IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 24, 2019, accepted September 9, 2019, date of publication September 16, 2019,
date of current version September 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941542

Label Correlation Guided Deep Multi-View Image Annotation

ZHE XUE[1], JUNPING DU[1], MIN ZUO[2], GUORONG LI[3], AND QINGMING HUANG[3], (Fellow, IEEE)

[1]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]National Engineering Laboratory for Agri-product Quality Traceability, Beijing Technology and Business University, Beijing 100048, China
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Min Zuo (zuomin@btbu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802028, Grant 61772083, Grant 61532006, and Grant 61877006, in part by the Open Project Program of National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University (BTBU), under Grant AQT-2019-YB9, and in part by the Science and Technology Major Project of Guangxi under Grant GuikeAA18118054.

ABSTRACT Automatic image annotation is an important technique which has been widely applied in many fields such as social network image analysis and retrieval, face recognition and so on. Multi-view image annotation aims to utilize multi-view complementary information to achieve more effective annotation results. However, the existing multi-view image annotation methods cannot well handle the complex and diversified multi-view feature, and the label correlation is also ignored. In this paper, we propose an image annotation method by integrating deep multi-view latent space learning and label correlation guided image annotation into a unified framework, which is termed as Label Correlation guided Deep Multi-view image annotation (LCDM) method. LCDM first learns a consistent multi-view representation via deep matrix factorization, which well captures multi-view complementary information. Then, label correlation is exploited to improve the discriminating power of the classifiers. We propose a unified objective function so that multi-view data representation and classifiers can be jointly learned. Extensive experimental results on various image datasets demonstrate the effectiveness of our method.

INDEX TERMS Deep matrix factorization, image annotation, label correlation, multi-view data, machine learning.

I. INTRODUCTION

A large number of image data are uploaded and disseminated on social network platforms every day. Manually labeling image contents is unpractical due to the huge amount of image data. Thus, automatic image annotation techniques are developed to label images according to their contents. Image annotation techniques have been applied in many fields such as social network image analysis and retrieval, face recognition, intelligent tourism and so on [1]. It is for this reason that image annotation has drawn more and more researchers' interest.

How to bridge the gap between low-level visual features and high-level semantics is the key of image annotation, and various image annotation methods have been developed in recent years. Some methods are based on generative model [2], [3], which calculate the joint distribution between images and labels and maximize the likelihood function. Some image annotation methods adopt nearest neighbor method to first find several nearest neighbours from the labeled images, and then derive the labels from the similar images [4], [5]. Moreover, the most common way of image annotation is to adopt multi-label learning method to predict image labels [6], [7]. Multi-label learning is to learn from a set of samples where each sample belongs to one or more classes. Image annotation can also be treated as a matrix completion problem [8], [9]. As the image labels are usually missing and noisy, matrix completion can complete the missing labels as well as correct the noisy labels [9].

The description ability of image features is a critical factor for image annotation. Powerful image descriptors are capable of improving the performance of image annotation. For image data, different features such as SIFT, HOG and LBP can be extracted, which constitute multi-view feature.

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh.

Compared to single type of feature, multi-view feature provides comprehensive descriptions for images so that better image annotations results can be obtained. Therefore, multi-view image annotation methods have been developed and better annotation results are obtained [10]–[15]. NMF-KNN [10] integrates nearest neighbor method and non-negative matrix factorization (NMF) for image annotation. It jointly factorizes multiple matrices so that multi-view feature and image labels can be associated in a consistent latent space. A multi-view label embedding method [15] is proposed for image annotation. By maximizing the correlations between multi-view feature space, label space and latent space, it can effectively predict image labels with many classes.

Although various multi-view image annotation methods have been developed, there are still some unresolved issues. First, multi-view data may be produced by complex data distributions. The traditional shallow models such as NMF cannot well capture the intrinsic data distribution. Compared with shallow models, deep model is able to extract high level data representation and capture the underlying data distribution. Second, the problem of missing and noisy image labels may lead to a biased estimation of the classifiers. Label correlation is an important clue to complete the missing labels and correct the error labels.

In this paper, we propose a Label Correlation guided Deep Multi-view image annotation method (LCDM) to predict image labels based on multi-view features. Our method first learns deep multi-view latent space via deep multi-view matrix factorization model to represent multi-view data, which effectively encodes the multi-view complementary information into a unified latent space. Then, based on the learned deep multi-view latent space, label classifiers are learned to predict image labels. To reduce the influence of unreliable label matrix and improve the discriminating power of classifiers, our method leverages label correlation to enhance the original label matrix and incorporates label correlation into the classifiers. Extensive experiments on several image annotation datasets demonstrate that our method outperforms the other image annotation methods. The contribution of this paper is summarised as follows:

1) We adopt deep multi-view matrix factorization to learn the unified multi-view representation. The importance of each view can be captured adaptively so that multi-view complementary information can be accurately preserved in the learned data representation.
2) We propose to learn a low-rank subspace from label matrix to explore label correlation. Then the low-rank subspace is used to enhance the original label matrix, so that the impact of missing and noisy labels can be effectively reduced.
3) We use label correlation to guide the training of classifiers. If two labels are closely related, then the corresponding classifiers should be similar. In this way, our method further improves the discriminating power of the classifiers.

II. RELATED WORK

Many methods have been developed for image annotation during the past two decades. Based on single type of visual feature, generative model based methods, discriminative model based methods, matrix completion methods and deep learning based methods are proposed to automatically label images [1]. Generative model based methods aim to maximize likelihood function of visual features and labels [2], [3]. Discriminative model based methods train classifiers for image labels, which convert image annotation to a multi-label learning problem [16], [17]. Considering the problem of missing and noisy labels, matrix completion technique is adopted for image annotation [8], [9], [18]. TMC [8] adopts matrix factorization to search for an optimal label matrix which jointly captures the visual correlation and label correlation. Recently, deep learning techniques are applied to image annotation task [19]–[21]. CNN-RNN [21] constructs a new network structure which combines recurrent neural network and convolutional neural network. Deep image representation and high-order label relations are jointly utilized to predict image labels more accurately.

Multi-view feature provides more diversified and complete description of images. Multi-view image annotation methods can exploit multi-view complementary information to achieve promising image annotation performance [22]–[27]. OGL [22] simultaneously learns an optimal similarity graph of images and propagate labels from labeled images to unlabeled ones. The learned graph can well preserve multi-view information and label information, which yields better label propagation results. MVML [23] jointly performs multi-view feature selection and multi-label learning for image annotation, where a block-row regularizer is used to capture discriminative features. LSA-MML [26] learns a predictive representation by enforcing the latent space of different views to be aligned, which can encode the complementary information of different views. A lifelong multi-task multi-view learning method [27] is developed to capture knowledge from different view-specific libraries, which provides a lifelong learning strategy and better classification performance can be obtained.

