figures, although the learning of x_{jk} are not divergent, the curves are not converging to fixed points. They are with oscillation in the learning. Test results show that when the variations of elements are with oscillating or divergence, the algorithm cannot obtain high clustering accuracy. On the other hand, by comparing the results in Fig. 6 and Fig. 9, in the case of $\lambda_{lx} = \lambda_{sx} = 0.00001$, the objective function converges but the variations of x_{jk} are with oscillating, which indicate that the convergence of objective function may not guarantee the convergence of learning algorithms.

C. CLUSTERING RESULTS

The experiments were running on Windows 10 operating system, Intel(R) Core(TM) i5-7200 CPU with 2.50 GHz, 2.71 GHz, 4.00 GB main memory. For the proposed algorithms, including decomposition and clustering, it takes about 2.35 minutes to process 100 images. For the COIL20 images, it totally takes about 35.8 minutes, for the LFW images, it takes about 315.06 minutes, for the Caltech101 images, it takes about 146.87 minutes, for the UFI images, it takes about 112.17 minutes, and for the cropped UFI images, it takes about 110.95 minutes to finish the learning and image clustering.

Fig. 10 shows one of the clustering results of images on UFI large dataset. We list the clustered images in a sequence with 10 images for each row, and we only show the clustering results on 200 facial images of this set. The results in Fig. 10 indicate that, under the given conditions, the proposed model can learn very good feature representations and clustering since almost all the images are clustered correctly, only about 12 facial images are in wrong positions in total 200 facial images (facial images with blue boxes are in wrong clusters).

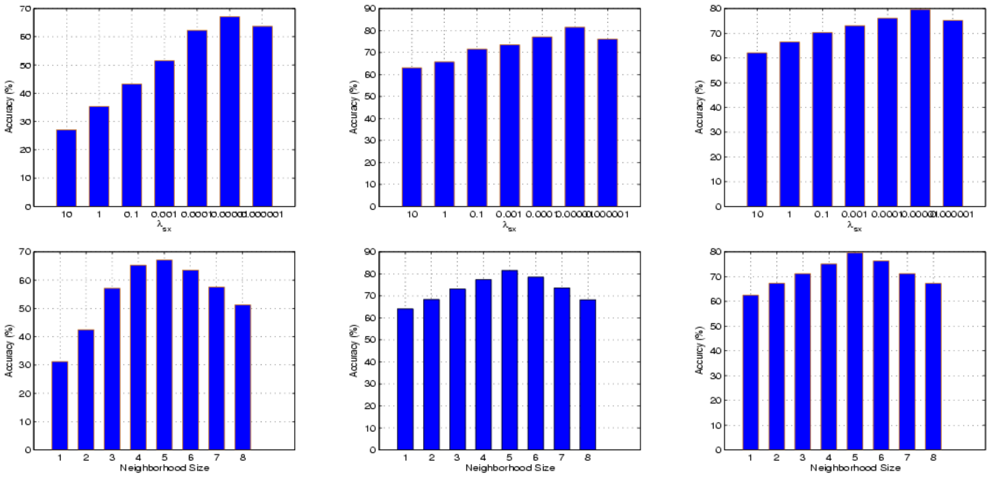


FIGURE 13. The variations of clustering accuracy for different settings of λ_{sx} and neighborhood size r_1 on three different datasets with fixed $\lambda_{lx} = 95$ and similarity neighborhood size $r_2 = 2$.

TABLE 1. The average clustering metric OFACC on different datasets (%).

Dataset	COIL20	LWF-	Caltech101	UFI-Cropped	UFI-Large
NCut	68.69±1.71	67.31±1.92	65.36±1.72	55.21±2.27	34.61±1.61
NMF	61.31±1.51	63.61±2.01	58.75±1.38	52.61±1.63	32.41±1.57
GNMF	73.41±1.75	72.14±1.57	71.37±1.72	61.14±1.75	37.14±1.79
RSNMF	77.32±2.31	74.52±1.72	73.72±1.67	67.17±1.58	41.52±2.53
LGNMF	79.67±2.74	76.75±2.05	74.13±1.78	66.83±1.93	47.83±1.65
MPNMF	**87.51±2.35**	79.61±1.52	76.63±1.74	75.67±1.55	60.67±1.38
Deep WSF	85.33±1.97	75.46±2.18	75.52±1.68	72.51±2.25	60.32±2.18
GR-DNN	86.75±1.52	78.31±1.58	77.25±1.41	**76.27±1.92**	62.71±1.79
The Proposed	87.35±1.62	**81.55±1.37**	**79.61±2.19**	75.21±1.35	**67.15±1.26**

TABLE 2. The average clustering metric of NMI on different datasets (%).

