IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 15, 2019, accepted May 5, 2019, date of publication May 9, 2019, date of current version May 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2915947

Discriminative Multiview Nonnegative Matrix Factorization for Classification

WEIHUA OU[1], (Member, IEEE), JIANPING GOU[2], (Member, IEEE), QUAN ZHOU[3], (Member, IEEE), SHIMING GE[4], (Senior Member, IEEE), AND FEI LONG[5], (Member, IEEE)

[1]School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550001, China
[2]School of Computer Science and Communication Engineering, Jiangsu University, Jiangsu 212013, China
[3]National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[4]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[5]School of Electrical and Information Engineering, Guizhou Institute of Technology, Guiyang 550003, China

Corresponding authors: Weihua Ou (ouweihuahust@gmail.com) and Quan Zhou (quan.zhou@njupt.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 61762021, Grant 61263005, Grant 61876093, Grant 61502208, and Grant 61863006, in part by the Nature Science Foundation of Guizhou Province under Grant [2017]1130 and Grant [2017]5726-32, in part by the Nature Science Foundation of Jiangsu Province under Grant BK20181393 and Grant BK20150522, in part by the 2014 Ph.D. Recruitment Program of Guizhou Normal University, Key Disciplines of Guizhou Province under Grant ZDXK[2016]8, in part by the Foundation of Guizhou Educational Department under Grant KY[2016]027, and in part by the Basic Research Program Key Project of Guizhou Province under Grant20191416.

ABSTRACT Multiview nonnegative matrix has shown many promising applications in computer vision and pattern recognition. However, most existing works focus on view consistency and ignore discrimination. In this paper, we introduce a novel discriminative multiview nonnegative matrix (DMultiNMF) algorithm to learn discriminative and consistent representations for facilitating classification. In this algorithm, we apply discriminative patch alignment to enhance the local discrimination in each view and utilize the large margin principle to improve global discrimination. At the same time, we use a shared representation to propagate information among the multiple views to ensure consistency. Apart from that, we measure the reconstruction errors utilizing the correntropy-induced metric to improve the robustness. The experiments on face recognition, handwritten digit recognition, Xmedia, and Wikipedia multiview data sets demonstrate the advantages of the proposed method compared with other algorithms like single view using concatenated views and substantially better than other multiview nonnegative matrix factorization algorithms.

INDEX TERMS Multiview learning, nonnegative matrix factorization, patch alignment, consistent representation, classification.

I. INTRODUCTION

For many real applications [1], [2], the objects are represented by different features or the objects are captured under different views. As shown in Figure 1, a dog is represented by different features, and a web page is depicted simultaneously by text, image and video. These different features are referred as multiple views and exhibit heterogeneous properties [3], because the statistical distributions and even the dimensionality are different to each other. Those heterogeneous properties impose many challenges for the traditional methods, which can only deal with different views separately. However, since they describe the same object from various views, those multiple view data are complementary to each other. Traditional methods cannot utilize those complementary information. Multiview learning is a new kind of machine learning method, which can utilize the complementary information amongst multiple views to improve the performance [4]. By exploring complementary properties of different views, multiview learning has obtained better performance than that of the single view learning method [5].

Multiview learning mainly includes co-training, multiple kernel and subspace based method. Among them, subspace based method is an important one, which aims to learn a common low-dimensional subspace based on the assumption that all the inputs views are generated from the common latent subspace [6], [7]. The representative subspace based method includes canonical correlation analysis (CCA) [8], which finds the basis vectors of common subspace by maximizing the correlation of the projection of different views in the latent space. As an unsupervised method, the discovered subspaces utilizing CCA have weak discriminative ability

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

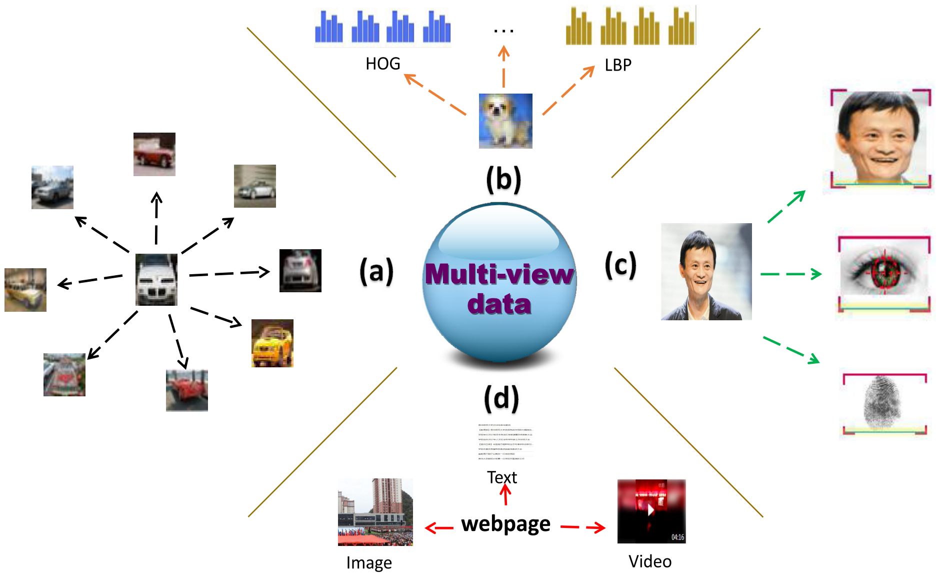


FIGURE 1. Examples of different kinds of multiview data. (a) different views of one object, (b) different kinds of features for one object, (c) different modules of one object, (d) different information sources for one topic.

due to the neglect of supervised information. To learn a discriminative common subspace, some supervised multiview learning methods have been proposed by incorporating label information, such as discriminative canonical correlation analysis (DCCA) [9], multiview fisher discriminant analysis (MFDA) [10], generalized multiview analysis (GMA) [4] and multiview discriminant analysis (MvDA) [11]. Benefitting from the supervised information, those discriminative methods usually outperform unsupervised methods.

However, in real applications, multiple view data are usually nonnegative, such as LBP [12], SIFT [13] and HOG [14] features. The existing approaches are not designed especially for nonnegative data. It is indispensable to study nonnegative multiview data representation learning method to deal with this kind of nonnegative multiview data. As an effective data representation method, nonnegative matrix factorization (NMF) [15]–[17] has been widely used in nonnegative features extraction [18], text analysis and face recognition [19]. However, standard NMF can only deal with single view data. Some extension of NMF to multiview have been developed in recent years. For example, Gao *et al.* [20] extended NMF to multiview clustering (MultiNMF), which encourages the data points from different views to be the common representation in latent low-dimensional space. Its results demonstrate that exploiting complementary information can improve the clustering performance [21]. Based on this ideas, many improvements have been made in [22]–[24], which mainly considers the local geometric structure in each view and even across views by exploring the intrinsic manifold. Furthermore, Kalayeh *et al.* [25] proposed a weighted MultiNMF considering the imbalance of dataset in real applications. Recently, Wang *et al.* [26] developed a diverse nonnegative matrix factorization (DiNMF), which considers the diversity of multiple views. To enhance the diversity and reduce redundancy among multiview representation, they designed a novel diversity regularization term. Many other related works can be referred to [27]–[32]. The main limitations of them are unsupervised methods, which are not suitable for labeled data.