As image labels are related to each other, label correlation is leveraged for image annotation [6], [28]–[31]. A graph-based image annotation method is proposed [28] to exploit both local correlations among different labels and global label consistency, where label correlation is treated as a constraint to guide image label prediction. MLMC [6] adopts label correlation and visual correlation for graph learning, which conducts image annotation by maximizing the label assignment consistency over the learned graph. An adaptive graph guided embedding method [30] is developed to utilize label correlation to learn an adaptive graph, and then image annotation is achieved by label propagation. Considering label-label correlation and label-feature correlation, LLSF [31] assumes the correlated labels share more features than uncorrelated ones, then a multi-label classification framework is proposed which also learns label specific features.

III. THE PROPOSED METHOD

A. NOTATIONS

Multi-view image data with n samples and V views can be represented by a set of matrices $\{X^v\}_{v=1}^V$. $X^v \in \mathbb{R}^{d_v \times n}$ is the feature matrix of the v-th view, d_v is the dimension of features from the v-th view. The image label set is denoted as $\{l_1, l_2, \ldots, l_c\}$, where c is the total number of labels. For the labeled images, each image may be annotated with several labels. We adopt label matrix $Y \in \mathbb{R}^{n \times c}$ to represent the relations between images and labels. $Y_{ij} = 1$ means image i is annotated with label l_j, otherwise $Y_{ij} = 0$. The objective of image annotation is to predict labels for the unlabeled images.

B. PRELIMINARIES

Given a non-negative matrix $X \in \mathbb{R}^{d \times n}$ representing n samples, NMF [32] decomposes X into two matrices,

$$\min \ \|X - ZH\|_F^2, \quad s.t.\ Z \geq 0,\ H \geq 0 \tag{1}$$

where $Z \in \mathbb{R}^{d \times k}$ is the basis matrix and $H \in \mathbb{R}^{k \times n}$ is the coefficient matrix. Since NMF provides interpretable and meaningful decomposition results, the coefficient matrix H can be used as a new data representation. However, the image data may be produced by complex data distributions, resulting in NMF cannot effectively capture the intrinsic data distributions. To solve this problem, deep semi-NMF model is developed for single view data [33] and multi-view data [34], which can reveal diversified data distributions and obtain high-level data representation. It decomposes data matrix X into m layers,

$$\begin{aligned} X &\approx Z_1 H_1^+, \\ X &\approx Z_1 Z_2 H_2^+, \\ &\vdots \\ X &\approx Z_1 Z_2 \ldots Z_m H_m^+, \end{aligned} \tag{2}$$

where $Z_i \in \mathbb{R}^{k_{i-1} \times k_i}$ is the basis matrix of the i-th layer, and $H_m \in \mathbb{R}^{k_t \times n}$ is coefficient matrix of the top layer. $(\cdot)^+$ is the hinge operation which is defined as $(a)^+ = \max(0, a)$.

Subspace clustering [35], [36] clusters data points that lie in a union of low-dimensional subspaces. Low-rank subspace clustering (LRSC) [37] aims to find a low-rank representation of data. It solves self-representation problem by finding the low-rank representation of data points as

$$\min_S \|S||_*, \quad s.t.\ X = XS + E, \tag{3}$$

where S is the low-rank subspace learned from data X, E is the error matrix, $\|S\|_*$ is the nuclear norm of S, which equals to the sum of its singular values. The low-rank subspace S can capture the correlation between data points and generate promising clustering results.

C. DEEP MULTI-VIEW LATENT SPACE LEARNING

To obtain unified data representation from multi-view data $\{X^v\}_{v=1}^V$, we adopt deep matrix factorization model to learn the basis matrices and coefficient matrices layer by layer, and the unified data representation is obtained by introducing a consistent coefficient matrix H across all the views. The objective function is proposed as

$$\begin{aligned} &\min_{H, \alpha^v} \sum_{v=1}^V (\alpha^v)^r \|X^v - Z_1^v Z_2^v \ldots Z_m^v H\|_F^2, \\ &s.t.\ \sum_{v=1}^V \alpha^v = 1, \quad \alpha^v > 0,\ H \geq 0 \end{aligned} \tag{4}$$

where Z_i^v is the basis matrix of the i-th layer for view v, m is the number of layers, α^v is the weight parameter to control the importance of the v-th view, H is the learned deep multi-view latent space. By solving problem (4), inter-view and intra-view correlations can be effectively captured and robust multi-view data representation H can be learned. Inter-view correlations are captured by enforcing each view to share a common representation H, so that inter-view complementary information can be preserved. Moreover, by using α^v, the view with smaller embedding loss is considered to be more accurate. Hence, inter-view correlations can be captured more accurately. Intra-view correlations are captured by deep matrix factorization on each view. By minimizing problem (4), the reconstruction error of each view can be reduced, so that H can well encode the intra-view correlations.

D. LABEL CORRELATION GUIDED IMAGE ANNOTATION

Image labels are always missing and noisy, which limits the performance of image annotation. To improve the effectiveness of image annotation, two factors should be considered. First, image labels are correlated with each other. Label correlation can be used to complete missing labels and correct the noisy labels. Second, the properties of classifiers should be consistent with label correlation. Each classifier predict labels based on specific features. If two labels are correlated, the features used for classification should be similar. The classifiers of two correlated labels share more features than the classifiers of two uncorrelated labels. In light of the two factors, we propose the following objective function for image annotation,

$$\begin{aligned} \min_{S,P} \ &\|Y - YS\|_F^2 + \beta \|S||_* + \eta \|PH - S^T Y^T\|_F^2 \\ &+ \lambda Tr(P^T LP) \quad s.t.\ S \geq 0 \end{aligned} \tag{5}$$

where the first two terms are to learn a low-rank subspace $S \in \mathbb{R}^{c \times c}$ from label Y. Since S captures the correlations of labels, we adopt the constraint $S \geq 0$ to ensure the solution is meaningful. The higher value of S_{ij}, the stronger the correlation between two labels. The third term is to predict image labels by linear classifier, and $P \in \mathbb{R}^{c \times k}$ is the classifier parameters. P_i is the i-th column of P, which represents the classifier for label l_i. Label correlation S is used to enhance the original image labels, and $S^T Y^T$ is used as the target to train the classifiers. The last term is a graph regularization constraint that imposed on the classifiers. We introduce the

affinity matrix of labels $W = \frac{S+S^T}{2}$, and its graph Laplacian is $L = D - W$, where D is the diagonal matrix defined as $D_{ii} = \sum_j W_{ij}$. By using the last term, if two labels l_i and l_j achieve higher correlation, then the corresponding classifier parameters P_i and P_j become more similar. β, η and λ are the parameters to control the importance of each term.

