Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 5987906, 11 pages
https://doi.org/10.1155/2018/5987906

Research Article

A Semisupervised Framework for Automatic Image Annotation Based on Graph Embedding and Multiview Nonnegative Matrix Factorization

Hongwei Ge,[1] **Zehang Yan**,[1] **Jing Dou**,[1] **Zhen Wang**,[2] **and ZhiQiang Wang**[1]

[1]*School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China*
[2]*School of Mathematical Science, Dalian University of Technology, Dalian 116023, China*

Correspondence should be addressed to Hongwei Ge; hwge@dlut.edu.cn

Received 25 February 2018; Accepted 4 June 2018; Published 27 June 2018

Academic Editor: Qian Zhang

Automatic image annotation is for more accurate image retrieval and classification by assigning labels to images. This paper proposes a semisupervised framework based on graph embedding and multiview nonnegative matrix factorization (GENMF) for automatic image annotation with multilabel images. First, we construct a graph embedding term in the multiview NMF based on the association diagrams between labels for semantic constraints. Then, the multiview features are fused and dimensions are reduced based on multiview NMF algorithm. Finally, image annotation is achieved by using the new features through a KNN-based approach. Experiments validate that the proposed algorithm has achieved competitive performance in terms of accuracy and efficiency.

1. Introduction

The advent of Internet age brings the explosive growth of image resources. Although managing and retrieving images by semantic tags is a common and effective way, there are still a large number of untagged or not fully tagged images. However, it is not easy to carry out manual annotation regarding the cost of human resources and the semantic nuances of annotation under the background of various cultures, religions, and languages. Moreover, the cognition bias caused by subjectivity could induce semantic discrepancies as well. Thus, how to design an efficient automatic image annotation algorithm to provide accurate labels for untagged images has been an urgent problem.

Automatic image annotation (AIA) refers to the process that computers automatically provide one or more semantic tags that can reflect the content of a specific image through algorithms. It is a mapping from images to semantic concepts, namely, the process of understanding images. Image annotation is based on image feature representations, and features utilized in different tasks have different representation abilities [1–3]. For example, global color and texture features have been successfully used in retrieving similar images [4], while local structure features perform well in tasks of object classification and matching [5, 6]. In general, features that depict images from different views can provide complementary information. Thus a rational fusion of multiview features contributes to more comprehensive depiction for images, which can be beneficial to image searching, classification, or other related tasks.

Many multiview learning algorithms have been proposed for operating some tasks such as classification, retrieval, and clustering based on multiview features. According to the levels of feature fusion, multiview learning methods can be grouped into two categories [7]: feature-level fusion such as MKL [8], SVM-2K [9], and CCA [10] and classifier-level fusion such as hierarchical SVM [11]. Some experimental studies show that classifier-level fusion outperforms simple feature concatenation, whereas sophisticated feature-level fusion usually performs better than classifier-level fusion [11, 12].

Recently, many image annotation algorithms use a variety of underlying features to improve annotation performance [8–10]. On one hand, the multiview features improve the

accuracy, but on the other hand the strategies decrease the efficiency and applicability of the algorithms because of the increase of feature dimensions. Moreover, many existing multiview learning algorithms are unsupervised; that is, they do not make use of the label information in the training set. Such fused features may not effectively contain the semantic relationship between samples. This paper proposes a semisupervised learning framework based on graph embedding and multiview NMF (GENMF). In GENMF, feature fusion and dimension reduction are firstly performed by the proposed graph embedded multiview NMF algorithm, and then the new obtained features are used to annotate images through KNN-based approach.

2. Related Works

Existing image annotation algorithms can be roughly divided into two categories [13]: model-based learning methods and database-based retrieval methods. Model-based methods explore the relationship between high-level semantic concepts and low-level visual features to discover a mapping function through machine learning or knowledge models for image annotation. Unlike model-based methods, database-based methods do not need to set up the mapping function based on the training set but directly provide a sequence of candidate labels according to the already annotated images in the database.

There are three kinds of model-based learning methods for image annotation: classification based methods, possibility based methods, and topic model-based methods. Classification based methods [14–16] treat tags as specific class labels and explore the mapping relations between low-level visual features and labels through machine learning methods. The essence of this kind of methods is transforming image annotation to image classification. Different classifiers are used to establish mapping functions between low-level features (from images or regions) and semantic concepts. Labels with the high confidence from the classifiers are annotated to images. Different from classification based methods, possibility based methods [17, 18] do not use classifiers to build the mapping functions but explore the relationship between the underlying features of the image and the semantic labels based on unsupervised probability and statistics models. They utilize the relations to calculate the joint probability of images and labels or the conditional probability of labels given an image and then estimate the possible labels through statistical inference. Topic model-based methods [19, 20] use latent topics to associate low-level visual features with high-level semantic concepts to implement image annotation.

The model-based methods have three difficulties in practical applications. First, the learning models trained on the datasets with finite image types and semantic labels can hardly reflect the characteristics of feature distributions in the real world, which leads to unsatisfactory annotation performance when facing new features and semantic labels. Second, the limited size of training sets may result in overfitting and low generalization ability of the models. Third, low-level features may often fail to express high-level semantic information because they belong to different feature spaces. Thus, it is also hard to establish a mapping model between image features and semantic concepts because of the semantic gap.