In this paper, we propose a novel discriminative multiview nonnegative matrix factorization (DMultiNMF) to learn discriminative and consistent representation for facilitating classification. As shown in Figure 2, the proposed algorithm includes training and testing. During the training stage, view-conditional basis matrix $W^{(i)}$ for each view and the classifier $\vec{w}$ are obtained. Specifically, we apply the local patch alignment to enhance the local discrimination and utilize the large margin to improve the global discrimination. At the same time, a shared representation is adopted to propagate information among the multiple views to ensure consistency. Apart from that, we utilize correntropy-induced metric to measure the reconstruction to improve the robustness. At testing, the view-conditional basis matrix is used to get the common representation and the classifier is used to predict the class label of test data. Because of the correntropy-induced metric, the objective function is nonconvex. Based on the half-quadratic optimization theory, we formulate it into many local quadratic optimization sub-problems and solve them via alternative update rules. The main contributions are summarized as follows:

- global discrimination: the large margin principle is adopted to improve the global discriminative ability of the latent representation and classifiers.
- local discrimination: the discriminative locality alignment is utilized for each view to enhance the local discrimination.
- correntropy-induced metric (CIM): CIM-based measurement is adopted to measure the reconstruction

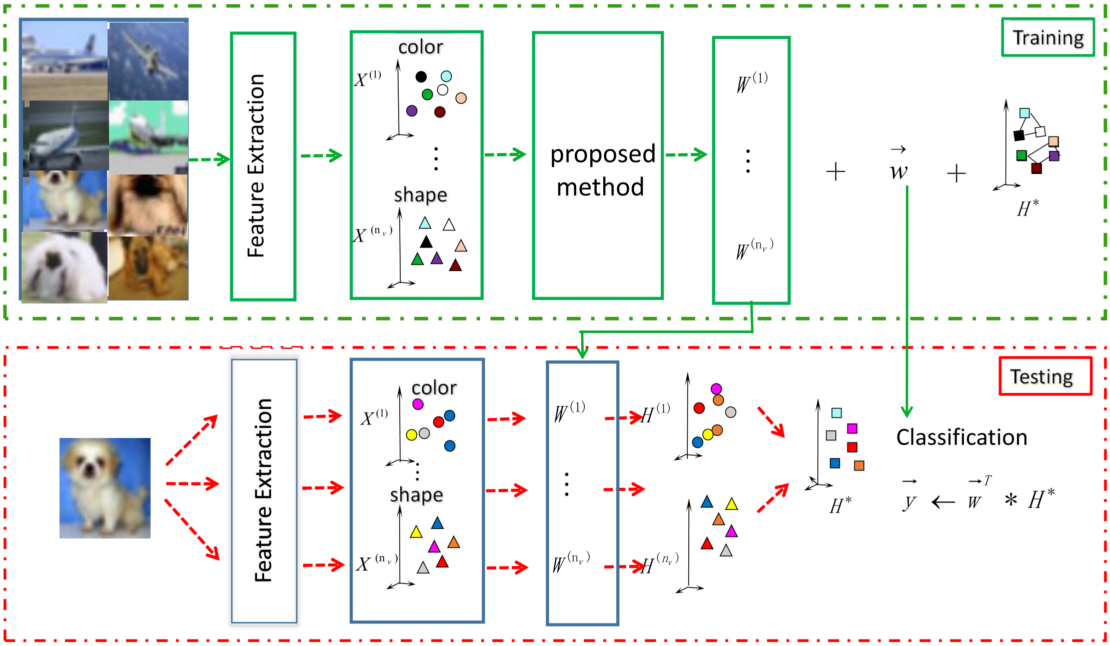


FIGURE 2. Flowchart of the proposed method, which includes training and testing stages. During the training stage, the view-conditional transformation matrix of each view $W^{(i)}$ and classifiers $\vec{w}$ are obtained. At the test step, the view-conditional transformation matrices $W^{(i)}$ are used to transform the original features into latent low-dimensional representation and the classifiers $\vec{w}$ are used to predict the labels of test samples.

errors, in which the weight for each entry can be learned adaptively.

- Experimental results demonstrate the proposed method obtained better performances than many existing methods.

This work is extension of the short version, which is published in the conference on security, pattern analysis, and cybernetics [33]. Compared to the short version, the locality discriminant alignment regularization term is designed, the optimization algorithm and complexity analysis are given, and more experimental results and details are presented.

The rest of paper is organized as follows. We introduce related works of basic NMF and multiview NMF in Section II. Then, we give the proposed method in Section III. After that, the details of optimization are shown in Sections IV. Finally, we report the experimental results in Section V and summarize this work in Section VI.

II. RELATED WORKS

In this section, we briefly introduce NMF and multi-view NMF.

A. NMF

Supposed $X \in \mathbb{R}_+^{m\times n}$ is a nonnegative data matrix, where m is the feature dimensionality and n is the data number, NMF approximates X by the production of two matrices W and H [16] with nonnegative constraints. The objective function of NMF can be formulated as follows [34]:

$$\min_{W,H} \| X - WH \|_F^2$$
$$s.t. \; W \geq 0, H \geq 0,$$

here $W \in \mathbb{R}_+^{m\times d}$ are the basis in the new space and $H \in \mathbb{R}_+^{d\times n}$ is the representation coefficient on basis W, d is the desired dimensionality. NMF has shown good performance in pattern recognition and computer vision [35], [36]. However, the objective function of NMF is not convex in both variables W and H. Fortunately, it is convex with respect to only W or H. It is proved that the local minimum can be found by the following multiplicative updates rules [37]:

$$W_{ik} \leftarrow W_{ik} \frac{(XH^T)_{ik}}{(WHH^T)_{ik}},$$

$$H_{kj} \leftarrow H_{kj} \frac{(W^T X)_{kj}}{(W^T WH)_{kj}}.$$

B. MULTI-VIEW NMF

Given a multi-view non-negative data matrix consisting of n samples with n_v different views as $\mathcal{D} = \{X^v = [\vec{x}_1^v, \cdots, \vec{x}_n^v] \in \mathbb{R}_+^{m_v\times n}\}_{v=1}^{n_v}$, multi-view NMF [20] factorizes each view X^v into $X^v \approx W^v H^v$ with view consistency constraint based on following objective function:

$$\sum_{v=1}^{n_v} \| X^v - W^v H^v \|_F^2 + \lambda_v \| H^v - H^* \|_F^2 \qquad (1)$$
$$s.t. \; W^v, H^v, H^* \geq 0, \| W_{*,r}^v \|_1 = 1, \forall 1 \leq r \leq d.$$

where $H^* = [\vec{h}_1^*, \vec{h}_2^*, \cdots, \vec{h}_n^*] \in \mathbb{R}_+^{d\times n}$ is the common latent representation, d is the dimensionality of the common latent space and m_v is the dimensionality of the v-th view.

III. PROPOSED METHOD

In this section, we present discriminative multi-view nonnegative matrix factorization for classification, which includes hinge loss, discriminative locality alignment for each view, reconstruction errors measurement and view consistency constraint.