E. THE OVERALL OBJECTIVE FUNCTION

By jointly conduct deep multi-view latent space learning and label correlation guided image annotation, we propose to minimize the overall objective function as follows,

$$\begin{aligned} J = &\sum_{v=1}^{V} (\alpha^v)^r \|X^v - Z_1^v Z_2^v \dots Z_m^v H\|_F^2 + \|Y - YS\|_F^2 \\ &+ \beta \|S\|_* + \eta \|PH - S^T Y^T\|_F^2 + \lambda Tr(P^T L P) \\ &s.t. \sum_{v=1}^{V} \alpha^v = 1, \quad \alpha^v \geq 0,\ H \geq 0,\ S \geq 0 \end{aligned} \tag{6}$$

Through deep multi-view latent space learning, our method is capable of learning high-level and robust multi-view representation H. By performing label correlation guided image annotation, our method can cope with the missing labels and enhance the discriminating power of classifiers P. By optimizing the overall objective function J, the two sub-problems can be solved jointly. During the optimization process, multi-view representation learning and classifiers learning can promote each other, so as to achieve better image annotation performance.

IV. OPTIMIZATION

Problem (6) can be effectively solved by an iterative block coordinate descent algorithm. In each iteration, only one variable is solved and keep the others unchanged. First, we adopt the pre-training method as in [33] to obtain proper Z_i^v and H in the deep matrix factorization model. Then, all the variables S, P, Z_i^v, H, and α^v are solved according to the update rules. The detailed pre-training strategy and update rules are introduced in the following part. The whole learning procedure for solving problem (6) is summarized in Algorithm 2.

A. PRE-TRAINING

The latent factors Z_i^v and H in the deep matrix factorization model are pre-trained layer by layer. For instance, for the v-th view, the first layer is trained through decomposition $X^v \approx Z_1^v H_1^v$, where $Z_1^v \in \mathbb{R}^{d_v \times k_1}$ and $H_1^v \in \mathbb{R}^{k_1 \times n}$. After that, the coefficient matrix H_1^v is further decomposed by $H_1^v \approx Z_2^v H_2^v$, where $Z_2^v \in \mathbb{R}^{k_1 \times k_2}$ and $H_2^v \in \mathbb{R}^{k_2 \times n}$. We keep decomposing H_i^v until all the layers are pre-trained, ie, $H_2^v \approx Z_3^v H_3^v, \cdots, H_{m-1}^v \approx Z_m^v H_m^v$. H is initialized by averaging the coefficient matrices of each view $\{H_m^v\}_{v=1}^V$. The merits of pre-training step are that it can effectively accelerate the convergence of the algorithm and obtain better solutions.

B. SOLVE Z

Let the derivative $\partial(J)/\partial(Z_i^v) = 0$, then the update rule for Z_i^v can be obtained by

$$Z_i^v \leftarrow \Psi^\dagger X^v \widetilde{H}_i^{v\dagger}, \tag{7}$$

where $\Psi = Z_1^v \cdots Z_{i-1}^v$, $\widetilde{H}_i^v = Z_{i+1}^v \cdots Z_m^v H$ is the reconstruction of the latent factor of the i-th layer. $(\cdot)^\dagger$ denotes the Moore-Penrose pseudo-inverse operator and $A^\dagger = (A^T A)^{-1} A^T$.

C. SOLVE H

We follow the method in [38] to derive the update rule for H. Keep the related parts from J and we have

$$\begin{aligned} J(H) = &\sum_{v=1}^{V} (\alpha^v)^r (\|X^v - Z_1^v Z_2^v \dots Z_m^v H\|_F^2) \\ &+ \eta \|PH - S^T Y^T\|_F^2 \end{aligned} \tag{8}$$

The partial derivative with respect to H is given as follows,

$$\begin{aligned} \frac{\partial J(H)}{\partial (H)} = &-2 \sum_{v=1}^{V} (\alpha^v)^r \Psi_Z^T (X^v - \Psi_Z H) \\ &+ 2\eta P^T (PH - S^T Y^T) \end{aligned} \tag{9}$$

where $\Psi_Z = Z_1^v Z_2^v \dots Z_m^v$. From the above formulations, we can derive the following update rule for H,

$$H_{ij} \leftarrow H_{ij} \odot \sqrt{\frac{(\Pi_1)_{ij}}{(\Pi_2)_{ij}}}, \tag{10}$$

where

$$\begin{aligned} \Pi_1 = &[\sum_{v=1}^{V} (\alpha^v)^r \Psi_Z^T X^v]^{pos} + [\eta P^T S^T Y^T]^{pos} \\ &+ [\sum_{v=1}^{V} (\alpha^v)^r \Psi_Z^T \Psi_Z H]^{neg} + [\eta P^T P H]^{neg}, \\ \Pi_2 = &[\sum_{v=1}^{V} (\alpha^v)^r \Psi_Z^T X^v]^{neg} + [\eta P^T S^T Y^T]^{neg} \\ &+ [\sum_{v=1}^{V} (\alpha^v)^r \Psi_Z^T \Psi_Z H]^{pos} + [\eta P^T P H]^{pos}. \end{aligned}$$

The operators $[\cdot]^{\text{pos}}$ and $[\cdot]^{\text{neg}}$ are defined as follows,

$$[A]_{jk}^{\text{pos}} = \frac{|A_{jk}| + A_{jk}}{2}, \quad [A]_{jk}^{\text{neg}} = \frac{|A_{jk}| - A_{jk}}{2}.$$

D. SOLVE S

Keeping the parts that are related to S from (6), the following problem are obtained

$$\begin{aligned} \min_S \ &\|Y - YS\|_F^2 + \beta \|S\|_* + \lambda Tr(P^T L P) \\ &+ \eta \|PH - S^T Y^T\|_F^2 \\ s.t.\ &S \geq 0 \end{aligned} \tag{11}$$

To make (11) easier to solve, we replace $Tr(P^T LP)$ by

$$Tr(P^T LP) = \frac{1}{2}\sum_{ij} W_{ij}\|P_i - P_j\|_2^2 = \frac{1}{2}Tr(QW) = \frac{1}{2}Tr(Q\frac{S+S^T}{2}) = \frac{1}{4}Tr(QS+QS^T) = \frac{1}{2}Tr(QS) \tag{12}$$

where P_i is the i-th row of P. $Q_{ij} = \|P_i - P_j\|_2^2$ and we have $Q = Q^T$. Then (11) is rewritten as

$$\begin{aligned} &\min_S \|Y - YS\|_F^2 + \beta\|S\|_* + \frac{\lambda}{2}Tr(QS) \\ &\quad + \eta\|PH - S^T Y^T\|_F^2 \\ &s.t.\ S \geq 0 \end{aligned} \tag{13}$$