Dataset	COIL20	LWF-	Caltech101	UFI-Cropped	UFI-Large
NCut	74.71±1.57	75.61±1.83	69.61±2.23	58.61±1.78	38.61±2.21
NMF	68.21±1.27	70.71±1.91	62.57±2.05	54.61±2.11	35.61±1.72
GNMF	81.72±1.79	82.14±1.47	77.54±2.35	63.19±1.81	39.14±1.29
RSNMF	85.73±1.74	83.38±1.69	80.61±2.37	71.59±1.75	45.47±1.92
LGNMF	86.91±1.93	85.81±1.59	79.83±1.31	69.69±2.36	50.87±1.58
MPNMF	89.51±1.57	86.36±2.01	81.63±1.83	**79.65±1.37**	67.16±1.95
Deep WSF	87.22±2.12	83.77±1.72	78.52±1.67	75.71±1.41	65.64±1.17
GR-DNN	89.35±1.62	85.46±1.76	80.32±2.35	78.83±1.74	67.53±1.93
The Proposed	**89.55±1.76**	**87.65±1.83**	**83.71±2.27**	79.53±1.67	**71.65±1.75**

In Fig. 11, sub-figures (a), (b) show the clustering results of images (we only select 100 clustered images to show the test results, 10 images for each object), and sub-figure (c) shows the variation curves of x_{jk} in the feature extracting on COIL20 dataset when $\lambda_{lx} = \lambda_{sx} = 0.00001$. The results indicate that all the images are clustered correctly. The right sub-figure shows that all the selected x_{jk} converge after 30 iteration learning. This figure is provided to show the relationship between the convergence and the clustering results. Comparing with the convergence curves in Fig. 9, when $\lambda_{lx} = \lambda_{sx} = 0.00001$, the learning does not diverge but the curves are oscillating. Test results show that in this case, the clustering results on UFI large dataset is not the best result. The best clustering results are obtained when $\lambda_{lx} = 95$, $\lambda_{sx} = 0.00001$ on this dataset. The last subfigure in Fig. 8 shows that in this case, all x_{jk} are converging gradually. Thus, the convergence of learning is significantly related to the clustering accuracy. Fig. 12 shows the clustered results of UFI-Large data when the feature data are mapped in to a two-dimensional space, from which we can see that in general, most images are clustered to the correct clusters. But few of them are in wrong clusters. Some images are not in any clusters, which are the outliers of clustering. Since its time consuming to label and reduce all the image data for low dimensional mapping and visualization, we only select about 600 images, 20 classes to show the test results. In fact, the results in Fig. 10 are also the visualization of clustered images on UFI-Large dataset. The only difference is that we line up the images one cluster by one cluster in one dimensional space, but not in two-dimensional data space. From left to right, Fig. 13 shows the relationship of clustering accuracy with the parameters λ_{sx} and neighborhood size r_1 (locally linear embedding) on the datasets UFI large, LFW, and Caltech respectively (Here the neighborhood size indicates the radius of a neighborhood). In this figure, we set fixed $\lambda_{lx} = 95$ and neighborhood size $r_2 = 2$ (local invariance). The sub-figures in the first-row show that when the similarity measurement parameter is fixed, the proposed algorithm obtains the best performance at the point $\lambda_{sx} = 0.00001$. Increase or decrease this parameter will degrade the accuracy of this algorithm. The sub-figures in the second-row show that in the case of the similarity neighborhood size $r_2 = 2$, the best performance is obtained at the point of the linear relation neighborhood size $r_1 = 5$. Since in the UFI large dataset, each person has about 8.2 facial images, considering all the images of each person to be in a locally linear patch may obtain the best clustering result. Thus, the radius of the neighborhood $r_1 = 5$ may have all facial images of one person in the same neighborhood. For other two sets, since the number of images for each object is the times of ten. Therefore, they can obtain the best performance in this case.