A. LARGE MARGIN

Given the i-th sample from all the views and the associated class label as $X_i = \{\vec{x}_i^1, \cdots, \vec{x}_i^{n_v}, y_i\}$, we aim to learn the corresponding latent common representation $\vec{h}_i^*$. Inspired by the work in [38], we use a linear classifier $\vec{w}$ to predict the label of the i-th sample and adopt the hinge loss function to measure the prediction error as follow:

$$\ell\left(y_i, \vec{w}^T\vec{h}_i^*\right) = \max\left(0, 1 - y_i\vec{w}^T\vec{h}_i^*\right).$$

To learn an effective classifier and discriminative latent representation, we propose to minimize the classification errors of all the training samples:

$$\min_{\vec{w}} \left\{\sum_{i=1}^{n} \ell\left(y_i, \vec{w}^T\vec{h}_i^*\right) + \gamma \|\vec{w}\|_2^2\right\}$$

which can be reformulated as below:

$$\min_{\vec{w}} \left\{Tr\left(\max\left(0, I - \vec{y}\vec{w}^T H^*\right)\right) + \gamma \|\vec{w}\|_2^2\right\}, \tag{2}$$

where $Tr(.)$ denotes the matrix trace, I is an identity matrix, $\vec{y} = (y_1, y_2, \cdots, y_n)^T$, $H^* = [\vec{h}_1^*, \vec{h}_2^*, \cdots, \vec{h}_n^*] \in \mathbb{R}_+^{d\times n}$, $\|\vec{w}\|_2^2$ is the regularization term to avoid overfitting and γ is the regularization parameter.

B. DISCRIMINATIVE LOCALITY ALIGNMENT IN EACH VIEW

Patch alignment framework (PAF) [39] is a general dimensionality reduction method, which consists of part optimization and whole alignment.

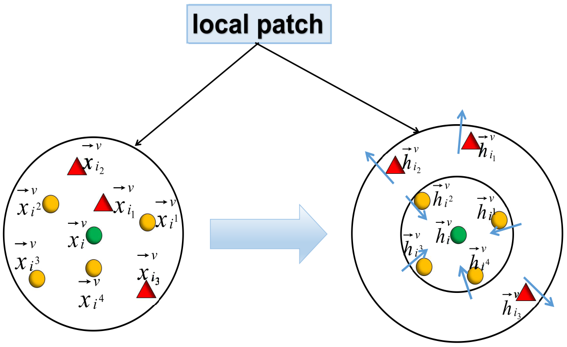


FIGURE 3. The illustration of discriminative locality alignment. For the given data point $\vec{x}_i^v$ marked with green circle, we select its neighbors marked with yellow circle from the same class, and select its neighbors marked with red triangle from different classes. After whole patch alignment, the data from the same class are as close as possible, while the data from different classes are as far as possible, as shown in the right part of the figure.

To construct the part optimization, we apply discriminative locality alignment to capture the local discriminative information in this paper. As shown in Figure 3, given data $\vec{x}_i^v$ in the v-th view, we select its k_1 nearest neighbors from the same class as $X_i^v(k_1) = \left[\vec{x}_{i^1}^v, \cdots, \vec{x}_{i^{k_1}}^v\right]$, and select k_2 nearest neighbors from the other classes as $X_i^v(k_2) = \left[\vec{x}_{i_1}^v, \cdots, \vec{x}_{i_{k_2}}^v\right]$. Based on those neighbors, we build the local patch for the sample $\vec{x}_i^v$ as $X_i^v = \left[\vec{x}_i^v, \vec{x}_{i^1}^v, \cdots, \vec{x}_{i^{k_1}}^v, \vec{x}_{i_1}^v, \cdots, \vec{x}_{i_{k_2}}^v\right] \in \mathbb{R}_+{}^{m_v\times(k_1+k_2+1)}$. For each patch, the corresponding representation in the low-dimensional space is denoted by $H_i^v = \left[\vec{h}_i^v, \vec{h}_{i^1}^v, \cdots, \vec{h}_{i^{k_1}}^v, \vec{h}_{i_1}^v, \cdots, \vec{h}_{i_{k_2}}^v\right] \in \mathbb{R}_+{}^{d\times(k_1+k_2+1)}$. To be discriminative, we maximize the average distances between samples from different classes and simultaneously minimize the average distances between samples from the same class. Thus, the part optimization can be defined as:

$$\ell_i^v = \min_{\vec{h}_i^v} \left(\sum_{j=1}^{k_1} \left\|\vec{h}_i^v - \vec{h}_{i^j}^v\right\|_2^2 - \sum_{p=1}^{k_2} \left\|\vec{h}_i^v - \vec{h}_{i_p}^v\right\|_2^2\right) \tag{3}$$

Define the coefficients vector $\vec{\varepsilon}_i^v$ as below

$$\vec{\varepsilon}_i^v = \Big[\overbrace{1, \cdots, 1,}^{k_1} \overbrace{-1, \cdots, -1,}^{k_2}\Big]^T,$$

we obtain

$$\begin{aligned}\ell_i^v &= \min_{\vec{h}_i^v} \left(\sum_{j=1}^{k_1+k_2} \left\|\vec{h}_{F_i^v(1)}^v - \vec{h}_{F_i^v(j+1)}^v\right\|_2^2 \left(\vec{\varepsilon}_i^v\right)_j\right)\\ &= \min_{H_i^v} Tr\left(H_i^v L_i^v \left(H_i^v\right)^T\right),\end{aligned} \tag{4}$$

where $F_i^v = [i, i^1, \cdots, i^{k_1}, i_1, \cdots, i_{k_2}]$ is the set of indices for measurements on the patch, and $L_i^v = \begin{bmatrix} \sum_{j=1}^{k_1+k_2}(\vec{\varepsilon}_i^v)_j & -(\vec{\varepsilon}_i^v)^T \\ -\vec{\varepsilon}_i^v & diag(\vec{\varepsilon}_i^v) \end{bmatrix}$. For the n data in v-th view, we construct n local patches $X_1^v, X_2^v, \cdots, X_n^v$ for them, and denote the corresponding low-dimensional representations for these local patches as $H_1^v, H_2^v, \cdots, H_n^v$, respectively. For whole alignment [40], we assume all these $H_i^v s$ are selected from the global coordinates $H^v = \left[\vec{h}_1^v, \vec{h}_2^v, \cdots, \vec{h}_n^v\right] \in \mathbb{R}_+{}^{d\times n}$, and $H_i^v = H^v S_i^v$, while S_i^v is defined below:

$$\left(S_i^v\right)_{pq} = \begin{cases} 1, & if\ p = F_i^v(q) \\ 0, & \text{otherwise} \end{cases}$$

By summing all the local patches, the global coordinate for the v-th view can be obtained as below:

$$\ell^v = \min_{H^v} Tr\left[H^v L^v \left(H^v\right)^T\right] \tag{5}$$

where $L^v = \sum_{i=1}^{n} S_i^v L_i^v \left(S_i^v\right)^T$.