Problem (13) can be solved by alternating direction method of multipliers (ADMM) [37]. We rewrite (13) as an unconstrained version:

$$\min_S \|Y - YS\|_F^2 + \beta\|S\|_* + \frac{\lambda}{2}Tr(QS) + \eta\|PH - S^T Y^T\|_F^2 + l_{R^+}(S) \tag{14}$$

where the indicator function $l_{R_+}(a)$ is defined as

$$l_{R^+}(a) = \begin{cases} 0 & \text{if } a \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Then, auxiliary variables are introduced and (14) is equivalent to the following problem

$$\begin{aligned} &\min_S \ \|Y - B_1\|_F^2 + \beta\|B_2\|_* + \frac{\lambda}{2}Tr(B_3) \\ &\quad + \eta\|PH - B_4\|_F^2 + l_{R^+}(B_5) \\ &s.t.\ YS = B_1, \quad S = B_2,\ QS = B_3,\ S^T Y^T = B_4,\ S = B_5 \end{aligned} \tag{15}$$

The augmented Lagrangian function of problem (15) is

$$\begin{aligned} &\mathcal{L}(S, B_1, B_2, B_3, B_4, B_5) \\ &= \|Y - B_1\|_F^2 + \beta\|B_2\|_* + \frac{\lambda}{2}Tr(B_3) \\ &\quad + \eta\|PH - B_4\|_F^2 + l_{R^+}(B_5) + \mu\|B_1 - YS - R_1\|_F^2 \\ &\quad + \mu\|B_2 - S - R_2\|_F^2 + \mu\|B_3 - QS - R_3\|_F^2 \\ &\quad + \mu\|B_4 - S^T Y^T - R_4\|_F^2 + \mu\|B_5 - S - R_5\|_F^2 \end{aligned} \tag{16}$$

We apply alternative minimization method to solve all the variables S, B_1, B_2, B_3, B_4 and B_5. In each step, only one variable is updated while keep the others fixed.

To solve S from (16), we set the partial derivative $\partial(\mathcal{L}(S, B_1, B_2, B_3, B_4))/\ \partial(S) = 0$ and obtain

$$\begin{aligned} S \leftarrow (2Y^T Y + Q^T Q + 2I)^{-1}(Y^T\xi_1 + \xi_2 \\ + Q^T\xi_3 + Y^T\xi_4^T + \xi_5) \end{aligned} \tag{17}$$

where $\xi_i = B_i - R_i$, and I is the identity matrix.

B_1 is solved by setting $\partial(\mathcal{L}(S, B_1, B_2, B_3, B_4))/\ \partial(B_1) = 0$,

$$B_1 \leftarrow \frac{1}{\mu+1}(Y + \mu(YS + R_1)) \tag{18}$$

Algorithm 1 The Algorithm to Solve S

Input: $Y, Q, H, P, \alpha^v, \beta, \lambda, \eta$.

Initialization: $\forall i, B_i = R_i = 0$
while *not converged* **do**
  $S \leftarrow (2Y^T Y + Q^T Q + 2I)^{-1}(Y^T\xi_1 + \xi_2 + Q^T\xi_3 + Y^T\xi_4^T + \xi_5)$
  $B_1 \leftarrow \frac{1}{\mu+1}(Y + \mu(YS + R_1))$
  $B_2 \leftarrow \Theta_{\beta/2\mu}(S + R_2)$
  $B_3 \leftarrow \frac{1}{4\mu}(4\mu(QS + R_3) - \lambda I)$
  $B_4 \leftarrow \frac{1}{\eta+\mu}(\eta PH + \mu(S^T Y^T + R_4))$
  $B_5 \leftarrow \max(S + R_5, 0)$
  update the Lagrange multipliers:
  $R_1 \leftarrow R_1 - (B_1 - YS)$;
  $R_2 \leftarrow R_2 - (B_2 - S)$;
  $R_3 \leftarrow R_3 - (B_3 - QS)$;
  $R_4 \leftarrow R_4 - (B_4 - S^T Y^T)$
  $R_5 \leftarrow R_5 - (B_5 - S)$;
end
Output: S.

To obtain B_2, we solve the following problem

$$\min_{B_2} \beta\|B_2\|_* + \mu\|B_2 - S - R_2\|_F^2 \tag{19}$$

Problem (19) can be solved by singular value thresholding operator [39]. Let $\Theta_\tau(X) = U\Lambda_\tau V^{\mathrm{T}}$, where $X = U\Lambda_\tau V^{\mathrm{T}}$ is the singular value decomposition, and $\Lambda_\tau(x) = \mathrm{sgn}(x)\max(|x| - \tau, 0)$ is the shrinkage operator. B_2 can be solved by

$$B_2 \leftarrow \Theta_{\beta/2\mu}(S + R_2). \tag{20}$$

Following the same method as solving B_1, the update rule of B_3 and B_4 can be obtained as

$$B_3 \leftarrow \frac{1}{4\mu}(4\mu(QS + R_3) - \lambda I), \tag{21}$$

$$B_4 \leftarrow \frac{1}{\eta+\mu}(\eta PH + \mu(S^T Y^T + R_4)). \tag{22}$$

Considering the non-negative constraint that imposed on B_5, we solve it using the following update rule

$$B_5 \leftarrow \max(S + R_5, 0). \tag{23}$$

Finally, the Lagrangian multipliers R_1, R_2, R_3, R_4 and R_5 are updated through ADMM algorithm. All the variables are solved by the above update rules, and the whole procedure for solving S is summarized in Algorithm 1.

E. SOLVE P

Fixing the related parts of P from J, we can obtain the following problem

$$J(P) = \lambda Tr(P^T LP) + \eta\|PH - S^T Y^T\|_F^2, \tag{24}$$

The derivative of $J(P)$ with respect to P is

$$\nabla_P J(P) = 2\lambda LP + 2\eta(PH - S^T Y^T)H^T, \tag{25}$$

Algorithm 2 The Optimization Algorithm of LDCM

Input: $\{X^v\}_{v=1}^V$, m, r, λ, β, and η.

Initialize Z_i^v by pre-training, $\alpha^v = \frac{1}{V}$.
while *not converged* **do**
  for $v = 1, ..., V$ *and* $i = 1, ...m$ **do**
    Update α^v by update rule (26) and (27);
    Update Z_i^v by $Z_i^v \leftarrow \Psi^\dagger X^v \widetilde{H}_i^{v\dagger}$;
  end
  Update H by $H \leftarrow H \odot \sqrt{\frac{\Pi_1}{\Pi_2}}$;
  Update S by Algorithm 1;
  Update P by $P \leftarrow P - \delta \nabla_P J(P)$;
end

Output: α^v, Z_i^v, H, S, P.

We adopt gradient descent method to solve P, and the step size δ is determined by Armijo line search [40].

F. SOLVE α

By using Lagrange multiplier method, we can solve the weight parameter α^v. For the case $r > 1$, the following update rule can be derived,

$$\alpha^v = \frac{(\rho^v)^{\frac{1}{1-r}}}{\sum_{v=1}^V (\rho^v)^{\frac{1}{1-r}}}, \tag{26}$$

where $\rho^v = \|X^v - Z_1^v Z_2^v \dots Z_m^v H\|_F^2$.