Table 1 and Table 2 show the average of clustering accuracy (ACC) and the normalized mutual information (NMI) for ten time running of each algorithm on each dataset. From the two tables, it is clear that, by comparing with the state of art algorithms, the proposed algorithms obtain the best performance in terms of ACC and NMI in complex scenes, although the MPMNMF and GR-DNN algorithms can learn close or better to the proposed algorithms on COIL20 and UFI cropped datasets. Since in these two sets, they do not have complex backgrounds, our algorithms obviously have the advantage of robustness in dealing with various scenes.

VII. CONCLUSION

In this paper, multi-scale manifold constrained NMF algorithms are proposed to exploit the intrinsic geometric features of images in different scenes, which can learn a state of art performance in image clustering. The experimental results confirm the efficiency of the proposed model. Analysis and the test results also show that the convergence of the objective function cannot guarantee the convergence of its corresponding learning algorithms. The convergence of objective function and the derived learning algorithms are significantly related to the accuracy of image clustering in various scenes. In the future, it is necessary to explore the deep structures of manifold learning for different applications. In general, the new design of manifold learning with flexible scales of neighborhoods size provides an efficient approach for extracting intrinsic geometric features of images in complex scenes.

REFERENCES

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, pp. 153–160.

[2] D. Kuang and H. Park, "Fast rank-2 nonnegative matrix factorization for hierarchical document clustering," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 739–747.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[4] N. Shahid, V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Robust principal component analysis on graphs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2812–2820.

[5] E. Hosseini-Asl, J. M. Zurada, and O. Nasraoui, "Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2486–2498, Dec. 2016.

[6] 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.

[7] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csiszár's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 621–625.

[8] A. Lemme, R. F. Reinhart, and J. J. Steil, "Online learning and generalization of parts-based image representations by non-negative sparse autoencoders," *Neural Netw.*, vol. 33, pp. 194–203, Sep. 2012.

[9] S. Yang, Z. Yi, X. He, and X. Li, "A class of manifold regularized multiplicative update algorithms for image clustering," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5302–5314, Dec. 2015.

[10] B. Liu, Z. Xu, S. Wu, and F. Wang, "Manifold regularized matrix completion for multilabel classification," *Pattern Recognit. Lett.*, vol. 80, pp. 58–63, Sep. 2016.

[11] D. Kim and J. P. Haldar, "Greedy algorithms for nonnegativity-constrained simultaneous sparse recovery," *Signal Process.*, vol. 125, pp. 274–289, Aug. 2016.

[12] W. Wu, S. Kwong, J. Hou, Y. Jia, and H. H. S. Ip, "Simultaneous dimensionality reduction and classification via dual embedding regularized nonnegative matrix factorization," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3836–3847, Aug. 2019.

[13] C. Leng, H. Zhang, G. Cai, I. Cheng, and A. Basu, "Graph regularized Lp smooth non-negative matrix factorization for data representation," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 2, pp. 584–595, Mar. 2019.

[14] J. Mei, Y. De Castro, Y. Goude, J.-M. Azais, and G. Hebrail, "Nonnegative matrix factorization with side information for time series recovery and prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 493–506, Mar. 2019.

[15] Y. Li, Y. Pan, and Z. Liu, "Multiclass nonnegative matrix factorization for comprehensive feature pattern discovery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 615–629, Feb. 2019.

[16] H. Li, J. Zhang, G. Shi, and J. Liu, "Graph-based discriminative nonnegative matrix factorization with label information," *Neurocomputing*, vol. 266, pp. 91–100, Nov. 2017.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14. Cambridge, MA, USA: MIT Press, 2002, pp. 585–591.

[18] 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, 2000.

[19] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[20] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 10, pp. 5591–5596, May 2003.

[21] A. Brun, C.-F. Westin, H. Knutsson, and M. Herberthson, "Fast manifold learning based on Riemannian normal coordinates," in *Proc. 14th Scand. Conon Image Anal.*, 2005, pp. 921–929.

[22] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1717–1729, Jul. 2013.

[23] S. Deutsch, S. Kolouri, K. Kim, Y. Owechko, and S. Soatto, "Zero shot learning via multi-scale manifold regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2017, pp. 5292–5299.