C. CORRENTROPY-INDUCED METRIC

The correntropy-induced metric (CIM) [41] has been successful in many applications, such as face recognition [42], feature extraction [43] and nonnegative matrix factorization [44] because of its robustness, as shown in Figure 4. In this paper,

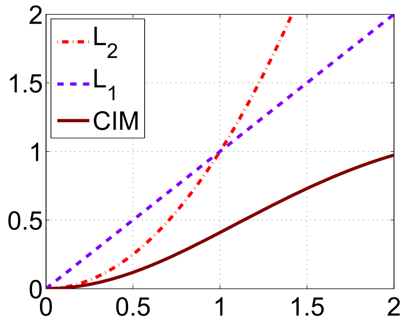


FIGURE 4. Comparisons of different errors measurement function. It is clear that the L_1-norm is linear to the errors, the L_2-norm is quadratic to the errors, while the correntropy-induced metric increases slower than the others, especially for large errors.

we adopt CIM to measure reconstruction errors for each view below:

$$\begin{aligned} &\min \mathcal{J}(X, WH) \\ &s.t. \quad W \geq 0, H \geq 0, \end{aligned} \tag{6}$$

where $\mathcal{J}(X, WH) = \sum_{i,j}\left(1 - g_\sigma\left(X_{ij} - (WH)_{i,j}\right)\right)$, and $g_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-x^2}{2\sigma^2}\right)$ is a gaussian function.

D. OBJECTIVE FUNCTION

Considering (2) (5) (6) and view consistency of different views simultaneously, we obtain the objective function of discriminative multiview NMF for classification as below:

$$\begin{aligned} \min\Bigg\{&\sum_{v=1}^{n_v}\left[\mathcal{J}(X^v, W^vH^v) + \lambda\|H^v - H^*\|_F^2 + \alpha\ell^v\right] \\ &+\beta Tr\left[\max\left(0, I - \vec{y}\vec{w}^TH^*\right)\right] + \gamma\|\vec{w}\|_2^2\Bigg\}, \\ \text{s.t.}\ \ & W^v, H^v, H^* \geq 0, \|W^v_{*,r}\|_1 = 1, r = 1, \cdots, d. \end{aligned} \tag{7}$$

where $\lambda, \alpha, \beta, \gamma$ are the regularization parameters.

IV. OPTIMIZATION

The objective function (7) is difficult to optimize because the presence of non-convex term in the objective function. Fortunately, it can be solved by the half-quadratic technique [45], which reformulates the non-convex term as an augmented objective function in an enlarged parameter space. According the property of convex conjugate function [46] and the half-quadratic theory [45], [47], we obtain augmented objective function as below:

$$\begin{aligned} \min\Bigg\{&\sum_{v=1}^{n_v}\Bigg\{\sum\left[P^v_{i,j}\left(X^v_{i,j} - (W^vH^v)_{i,j}\right)^2 + \varphi(P^v_{i,j})\right] \\ &+\lambda\|H^v - H^*\|_F^2 + \alpha Tr\left[H^vL^v\left(H^v\right)^T\right]\Bigg\} \\ &+\beta Tr\left(\max\left(0, I - \vec{y}\vec{w}^TH^*\right)\right) + \gamma\|\vec{w}\|_2^2\Bigg\}, \\ \text{s.t.}\ & W^v, H^v, H^*, P^v \geq 0, \|W^v_{*,r}\|_1 = 1, r = 1, \cdots, d. \end{aligned} \tag{8}$$

where $P_{i,j} = g_\sigma(X_{i,j} - (WH)_{i,j})$ [43] is the associated auxiliary variable and $\varphi(.)$ is the conjugate function of $g_\sigma(x)$. Similar to [20], we formulate the equation constraints imposing on the basis matrix W^v into following diagonal matrix:

$$Q^v = Diag\left(\sum_{i=1}^{m_v}W^v_{i,1}, \sum_{i=1}^{m_v}W^v_{i,2}, \cdots, \sum_{i=1}^{m_v}W^v_{i,d}\right). \tag{9}$$

Thus, problem (8) can be reformulated as:

$$\begin{aligned} \min\Bigg\{&\sum_{v=1}^{n_v}\Bigg\{\sum\left[P^v_{i,j}\left(X^v_{i,j} - (W^vH^v)_{i,j}\right)^2 + \varphi(P^v_{i,j})\right] \\ &+\lambda\|Q^vH^v - H^*\|_F^2 + \alpha Tr\left[H^vL^v\left(H^v\right)^T\right]\Bigg\} \\ &+\beta Tr\left(\max\left(0, I - \vec{y}\vec{w}^TH^*\right)\right) + \gamma\|\vec{w}\|_2^2\Bigg\}, \\ \text{s.t.}\ \ & W^v, H^v, H^*, P^v \geq 0. \end{aligned} \tag{10}$$

A. OPTIMIZE P^v FOR GIVEN W^v, H^v, AND H^*

Given W^v, H^v, H^* and $\vec{w}$, problem (10) can be solved separately with respect to $P^v_{i,j}$ for each view v, and the solution is given as below:

$$P^v_{i,j} = g_\sigma(X^v_{i,j} - (W^vH^v)_{i,j}), \quad 1 \leq v \leq n_v. \tag{11}$$

B. OPTIMIZE W^v AND H^v FOR GIVEN P^v AND H^*

When H^* is fixed, for each given v, the computation of W^v and H^v do not depend on $W^{\bar{v}}$ and $H^{\bar{v}}$ for any $\bar{v} \neq v$. Thus, we use X, W, H and Q to represent X^v, W^v, H^v and Q^v for the brevity. Therefore, we obtain the function with respect to W^v and H^v as following:

$$\begin{aligned} \min_{W,H}\Bigg\{&\sum\left[P_{i,j}\left(X_{i,j} - (WH)_{i,j}\right)^2\right] + \lambda\|QH - H^*\|_F^2 \\ &+\alpha Tr\left[HLH^T\right]\Bigg\}. \ \text{s.t.}\ W, H \geq 0, \end{aligned} \tag{12}$$

1) OPTIMIZE W FOR GIVEN H, P AND H^*

Given H, the problem (12) can be solved as follows by optimizing each row of W separately,

$$\begin{aligned} \mathcal{L}(W, \Theta) = &\sum_{i=1}^{m}\left(X_{i,*} - W_{i,*}H\right)A_i\left(X_{i,*} - W_{i,*}H\right)^T \\ &+\lambda U + \alpha Tr\left(HLH^T\right) + Tr\left(\Theta^TW\right), \end{aligned} \tag{13}$$

where $A_i = diag(P_{i,*}) \in \mathbb{R}^{n\times n}$, $\Theta = [\theta_{i,k}] \in \mathbb{R}^{m\times d}$ is the Lagrange multipliers for the non-negative constraints $W \geq 0$, and $U = \text{Tr}\left(QH(QH)^T - 2QH(H^*)^T\right)$ is the constraint term in $\|QH - H^*\|_F^2$. The partial derivatives of $\mathcal{L}(W, \Theta)$ with respect to $W_{i,k}$ is presented below:

$$\frac{\partial\mathcal{L}(W, \Theta)}{\partial W_{i,k}} = 2\left(W_{i,*}HA_iH^T - X_{i,*}A_iH^T\right)_k + \lambda S_{i,k} + \theta_{i,k}. \tag{14}$$

where $S_{i,k}$ is the derivative of U and is calculated as below:

$$S_{i,k} = \frac{\partial U}{\partial W_{i,k}} \tag{15}$$

$$= \frac{\partial \left\{ \mathrm{Tr}(QH(QH)^T) - 2QH(H^*)^T \right\}}{\partial W_{i,k}}$$

$$= 2\left(\sum_{l=1}^{m} W_{l,k} \left(\sum_{j=1}^{n} H_{i,j} H_{k,j} \right) - \sum_{j=1}^{n} H_{i,j} H^*_{k,j} \right).$$