For the case $r = 1$, we can obtain the following update rule:

$$\alpha^v = \begin{cases} 1 & v = \underset{i}{argmin}\ \rho^i \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

G. OUT-OF-SAMPLE EXTENSION

To predict labels for unlabeled images, we adopt the following steps. Given unlabeled multi-view feature $\{X^v\}_{v=1}^V$, we first obtain data representation for unlabeled data by using Eq.(10) where we set $\eta = 0$, and we can obtain $\hat{H}$. Then, the predicted label matrix $\hat{Y}$ is given by $\hat{Y} = P\hat{H}$.

H. COMPUTATIONAL COMPLEXITY

For pre-training step, the computational complexity is of order $\mathcal{O}(mVt_1(ndk + nk^2 + kn^2))$, where m is the number of layers, k is the maximum dimensionality of the layers, n is the number of images, d is the dimensionality of feature, and t_1 is the number of iterations. For fine-tuning step, the main computational cost is dominated by updating S, Z_i^v, H and P. The complexity for fine-tuning is of order $\mathcal{O}\big(mVt_2(ndk + nk^2 + kn^2 + kc^2 + t_3(c^3 + nc^2 + ckn))\big)$, where c is the number of labels, t_2 and t_3 are the number of iterations of Algorithm 2 and Algorithm 1, respectively.

V. EXPERIMENTS

We conduct image annotation experiments on four datasets to verify the effectiveness of the proposed method LCDM. The datasets, compared methods and experimental settings are introduced first. Then, we present the performance comparison of all the methods on each dataset. Finally, we present parameter sensitivity analysis to further evaluate the performance of the proposed method.

A. DATASETS

1) Corel5k [41]. It consists of 5,000 images from 50 classes. 260 keywords are contained in the vocabulary. 4500 samples are used as the training set and the rest 500 images are used for testing.
2) ESP Game [42]. It consists of 20,770 images collected from ESP online labeling game. The dataset contains 268 keywords. 18,689 images are used for training and the rest images are used for testing.
3) NUS-WIDE [43]. It contains 55,615 images collected from Flickr. Images that are annotated less than 3 labels and labels whose occurrence numbers are smaller than 100 are removed to improve the quality of the dataset. The remaining 13,000 images constitute the dataset, where 10,000 samples are randomly chosen for training and the remaining are used for testing.
4) IAPRTC-12 [44]. Its images cover many scenes including sports, landscapes, animals, buildings and other aspects in our life. The dataset contains 19,267 images with 291 keywords. 17,665 images are used for training and 1,962 images are used for testing.

B. COMPARED METHODS AND EXPERIMENTAL SETTINGS

To fully demonstrate the effectiveness of our method, we compare our method LCDM with several representative image annotation methods. The first five methods are single view image annotation methods (1-5), while the last six methods are multi-view image annotation methods (6-11). We introduce each method in detail as follows.

1) FastTag [45]: An image tagging method which can quickly predict image tags via combining two linear mappings in a convex loss function.
2) LSG [46]: An image annotation method which models label correlation using a graph, and the topological constraints are utilized for multi-label learning.
3) LSR [9]: A label completion method which is based on label matrix and image matrix reconstruction.
4) TMC [8]: A label completion method which recovers tag matrix according to visual and semantical correlation of images.
5) GLOCAL [47]: A multi-label learning method which exploits global and local label correlations based on a latent label subspace and label manifolds.
6) NMF-KNN [10]: It utilize nearest neighbour model and matrix factorization technique to label images.

7) OGL [22]: An optimal graph is learned from different views and image labels, and then tags are propagated from labeled images to unlabeled ones.
8) lrMVL [48]: A consistent representation of multi-view data is learned to predict image labels by a low-rank matrix completion method.
9) MVLR [49]: A multi-view linear regression model which can be used for image annotation, and a closed-form solution of the parameters can be obtained.
10) OPSL [50]: A multi-view image annotation method based on optimal predictive subspace learning, where both image representation and label predictors can be jointly learned.
11) iMVWL [51]: A multi-view multi-label learning method which learns a unified subspace and a predictor. Both multi-view correlation and label correlation can be captured in this model.

To construct multi-view feature for images, we extract different kinds of visual features. For Corel5k, ESP Game and IAPRTC-12 datasets, we adopt seven visual features [50], [52]: DenseHueV3H1, DenseHue, HarrisHueV3H1, HarrisHue, DenseSift, HarrisSift and Gist. For NUS dataset, six types of features are used: color correlation, color moments, color histogram, SIFT, edge direction histogram and wavelet texture. For single view image annotation methods, multi-view feature cannot be directly utilized. Thus, we perform PCA on the feature of each view and then concatenate the obtained results as the new feature for single view methods.

The parameters of the compared methods are determined as suggested in the corresponding literatures. The parameters of LCDM are determined by cross-validation. $1/10$ of training data are used as the validation set. η is tuned from $\{0.0005, 0.001, \ldots, 10, 50\}$, λ is tuned from $\{10^{-6}, 10^{-5}, \ldots, 1, 10\}$, β is tuned from $\{0.01, 0.1, 1\}$, r is tuned from $\{1, 1.5, 2, 5, 10, 50, 100\}$, the structure of our model is tuned from {(100), (150-100), (200-150-100), (250-200-150-100)}, where (200-150-100) is a 3-layer model and the dimensions of the first, second and top layer are 200, 150 and 100, respectively. The detailed parameter sensitivity analysis of LCDM is presented in Section V-D. Since different initializations of LCDM would obtain different solutions, we repeat training and testing of LCDM ten times and the averaged performance are reported.

To evaluate image annotation performance of each method, we annotate five most relevant labels to each image in the experiments. As in [8], [22], [53], four commonly used evaluation measures, average precision (P), average recall (R), F1-score (F1) and Mean Average Precision (MAP) are used for performance evaluation. We first calculate the evaluation measures for each image, and then report the averaged results over all the images.

C. EXPERIMENTAL RESULTS

We conduct image annotation experiments on four datasets, and the image annotation results are presented in Table 1.

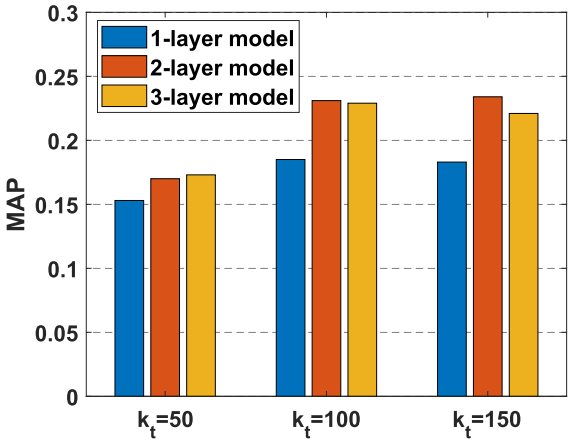

FIGURE 1. The image annotation performance with different dimensions of layers on NUS dataset.