[24] L. Rossi, A. Torsello, and E. R. Hancock, "Unfolding kernel embeddings of graphs: Enhancing class separation through manifold learning," *Pattern Recognit.*, vol. 48, no. 11, pp. 3357–3370, Nov. 2015.

[25] C.-S. Lee, A. Elgammal, and M. Torki, "Learning representations from multiple manifolds," *Pattern Recognit.*, vol. 50, pp. 74–87, Feb. 2016.

[26] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 417–429, Mar. 2017.

[27] S. Yang, L. Li, S. Wang, W. Zhang, and Q. Huang, "A graph regularized deep neural network for unsupervised image representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1203–1211.

[28] Y. Meng, R. Shang, F. Shang, L. Jiao, S. Yang, and R. Stolkin, "Semi-supervised graph regularized deep NMF with bi-orthogonal constraints for data representation," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2019.2939637.

[29] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, "Manifold regularized sparse NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2815–2826, May 2013.

[30] 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.

[31] T. Zhang, B. Fang, Y. Y. Tang, G. He, and J. Wen, "Topology preserving non-negative matrix factorization for face recognition," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 574–584, Apr. 2008.

[32] B. Liu, Y. Li, and Z. Xu, "Manifold regularized matrix completion for multi-label learning with ADMM," *Neural Netw.*, vol. 101, pp. 57–67, May 2018.

[33] Q. Gu, C. Ding, and J. Han, "On trivial solution and scale transfer problems in graph regularized NMF," *Int. J. Control Automat.*, vol. 2011, pp. 1288–1293, Jun. 2011.

[34] S. Yang, L. Zhang, X. He, and Z. Yi, "Learning manifold structures with subspace segmentations," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2019.2895497.

[35] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[36] H. Gao, F. Nie, and H. Huan, "Local centroids structured non-negative matrix factorization," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 1905–1911.

[37] 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.

[38] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[39] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Sarasota FL, USA, Dec. 1994, pp. 138–142.

[40] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, Nov. 2017.

[41] Z. Ding and Y. Fu, "Robust multiview data analysis through collective low-rank subspace," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1986–1997, May 2018.

[42] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao, "Multi-view clustering via multi-manifold regularized non-negative matrix factorization," *Neural Netw.*, vol. 88, no. 4, pp. 74–89, Apr. 2017.

[43] X. Wang, T. Zhang, and X. Gao, "Multiview clustering based on non-negative matrix factorization and pairwise measurements," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3333–3346, Sep. 2019.

[44] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library," Dept. of Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.

[45] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[46] L. Lenc and P. Král, "Unconstrained facial images: Database for face recognition under real-world conditions," in *Proc. 14th Mexican Int. Conf. Artif. Intell. (MICAI)*, Cuernavaca, Mexico. Cham, Switzerland: Springer, 2015, pp. 25–31.

[47] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE. CVPR*, Jun./Jul. 2004, p. 178.

[48] S. Ameri, *Differential-Geometrical Methods in Statistics*. New York, NY, USA: Springer-Verlag, 1985, pp. 73–89.

QIAOQIN LI received the M.S. and Ph.D. degrees in computer applications from the University of Electronic Science and Technology of China, China, in 2000 and 2010, respectively. She is currently an Associate Professor with the School of Software and Information Engineering, University of Electronic Science and Technology of China. Her current research interests include artificial neural networks, machine learning, and the Internet of Things applications.

YONGGUO LIU received the B.S. degree in mechanical engineering from the Sichuan Institute of Light Industry and Chemical Technology, in 1997, the M.S. degree in mechanical engineering from Sichuan University, in 2000, and the Ph.D. degree in computer science from Chongqing University, in 2003. He finished his Postdoctoral Research with Shanghai Jiao Tong University, in 2005. He joined the University of Electronic Science and Technology of China, in 2005, and is currently a Full Professor with the School of Information and Software Engineering. His research interests include medical informatics and data mining.

SHANGMING YANG received the M.S. degree from St. Cloud State University, St. Cloud, MN, USA, in 2000, and the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, in 2009. He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His current research interests include artificial neural networks, data mining, and machine learning.

• • •