Setting (15) to zero and utilizing the KKT conditions $\theta_{i,k} W_{i,k} = 0$, we can get following equation for $W_{i,k}$,

$$\left(2\left(W_{i,*} H A_i H^T - X_{i,*} A_i H^T \right)_k + \lambda S_{i,k} \right) W_{i,k} = 0, \tag{16}$$

This equation leads to the update rule below for $W_{i,k}$:

$$W_{i,k} \leftarrow \tag{17}$$

$$\frac{W_{i,k} \left[\left(X_{i,*} A_i H^T \right)_k + \lambda \sum_{j=1}^{n} H_{i,j} H^*_{k,j} \right]}{\left(W_{i,*} H A_i H^T \right)_k + \lambda \sum_{l=1}^{m} W_{l,k} \left(\sum_{j=1}^{n} H_{i,j} H_{k,j} \right)}.$$

2) OPTIMIZE H FOR GIVEN W, P AND H^*

For the computation of H, we first normalize the column vectors of W using Q as following:

$$W \leftarrow WQ^{-1}, H \leftarrow QH, \tag{18}$$

Thus, the problem (12) is equivalent to minimize following objective function:

$$\mathcal{L}(H, \Psi) = \sum_{j=1}^{n} \left\{ \left(X_{*,j} - WH_{*,j} \right)^T B_j \left(X_{*,j} - WH_{*,j} \right) \right\}$$

$$+ \lambda \|H - H^*\|_F^2 + \alpha Tr\left(HLH^T \right) + \mathrm{Tr}(\Psi^T H), \tag{19}$$

where $B_j = diag(P_{*,j}) \in \mathbb{R}^{m \times m}$, $\Psi = [\psi_{k,j}] \in \mathbb{R}^{d \times n}$ is the Lagrange multipliers for the non-negative constraints $H \geq 0$. The partial derivatives of $\mathcal{L}(H, \Psi)$ with respect to $H_{k,j}$ is presented below:

$$\frac{\partial \mathcal{L}(H, \Psi)}{\partial H_{k,j}} = -2\left(W^T B_j X_{*,j} \right)_k + 2\left(W^T B_j W H_{*,j} \right)_k$$

$$+ 2\lambda(H - H^*)_{k,j} + 2\alpha \, (HL)_{k,j} + \psi_{k,j}. \tag{20}$$

Setting (20) to zero and utilizing the KKT conditions $\psi_{k,j} H_{k,j} = 0$, we can get following equation for $H_{k,j}$,

$$\left\{ -2\left(W^T B_j X_{*,j} \right)_k + 2\left(W^T B_j W H_{*,j} \right)_k \right.$$

$$\left. + 2\lambda(H - H^*)_{kj} + 2\alpha \, (HL)_{k,j} \right\} H_{k,j} = 0, \tag{21}$$

where $L = L^+ - L^-, L^+_{i,j} = \left(|L_{i,j}| + L_{i,j} \right)/2, L^-_{i,j} = \left(|L_{i,j}| - L_{i,j} \right)/2$. Equation (21) leads to the update rule for $H_{k,j}$:

$$H_{k,j} \leftarrow H_{k,j} \frac{\left(W^T B_j X_{*,j} + \alpha \left(HL^- \right)_{*,j} + \lambda H^*_{*,j} \right)_k}{\left(W^T B_j W H_{*,j} + \alpha \left(HL^+ \right)_{*,j} + \lambda H_{*,j} \right)_k}$$

$$H_{k,j} \leftarrow H_{k,j} \frac{\left(W^T (X \odot P) + \alpha \left(HL^- \right)_{*,j} + \lambda H^* \right)_{k,j}}{\left(W^T (WH \odot P) + \alpha \left(HL^- \right)_{*,j} + \lambda H \right)_{k,j}}. \tag{22}$$

C. OPTIMIZE $\vec{w}$ FOR GIVEN W^v, H^v, H^* AND P^v

For $Tr\left[\max(0, I - \vec{y}\vec{w}^T H^*)) \right]$, we introduce an auxiliary matrix Z to reformulate it into $Tr\left(Z\left(I - \vec{y}\vec{w}^T H^* \right) \right)$, where $Z = diag(z_1, z_2, \cdots, z_n)$, and

$$z_i = \begin{cases} 1, & \text{if } 1 - y_i \vec{w}^T h_i^* > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Taking the derivative of the objective function with respect to $\vec{w}$, we obtain:

$$\frac{\partial \left[\beta Tr\left(Z\left(I - \vec{y}\vec{w}^T H^* \right) \right) + \gamma \|\vec{w}\|_2^2 \right]}{\partial \vec{w}} = -\beta H^* Z \vec{y} + 2 * \gamma \vec{w}, \tag{24}$$

Setting above equation to zero, we get the closed solution for $\vec{w}$:

$$\vec{w} = \beta H^* Z \vec{y} / (2 * \gamma). \tag{25}$$

D. OPTIMIZE H^* FOR GIVEN W^v, H^v, $\vec{w}$ AND P^v

Taking the derivative of the objective function with respect to H^* and set it to zero, we obtain

$$\frac{\partial O}{\partial H^*} = \sum_{v=1}^{n_v} \lambda(-2Q^v H^v + 2H^*) - \beta \vec{w}\vec{y}^T Z^T = 0, \tag{26}$$

where

$$O = \sum_{v=1}^{n_v} \lambda \|Q^v H^v - H^*\|_F^2 + \beta Tr\left(Z\left(I - \vec{y}\vec{w}^T H^* \right) \right).$$

Solving equation (26), we get the closed solution for H^*:

$$H^* = \frac{\sum_{v=1}^{n_v} \lambda Q^v H^v + \beta \vec{w}\vec{y}^T Z^T / 2}{n_v \lambda}. \tag{27}$$

From equation (27), it is obvious that the H^* is the linear combination of different view's representation and the classification errors. The whole procedure is summarized in algorithm 1. After the training, we utilize the view-conditional transformation matrix W^v to learn the common representation for test samples. The test algorithm is shown in algorithm 2.