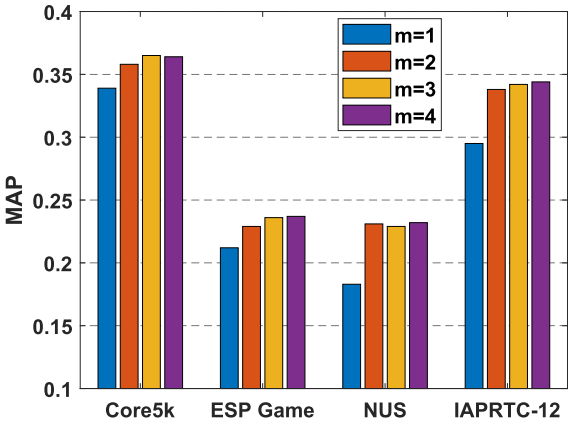

FIGURE 2. The image annotation performance with different number of layers.

It should be noted that we adopt the same dataset and feature as [50], thus Table 1 shares some common results with [50]. From Table 1, we can observe that LCDM outperforms the other image annotation methods on all the datasets, which demonstrates the effectiveness of the proposed method. Compared to the best results that achieved by the other methods, LCDM improves the performance by 2.1% in F1 and 3.3% in MAP for Corel5k dataset, 2.8% in F1 and 3.5% in MAP for ESP Game dataset, 4.6% in F1 and 4.3% in MAP for NUS dataset, 3.7% in F1 and 4.1% in MAP for IAPRTC-12 dataset. From the experimental results presented in Table 1, we would like to highlight some other aspects of the experimental results.

- In general, single view image annotation methods FastTag, LSG, LSR, TMC and GLOCAL perform not as well as multi-view image annotation methods such as NMF-KNN, OGL, OPSL, iMVWL and LCDM. The reason is that multi-view image annotation methods are capable of utilizing the complementary information of multi-view data, so that more complete image descriptions can be obtained. Although single view methods also use multi-view features by concatenating the results learned by

TABLE 1. P, R, F1 and MAP results on all the datasets.

Methods	Corel5k				ESP Game				NUS				IAPRTC-12			
	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP
FastTag [45]	32.2	45.7	37.8	25.3	29.0	32.1	30.5	12.2	58.0	26.6	36.5	11.2	35.9	33.1	34.4	25.8
LSG [46]	30.3	41.9	35.2	21.6	25.3	29.5	27.2	10.7	50.6	21.8	30.5	9.6	32.4	30.9	31.6	21.2
LSR [9]	33.1	46.8	38.8	24.8	28.5	32.4	30.3	14.9	52.8	24.2	33.2	13.6	34.6	32.2	33.4	22.7
TMC [8]	31.7	37.1	33.9	17.3	21.1	23.2	22.1	9.8	39.2	17.9	24.6	9.4	30.0	28.4	29.2	19.8
GLOCAL [47]	31.6	43.1	36.5	22.6	27.8	30.5	29.1	11.5	55.2	23.9	33.4	11.0	33.9	31.7	32.8	23.5
NMF-KNN [10]	35.0	49.6	41.0	26.2	28.4	31.6	29.4	13.7	51.6	23.8	32.5	10.5	34.6	33.4	34.1	24.4
OGL [22]	34.7	49.0	40.7	27.5	31.0	34.1	32.5	17.0	57.2	26.2	35.9	13.3	39.3	36.6	37.9	28.3
lrMVL [48]	29.9	42.0	34.9	20.4	25.9	28.5	27.1	10.3	48.6	22.3	30.6	9.4	29.1	27.0	28.0	20.4
MVLR [49]	25.9	37.2	30.5	16.9	24.5	27.2	25.8	9.5	37.7	17.3	23.7	7.9	28.1	26.7	27.4	19.2
OPSL [50]	38.3	55.0	45.2	33.2	33.5	36.9	35.1	20.1	59.2	27.2	37.3	18.9	42.5	40.0	41.2	30.1
iMVWL [51]	36.9	50.5	42.6	29.8	32.7	34.6	33.6	19.4	58.1	26.6	36.5	18.3	39.5	36.2	37.8	28.6
LCDM	**40.2**	**57.5**	**47.3**	**36.5**	**36.1**	**39.8**	**37.9**	**23.6**	**62.9**	**31.4**	**41.9**	**23.2**	**46.7**	**43.3**	**44.9**	**34.2**

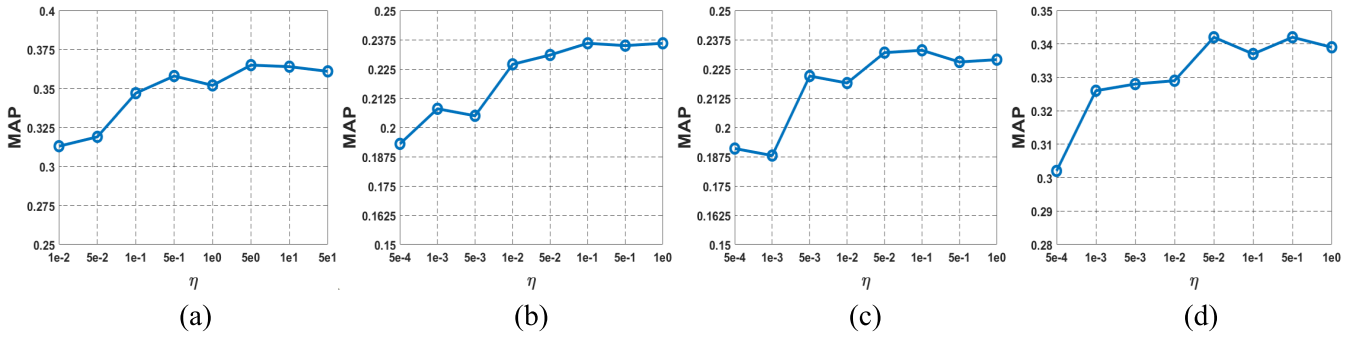

FIGURE 3. Sensitivity analysis of η on each dataset. (a) Corel5k dataset. (b) ESP Game dataset. (c) NUS dataset. (d) IAPRTC-12 dataset.