E. COMPLEXITY ANALYSIS

We adopt standard NMF as baseline to analyze the complexity of the proposed method. As we known, the complexity

Algorithm 1 Discriminative Multi-View NMF (DMultiNMF) for Classification (Training)

```
Input: Training data $\mathcal{D} = \{X^1, \cdots, X^{n_v}\} \in \mathbb{R}_+^{m_v \times n}$, $\vec{y} = (y_1, \cdots, y_n)^T$, parameters $\lambda, \alpha, \beta, \gamma, k_1, k_2, d$.
Output: $\{W^1, \cdots, W^{n_v}\}$, $\{H^1, \cdots, H^{n_v}\}$, $H^*$, and $\vec{w}$.
Normalize each view $X^v$ such that $\|X^v_{*,i}\|_1 = 1$.
Initialize randomly $\vec{w}$, $W^v$, $H^v$ and $H^*$.
repeat
    for $v = 1$ to $n_v$ do
        repeat
            $\sigma_v^2 = \frac{1}{2m_v n} \sum_{i=1}^{m_v} \sum_{j=1}^{n} (X^v_{i,j} - (W^v H^v)_{i,j})^2$;
            Update $P^v$ by $P^v_{i,j} = g_{\sigma_v}(X^v_{i,j} - (W^v H^v)_{i,j})$;
            Fixed $H^*$, $H^v$, $P^v$and update $W^v$ by Eq.(17);
            Fixed $W^v$, $H^*$, $P^v$ and update $H^v$ by Eq.(22);
        until The stop condition is satisfied.
    end for
    Fixed $H^*$, and update $\vec{w}$ by Eq.(25);
    Fixed $\vec{w}$, $H^v$, and update $H^*$ by Eq.(27);
until The stop condition is satisfied.
```

Algorithm 2 Discriminative Multi-View NMF (DMultiNMF) for Classification (Testing)

```
Input: Test data $\mathcal{D} = \{X^v = \left[\vec{x}^v_1, \cdots, \vec{x}^v_n\right] \in \mathbb{R}_+^{m_v \times n}\}_{v=1}^{n_v}$, $\{W^1, \cdots, W^{n_v}\}$, $\vec{w}$, and $\lambda$.
Output: class label = $\arg\max \left\{\vec{w}^T H^*\right\}$
Normalize each view $X^v$ such that $\|X^v_{*,i}\|_1 = 1$.
Initialize randomly $H^v$ and $H^*$.
repeat
    for $v = 1$ to $n_v$ do
        repeat
            $\sigma_v^2 = \frac{1}{2m_v n} \sum_{i=1}^{m_v} \sum_{j=1}^{n} (X^v_{i,j} - (W^v H^v)_{i,j})^2$;
            Update $P^v$ by $P^v_{i,j} = g_{\sigma_v}(X^v_{i,j} - (W^v H^v)_{i,j})$;
            Update $H^v$ by Eq.(22) with $\alpha = 0$;
        until The stop condition is satisfied.
    end for
    Update $H^*$ by Eq.(27) with $\beta = 0$;
until The stop condition is satisfied.
```

of NMF's multiplicative update rules in each iteration is $O(mdn)$, where m is the feature dimensionality, d is the low-dimensional subspace dimensionality and n is the sample number. In our method, for each view, the complexity of each updating is $O(m_v dn)$, which is the same as the NMF. Denote the iterations of inner loop as t_{in}, then the complexity is $O(t_{in} n_v mnd)$, where n_v is the view number. Assume the iterations of outer loop is t_{out}, then the overall time complexity for the proposed method is $O(t_{out} t_{in} n_v mnd)$. Besides this multiplicative updates, the algorithm also needs $O(n^2 m)$ to construct the local patch during the training stage. In all, the overall running time of the proposed method in test stage is linear with respect to the reduced dimensionality and view number.

V. EXPERIMENTS

In this section, we evaluate our proposed method on four multiview datasets and compare to following representative methods:

- **Single view.** It runs each view separately using NMF. We report the best, the worst and the average results denoted by BSV, WSV and AVG respectively.
- **Feature concatenation (FC).** It runs NMF directly on the concatenated features from all views.
- **MultiNMF [20].** It requires all the representation of different views to approximate a latent low-dimensional representation.
- **MultiCIM.** It is the same as Multi-NMF except for the use of CIM in the reconstruction errors measurement.
- **MultiRNMF [23].** It takes the local geometric structure of each view into consideration, and utilizes locally linear embedding to represent the local geometric structure.
- **DMultiNMF-1.** It considers view consistency and maximize the margin, which doesn't consider the patch alignment for each view.
- **DMultiNMF-2.** It is the extension of DMultiNMF-1 by exploring the locality discriminative information in each view.

A. DATASETS AND EVALUATION

- **ORL dataset.** The ORL dataset consists of 40 subjects and 10 different images for each subject with totally 400 images. The images are grayscale and have been normalized to 32 × 32 pixels. The first view is the raw pixel values, i.e., $X^1 \in \mathbb{R}_+^{1024 \times 400}$, and the second view is the local binary pattern, i.e., $X^2 \in \mathbb{R}_+^{59 \times 400}$.
- **Handwritten digit dataset (UCI):** This handwritten digits (0-9) data is from the UCI repository, which consists of 2000 samples of 10 different object classes, with view-1 being the 76 Fourier coefficients of the character shapes, view-2 being the 240 pixel averages in 2 × 3 windows.
- **XMedia dataset [48]:** This dataset consists of 5,000 texts, 5,000 images, 500 videos, 1,000 audio clips and 500 3D models. The dataset is evenly split into 20 categories, which are insect, bird, wind, dog, tiger, explosion, elephant, flute, airplane, drum, train, laughter, wolf, thunder, horse, autobike, gun, stream, piano and violin. We use text as the view 1 and image as the view 2. All the features are obtained from the author provided from their website[1].
- **Wikipedia dataset [49]:** This dataset is a document corpus with paired texts and images. It collected 2700 articles since 2009. Each featured article is categorized into one of 29 categories, which were assigned to both the text and image components of each article. The 10 most populated categories are considered because

[1] http://www.icst.pku.edu.cn/mipl/

TABLE 1. Summary of the wikipedia dataset.

Category	Training	Testing	Total
Art & architecture	138	34	172
Biology	272	88	360
Geography&places	244	96	340
History	248	85	333
Literature&theatre	202	65	267
Media	178	58	236
Music	186	51	237
Royalty & nobility	144	41	185
Sport & recreation	214	71	285
Warfare	347	104	451

TABLE 2. Parameters setting.

parameters	α	β	γ	λ
ORL	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Digit	10^{-3}	10^{-3}	10^{-2}	10^{-1}
XMedia	10^{-4}	10^{-4}	10^{-2}	10^{-4}
Wikipedia	10^{-4}	10^{-4}	10^{-3}	10^{-3}

some of the categories are very scarce. By removing the sections without any image, the final corpus contains a total of 2866 documents. These text-image pairs annotated with a label from the vocabulary of 10 semantic classes. A random split was used to produce a training set of 2173 documents, and a test set of 693 documents, as summarized in Table 1. We use text as the view 1 and image as the view 2.

Evaluation: We use classification accuracy (CA) [5] to measure the classification performance as follows:

$$CA = \frac{\text{number of correctly classified data points}}{\text{number of test data points}}$$

B. EXPERIMENTAL RESULTS

Some parameters are involved in the proposed algorithm, $k_1 = k_2 = 5, d = 10$ and all the other parameters are set empirically in our experiments as Table 2.