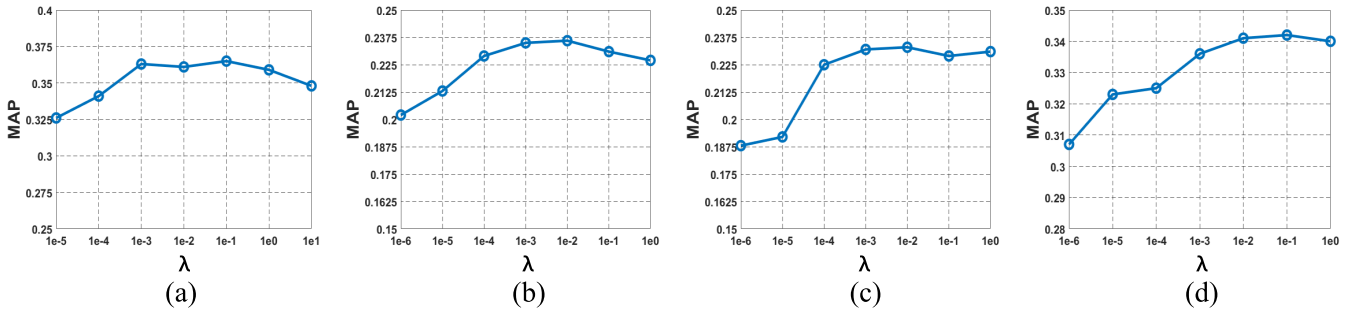

FIGURE 4. Sensitivity analysis of λ on each dataset. (a) Corel5k dataset. (b) ESP Game dataset. (c) NUS dataset. (d) IAPRTC-12 dataset.

PCA, the multi-view complementary information cannot be well leveraged so that their annotation performance are limited.

- Leveraging label correlation can effectively address the problem of the missing and noisy image labels. OPSL encodes the label correlation into the learned subspace. iMVWL and LCDM utilize low-rank property of label correlation matrix to predict image labels. These methods can achieve more competitive image annotation performance than the other methods.
- TMC, NMF-KNN and LCDM are image annotation methods based on matrix factorization. NMF-KNN outperforms TMC because NMF-KNN can effectively utilize multi-view feature to predict image labels, while TMC cannot properly use multi-view feature. LCDM achieves better performance than TMC and NMF-KNN because deep multi-view matrix factorization can learn high level and robust data representation that shallow model cannot obtain.
- The proposed method LCDM generally achieves better performance compared to the other methods. The main reasons are summarized as follows. First, LCDM learns a deep multi-view latent space from the diversified image samples with complex data distributions.

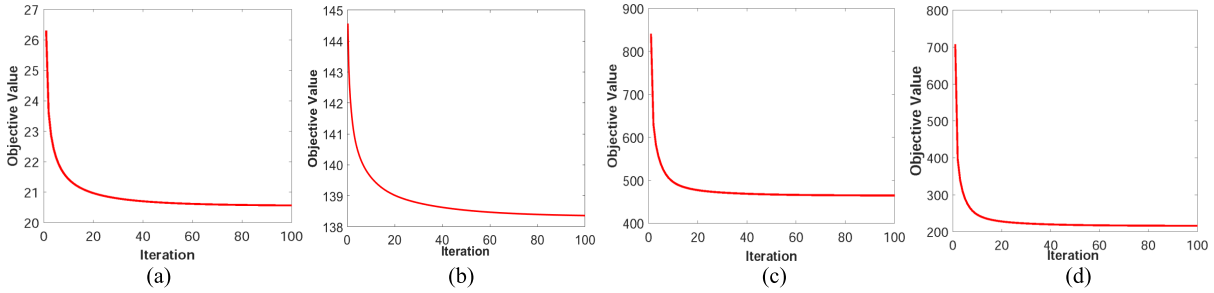

FIGURE 5. Convergence curves on each dataset. (a) Corel5k dataset. (b) ESP Game dataset. (c) NUS dataset. (d) IAPRTC-12 dataset.

Deep matrix factorisation is capable of extracting intrinsic data distribution so that multi-view complementary information can be well preserved in the latent space. Second, label correlation is used for enhancing the original label matrix as well as guiding the learning of the classifiers. Thus, the accuracy of label matrix can be improved and more discriminative classifiers are obtained. All these factors make LCDM method achieve more promising image annotation performance.

D. PARAMETER ANALYSIS

To demonstrate the performance of LCDM under different parameter settings, we conduct parameter sensitivity experiments and the image annotation performance on MAP are shown in Fig.1, Fig.2, Fig.3 and Fig.4.

To study how the performance of LCDM vary with different dimensions of layers, we construct 1-layer model, 2-layer model, 3-layer model, and make the dimension of each layer increase by 50. We set the dimension of the top layer (k_t) to 50, 100, 150 and obtain 9 models, ie, $k_t = 50$: (50), (100-50), (150-100-50); $k_t = 100$: (100), (150-100), (200-150-100); $k_t = 150$: (150), (200-150), (250-200-150). The performances of each model on NUS dataset are shown in Fig.1. It can be observed that $k_t = 100$ and $k_t = 150$ achieve better performance than $k_t = 50$. This is because $k_t = 50$ cannot fully preserve the information of multi-view data. $k_t = 150$ achieves comparable results as $k_t = 100$, however, this model contains some redundant dimensions. Hence, $k_t = 100$ is used in our experiments.

To evaluate how the performance of LCDM change with the number of layers m, we test four models: 1-layer model (m=1), 2-layer model (m=2), 3-layer model (m=3) and 4-layer model (m=4). The image annotation results are shown in Fig.2. We can observe that $m = 2, 3, 4$ achieve better performance than $m = 1$ on all the datasets. It indicates that compared with shallow model ($m = 1$), deep models ($m = 2, 3, 4$) are capable of capturing the intrinsic distribution of multi-view data and exploiting multi-view complementary information, so that more effective multi-view data representation can be obtained.

Next, we study two important parameters η and λ. η controls the weight of linear classification, and the sensitivity analysis on η are shown in Fig.3. We can observe that better performance can be obtained for $\eta > 0.1$ on Corel5k dataset and $\eta > 0.01$ on ESP Game, NUS and IAPRTC-12 datasets. λ controls the weight of label correlation constraint, and the sensitivity analysis on λ are shown in Fig.4. When λ is small, the performance of LCDM is limited. This is because the label correlation cannot well guide the learning of classifiers. If λ is too large, the performance of LCDM is also limited on Corel5k, ESP Game datasets. The reason is that the label correlation is so strong that it influences the classifiers to obtain proper parameters. Hence, the appropriate range of λ is [0.001, 1] for Corel5k dataset, and [0.001, 0.1] for ESP Game, NUS, IAPRTC-12 datasets.

Finally, we analyze the convergence speed of the proposed method. The convergence curves of LCDM on all the datasets are shown in Fig.5. We can see that the proposed method is efficient and it usually converges in 60 iterations. The results from Fig.5 verity the effectiveness and correctness of the proposed optimization algorithm.

VI. CONCLUSION

In this paper, a label correlation guided deep multi-view image annotation method LCDM is proposed. LCDM incorporates deep multi-view latent space learning and label correlation guided image annotation into a unified objective function, which can jointly learn multi-view data representation and classifiers. The experimental results on four image datasets demonstrate that the proposed method outperforms the other image annotation methods and promising image annotation performance are obtained.

REFERENCES

[1] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognit.*, vol. 79, pp. 242–259, Jul. 2018.
[2] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Informaion Retr.*, 2003, pp. 119–126.
[3] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun./Jul. 2004, p. II.
[4] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Computer Vision—ECCV 2008* (Lecture Notes in Computer Science), vol. 5304, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Germany: Springer, 2008, pp. 316–329.
[5] L. Wu, S. C. H. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 135–144.
[6] H. Wang and J. Hu, "Multi-label image annotation via maximum consistency," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2337–2340.

[7] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "Multi-feature metric learning with knowledge transfer among semantics and social tagging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2240–2247.
[8] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.
[9] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1618–1625.
[10] 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. (CVPR)*, Jun. 2014, pp. 184–191.
[11] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "S3MKL: Scalable semi-supervised multiple kernel learning for image data mining," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 163–172.
[12] Z. Xue, G. Li, and Q. Huang, "Joint multi-view representation learning and image tagging," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, 1–7.
[13] Q. Wang, Y. Guo, J. Wang, X. Luo, and X. Kong, "Multi-view analysis dictionary learning for image classification," *IEEE Access*, vol. 6, pp. 20174–20183, 2018.
[14] W. Ou, J. Gou, Q. Zhou, S. Ge, and F. Long, "Discriminative multiview nonnegative matrix factorization for classification," *IEEE Access*, vol. 7, pp. 60947–60956, 2019.
[15] P. Zhu, Q. Hu, Q. Hu, C. Zhang, and Z. Feng, "Multi-view label embedding," *Pattern Recognit.*, vol. 84, pp. 126–135, Dec. 2018.
[16] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.
[17] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.
[18] Z. Qin, C.-G. Li, H. Zhang, and J. Guo, "Improving tag matrix completion for image annotation and retrieval," in *Proc. Vis. Commun. Image Process. (VCIP)*, Dec. 2015, pp. 1–4.
[19] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3460–3469.
[20] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2960–2968.
[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2285–2294.
[22] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," *Trans. Cybern.*, vol. 44, no. 7, pp. 1225–1236, 2014.
[23] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 450–461, Feb. 2015.
[24] S. Kucheryavskiy, "Analysis of NIR spectroscopic data using decision trees and their ensembles," *J. Anal. Test.*, vol. 2, no. 3, pp. 274–289, 2018.
[25] T. Shu, B. Zhang, and Y. Y. Tang, "Multi-view classification via a fast and effective multi-view nearest-subspace classifier," *IEEE Access*, vol. 7, pp. 49669–49679, 2019.
[26] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang, "Latent semantic aware multi-view multi-label classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[27] G. Sun, Y. Cong, J. Li, and Y. Fu, "Robust lifelong multi-task multi-view representation learning," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Nov. 2018, pp. 91–98.
[28] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, "Graph-based semi-supervised learning with multi-label," in *Proc. IEEE Int. Conf. Multimedia Expo*, Apr./Jun. 2008, pp. 1321–1324.
[29] S. Hugelier, M. Sliwa, and C. Ruckebusch, "A perspective on data processing in super-resolution fluorescence microscopy imaging," *J. Anal. Test.*, vol. 2, no. 3, pp. 193–209, 2018.
[30] L. Wang, Z. Ding, and Y. Fu, "Adaptive graph guided embedding for multi-label annotation," in *Proc. IJCAI*, 2018, pp. 2798–2804.
[31] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label specific features for multi-label classification," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 181–190.
[32] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 556–562.
[33] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep semi-nmf model for learning hidden representations," in *Proc. ICML*, 2014, pp. 1692–1700.
[34] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Proc. AAAI*, 2017, pp. 2921–2927.
[35] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.
[36] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[37] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[38] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[39] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[40] D. P. Bertsekas, "Nonlinear programming," *J. Oper. Res. Soc.*, vol. 48, no. 3, p. 334, 1997.
[41] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Computer Vision—ECCV 2002* (Lecture Notes in Computer Science), vol. 2353, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Germany: Springer, 2002, pp. 97–112.
[42] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2004, pp. 319–326.
[43] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 48.
[44] M. Grubinger, "Analysis and evaluation of visual information systems performance," Ph.D. dissertation, School Comput. Sci. Math., Fac. Health, Eng. Sci., Victoria Univ., Melbourne, VIC, Australia, 2007.
[45] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1274–1282.
[46] X. Cai, F. Nie, W. Cai, and H. Huang, "New graph structured sparsity model for multi-label image annotations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 801–808.
[47] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, Jun. 2018.
[48] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.
[49] S. Zheng, X. Cai, C. Ding, F. Nie, and H. Huang, "A closed form solution to multi-view low-rank regression," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.
[50] Z. Xue, G. Li, and Q. Huang, "Joint multi-view representation and image annotation via optimal predictive subspace learning," *Inf. Sci.*, vols. 451–452, pp. 180–194, Jul. 2018.
[51] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Incomplete multi-view weak-label learning," in *Proc. IJCAI*, 2018, pp. 2703–2709.
[52] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 309–316.
[53] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Comput. Soc. Tech. Committee Data Eng.*, vol. 24, no. 4, pp. 35–43, Jan. 2001.

ZHE XUE received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Assistant Professor with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing. His research interests include machine learning, computer vision, and multimedia data mining.

JUNPING DU received the Ph.D. degree in computer science from the University of Science and Technology Beijing. She held a Postdoctoral Fellowship with the Department of Computer Science, Tsinghua University. She joined the School of Computer Science, Beijing University of Posts and Telecommunications, in 2006, where she is currently a Professor. She has served as the Chair and the Co-Chair of IPC for many international and domestic academic conferences. She has been the Vice General Secretary of the Chinese Association for Artificial Intelligence. She was a Visiting Professor with the Department of Computer Science, Aarhus University, Aarhus, Denmark, from 1996 to 1997. Her current research interests include artificial intelligence, data mining, motion image processing, social network analysis and search, and computer applications.

MIN ZUO received the Ph.D. degree in computer science from the University of Science and Technology Beijing, in 2011. He is currently a Professor with the School of Computer and Information Engineering, Beijing Technology and Business University. His research interests include artificial intelligence, and big data technology and application.

GUORONG LI received the B.S. degree in computer science from the Renmin University of China, in 2006, and the Ph.D. degree in computer science from the Graduate University of the Chinese Academy of Sciences, in 2012. She is currently an Associate Professor with the University of Chinese Academy of Sciences. Her research interests include object tracking, pattern recognition, cross-media analysis, and multi-label learning.

QINGMING HUANG received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, China, in 1988 and 1994, respectively. He is currently a professor with the University of Chinese Academy of Sciences and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has authored or coauthored more than 300 academic articles in prestigious international journals and top-level international conferences. His research interests include multimedia computing, image processing, computer vision, and pattern recognition. He has served as the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, and PSIVT. He is also an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Acta Automatica Sinica*, and a Reviewer of various international journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON MULTIMEDIA.