1) RESULTS ON ORL AND HANDWRITTEN DIGITS DATASETS

In this section, we conduct classification experiments on ORL dataset and handwritten digit dataset, respectively. On both dataset, we randomly select 80% data for training from each class and select the rest as testing. All the methods are implemented 15 times independently on each dataset, then the average performances are reported. As shown in Figures 5 and 6, it is obvious that the performances of multiview NMF algorithm are much better than that of single view NMF. Among all the multiview NMF, the DMultiNMF-1 and DMultiNMF-2 are better than that of the other multiview NMF methods. The classification accuracy of the proposed method on two datasets is more than 91% and 95%, respectively. This is mainly due to supervised information is utilized in those two methods. Compared to DMultiNMF-1, the DMultiNMF-2 is slightly better, which demonstrates the discriminative locality alignment is effective.

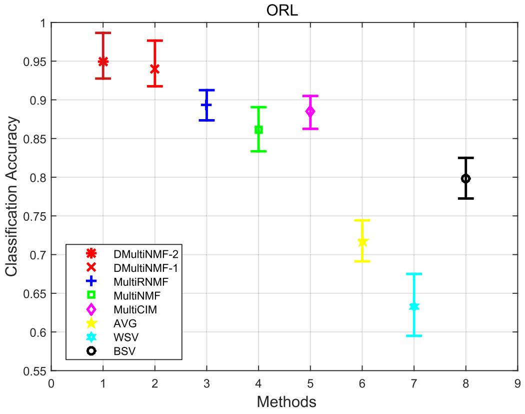


FIGURE 5. Classification results of different algorithms on ORL dataset.

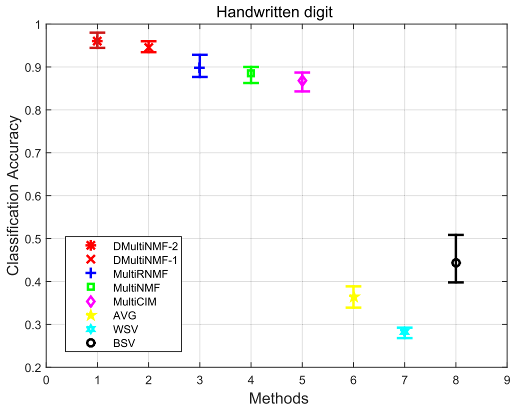


FIGURE 6. Classification results of different algorithms on handwritten digits dataset.

2) RESULTS ON XMEDIA AND WIKIPEDIA DATASETS

In this subsection, we consider two datasets: XMedia dataset, wikipedia dataset. For the XMedia dataset, we randomly select 80% data for training from each class and select the rest as testing. All the methods are implemented 15 times independently on each dataset, then the average performances are reported. The average classification results are shown in Table 3. It is clear that the proposed method gets better classification results compared to other NMF-based algorithms on those two datasets.

3) THE INFLUENCE OF PARAMETERS

We conduct experiments on ORL and handwritten digit (UCI) datasets to discuss the influence of parameter d. As shown in Figure 7, when the dimension d is less than 70, the classification performance of the proposed method keep stable and it can reach 95%. However, the classification accuracy fall downs when the dimension d is more than 70. In a word, the proposed method is stable in a large range between 10 and 70 for the dimension d. Furthermore, we show the confusion matrix on the handwritten digit (UCI) dataset in the Figure 8, where the value of d is 30. As we all know that the bigger value of d will increase the calculation cost. Therefore, we set d to 10 in the experiments.

TABLE 3. Classification results on XMedia and Wikipedia datasets.

Methods	BSV	WSV	FC	MultiNMF	MultiCIM	MultiRNMF	DMutliNMF-1	**DMultiNMF-2**
XMedia	0.964±0.01	0.933±0.02	0.947±0.01	0.955±0.01	0.963±0.01	0.972±0.01	0.987±0.01	**0.993±0.01**
Wikipedia	0.950±0.01	0.914±0.01	0.919±0.01	0.921±0.01	0.953±0.01	0.956±0.01	0.960±0.01	**0.968±0.01**

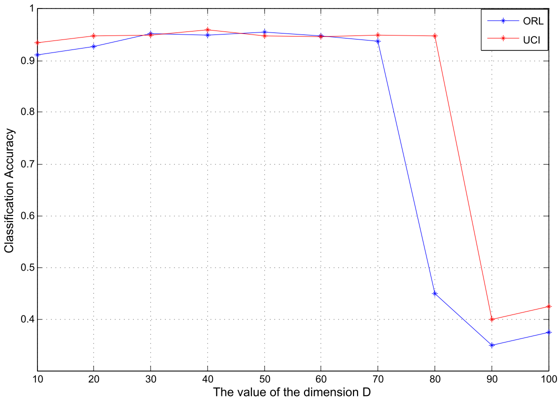
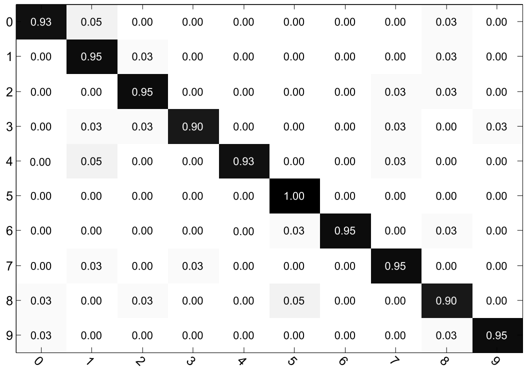


FIGURE 7. Classification results with different *d* on ORL and handwritten digit (UCI) dataset.



	0	1	2	3	4	5	6	7	8	9
0	0.93	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
1	0.00	0.95	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00
2	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.03	0.03	0.00
3	0.00	0.03	0.03	0.90	0.00	0.00	0.00	0.03	0.00	0.03
4	0.00	0.05	0.00	0.00	0.93	0.00	0.00	0.03	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.03	0.95	0.00	0.03	0.00
7	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.95	0.00	0.00
8	0.03	0.00	0.03	0.00	0.00	0.05	0.00	0.00	0.90	0.00
9	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.95

FIGURE 8. The confusion matrix obtained on handwritten digit (UCI) dataset.

VI. CONCLUSION

In this paper, we propose a discriminative multiview nonnegative matrix factorization approach for classification. The locality discriminant information in each view and sample label information are considered simultaneously. The multiplicative update rules are developed and the complexity analysis of the algorithm is also given. The experimental results demonstrate the multiview methods are much better than the traditional single view NMF method. Among all the multiview NMF methods, the experimental results show that exploring locality discriminant information and supervised information can improve the performances. The disadvantage of this method is selection of hyper parameters. In the future, we will study it deeply and consider the semi-supervision situation.

REFERENCES

[1] X. You, Q. Li, D. Tao, W. Ou, and M. Gong, "Local metric learning for exemplar-based object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1265–1276, Aug. 2014.

[2] X. Li, Q. Liu, Z. He, H. Wang, C. Zhang, and W.-S. Chen, "A multi-view model for visual tracking via correlation filters," *Knowl.-Based Syst.*, vol. 113, pp. 88–99, Dec. 2016.

[3] C. Xu, D. Tao, and C. Xu. (2013). "A survey on multi-view learning." [Online]. Available: https://arxiv.org/abs/1304.5634

[4] A. Sharma, A. Kumar, H. Daume, III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2160–2167.

[5] Q. Wang, H. Lv, J. Yue, and E. Mitchell, "Supervised multiview learning based on simultaneous learning of multiview intact and single view classifier," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 2293–2301, 2016.

[6] Z. Wang, X. Sun, L. Sun, and Y. Huang, "Multiview discriminative geometry preserving projection for image classification," *Sci. World J.*, vol. 2014, no. 11, 2014, Art. no. 924090.

[7] S. Yi, Z. He, Y.-M. Cheung, and W.-S. Chen, "Unified sparse subspace learning via self-contained regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2537–2550, Oct. 2017.

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[9] T.-K. Kim, J. Kittler, and R. Cipolla, "Learning discriminative canonical correlations for object recognition with image sets," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 3953. Berlin, Germany: 2006, pp. 251–262.

[10] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Constructing nonlinear discriminants from multiple data views," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2010, pp. 328–343.

[11] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2012.

[12] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 404–420.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[15] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[16] C. Ding and D. Kong, "Nonnegative matrix factorization using a robust error function," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 2033–2036.

[17] T. Liu, M. Gong, and D. Tao, "Large-cone nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 9, pp. 2129–2142, Sep. 2017.

[18] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "MahNMF: Manhattan non-negative matrix factorization," *J. Mach. Learn. Res.*, vol. 1, no. 5, pp. 11–43, 2012.

[19] W. Ou, G. Li, S. Yu, G. Xie, F. Ren, and Y. Tang, "Robust discriminative nonnegative patch alignment for occluded face recognition," in *Proc. 22nd Int. Conf. Neural Inf. Process. (ICONIP)*, Istanbul, Turkey, Nov. 2015, pp. 207–215.

[20] J. Gao, J. Han, J. Liu, and C. Wang, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. 13th SIAM Int. Conf. Data Mining*, 2013, pp. 252–260.

[21] W. F. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview Hessian discriminative sparse coding for image annotation," *Comput. Vis. Image Understand.*, vol. 118, pp. 50–60, Jan. 2014.

[22] X. Zhang, L. Zhao, L. Zong, H. Yu, and X. Liu, "Multi-view clustering via multi-manifold regularized nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 1103–1108.

[23] W. Ou, S. Yu, G. Li, J. Lu, K. Zhang, and G. Xie, "Multi-view non-negative matrix factorization by patch alignment framework with view consistency," *Neurocomputing*, vol. 204, pp. 116–124, Sep. 2016.

[24] D. Hidru and A. Goldenberg. (2014). "EquiNMF: Graph regularized multiview nonnegative matrix factorization." [Online]. Available: https://arxiv.org/abs/1409.4018

[25] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 184–191.

[26] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang, "Diverse non-negative matrix factorization for multiview data representation," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2620–2632, Sep. 2017.

[27] X. Wang, T. Zhang, and X. Gao, "Multiview clustering based on non-negative matrix factorization and pairwise measurements," *IEEE Trans. Cybern.*, to be published.

[28] G. Li and W. Ou, "Pairwise probabilistic matrix factorization for implicit feedback collaborative filtering," *Neurocomputing*, vol. 204, pp. 17–25, Sep. 2016.

[29] X. You, W. Ou, C. L. P. Chen, Q. Li, Z. Zhu, and Y. Tang, "Robust nonnegative patch alignment for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2760–2774, Nov. 2015.

[30] S. Yi, Z. Lai, Z. He, Y.-M. Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognit.*, vol. 61, pp. 524–536, Jan. 2017.

[31] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.

[32] J. Xu, S. Yu, X. You, M. Leng, X.-Y. Jing, and C. L. P. Chen, "Multiview hybrid embedding: A divide-and-conquer approach," *IEEE Trans. Cybern.*, to be published.

[33] F. Long, W. Ou, K. Zhang, Y. Tan, Y. Xue, and G. Li, "Discriminative multiview nonnegative matrix factorization with large margin for image classification," in *Proc. Int. Conf. Secur., Pattern Anal. (SPAC)*, Shenzhen, China, Dec. 2017, pp. 37–42.

[34] 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.

[35] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC Bioinform.*, vol. 14, no. 1, p. 107, 2013.

[36] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. ACM 26th Annu. Int. SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 267–273.

[37] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 13, no. 6, pp. 556–562.

[38] X. Fan and K. Tang, "Enhanced maximum AUC linear classifier," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Yantai, China, Aug. 2010, pp. 1540–1544.

[39] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[40] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Non-negative patch alignment framework," *IEEE Trans. Neural Netw.*, vol. 22, no. 8, pp. 1218–1230, Aug. 2011.

[41] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[42] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[43] X.-T. Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *Proc. ACM ICML*, 2009, pp. 1193–1200.

[44] L. Du, X. Li, and Y.-D. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization," in *Proc. IEEE ICDM*, Dec. 2012, pp. 201–210.

[45] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[46] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[47] M. Nikolova and R. H. Chan, "The equivalence of half-quadratic minimization and the gradient linearization iteration," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1623–1627, Jun. 2007.

[48] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[49] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia*, 2010, pp. 251–260.

WEIHUA OU (M'14) received the M.S. degree in mathematics from Southeast University, Nanjing, China, in 2006, and the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology (HUST), China, in 2014, respectively. From 2016 to 2017, he was a Postdoctoral Researcher with the University of Technology Sydney with Prof. D. Tao. He is currently an Associate Professor with the School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China. He has authored more than 40 articles, including IEEE TNNLS, TCSVT, PR, and *Neurocomputing*. His current research interests include sparse (low-rank) representation, multiview learning, cross-modal retrieval, image processing, and computer vision.

JIANPING GOU received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He is currently an Associate Professor with the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, Jiangsu, China. His current research interests include pattern recognition and machine learning.

QUAN ZHOU (M'13) received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2013. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications. He has published more than 30 articles in top journals (e.g., IEEE TIP, PR, EL, and *Sensors*) and conference (e.g., ICIP, ICASSP, ACCV, and ICPR) in image processing and computer vision. His research topics include image labeling and scene understanding, visual attention and saliency detection, and face identification and recognition. He currently serves as a Reviewer for a series of SCI journals, including the IEEE TIP, IEEE TSP, IEEE TC, IEEE TCSVT, and *Neurocomputing*.

SHIMING GE received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei. He was a Senior Researcher with Shanda Innovations, a Researcher with Samsung Electronics, and a Research Fellow with the Nokia Research Center. His home page is http://www.escience.cn/people/geshiming. His research fields include AI security, deep learning, and computer vision. He is currently an Associate Professor and a Doctoral Supervisor with the Institute of Information Engineering, Chinese Academy of Sciences.

FEI LONG (M'10) received the Ph.D. degree from Southeast University, in 2006. He is currently a Professor with the School of Electrical and Information Engineering, Guizhou Institute of Technology. His research interests include image processing, intelligent control, and data mining.