The essence of retrieval based method is directly providing a list of candidate labels for the images to be tagged based on the existing datasets with complete and valid label information. Most common retrieval methods are based on KNN [21–23]: they retrieve k images with the highest similarity to the input image from the database, and the labels of the k images are sorted based on the statistical relationship or weighted statistical relationship to generate the candidate labels of the input images. The other category is graph-based methods [24–27] that utilize image feature distance to establish relevant graphs of samples. Based on the assumption that neighboring images in the relevant graph have similar labels (label smoothness), the similarity between nodes and the global structural characteristics of the relevant graph are used to propagate and enrich the node information including labels and classes. This kind of semisupervised learning methods is suitable for not fully tagged datasets existing on the Internet.

Traditional graph-based methods usually label images by aggregating multiple features into one feature and building a relation graph based on this feature. In [25], it is pointed out that traditional methods cannot effectively capture the unique information for each feature and proposes to utilize different features to establish relation subgraphs and then link these subgraphs to form a supergraph. Based on the supergraph, label propagation is achieved through the graph-based method. In [26], different feature graphs are built based on different features of the images and then the relationship between images is constructed through the graph-based method based on different feature graphs. Furthermore, the relationship between images and different features can be also constructed. Finally, the two relationships, namely, the relation between images and the relation between images and different features, can be fused by a designed objective function to obtain good candidates for the labels.

In [27], a graph learning KNN (GLKNN) is proposed by combining KNN-based method and graph-based method. GLKNN first uses graph-based method to propagate the labels of the K nearest neighbors to the new image and obtain one sequence of candidate labels, then GLKNN employs the naive-Bayes nearest neighbor algorithm to establish the relationship between labels and image features for obtaining another sequence of candidate labels. Finally, the two candidate label sequences are linearly combined as the final predicted labels. In [28], graph embedding discriminant analysis is applied to classify marine fish marine fish species by constructing intraclass similarity graph and interclass penalty graph. Although the algorithm improves the performance of classification and clustering by utilizing class labels to build graph embedded term, the traditional graph embedding algorithm is not suitable for multilabel problems with multilabel images because there is no intraclass and interclass relationship. In [21, 22], different models based on metric methods are proposed to enhance the representation ability of features and further improve the performance of image

annotation. However, the metric based feature processing only linearly embeds the original features and does not reduce the feature dimension. In [13], multiple features are fused by concatenation, which ignores the manifold characters of different features and high feature dimension results in low efficiency of the algorithm.

For reducing the dimensions of each feature for annotation, an extended local sensitive discriminant analysis algorithm is proposed by constructing relevant and irrelevant graphs in [29]. Generally, feature dimension reduction methods based on NMF decomposition are for single-view features. References [30, 31] extend this method to multiview features by simply concatenating multiple vectors into one feature vector before further dimension reduction. However, this concatenation way can cause vector dimension disaster. Besides, multiview features are descriptions from different views for images so that simple connection does not make good sense. Then a multiview NMF model based on shared coefficient matrix is developed for capturing the latent feature patterns in multiview features [32], where different view features have their own basis matrices and share a coefficient matrix. The proposed model is used for solving classification and clustering problems and is not suitable for multilabel problems with multilabel images.

Based on the above reviews, this paper proposes a semisupervised learning model based on multiview NMF and graph embedding. A novel multiview NMF algorithm based on graph embedding is developed to fuse the multiview features and reduce the dimension of the fused features by designing appropriate graph embedded regularization terms. Then, the image annotation is performed by using the new features through a KNN-based algorithm.

3. The Proposed Methods

In this section, we elaborate the proposed semisupervised framework for automatic image annotation. First, the graph embedding terms for multilabel problems are constructed through semantic similarity matrix. Second, an objective function is established by adding graph embedded semantic constraints. Third, the update rules for optimizing are derived in detail. Finally, the overall framework of the algorithm is presented.

3.1. Graph Embedding for Multilabel Problem. The traditional graph embedding model is introduced for classification problems, in which each sample has only one label, so that the Laplacian matrices L and L^p can be given according to whether they belong to the same category or not. However, for multilabel problems, a sample usually contains multiple category labels. Therefore, traditional graph embedding methods cannot be directly applied to multilabel problems. In this paper, we give a relation matrix according to whether samples are related or not. By setting appropriate thresholds, the relevant matrix and the irrelevant matrix can be obtained, and they can be used to calculate Laplacian matrices L and L^p, respectively.

Let $\{x_i, y_i\}$ denote the i-th sample and $Y \in \mathbb{R}^{n_1 * m}$ denote label matrix, where n_1 is the number of samples in the training set, m is the number of labels, y_i represents the i-th row of Y, and $y_{:i}$ represents the i-th column. The semantic similarity between sample i and sample j can be formulated as $y_i C y_j$, where C is a priori label relation matrix similar to that in [33].

$$C_{ij} = \cos\left(y_{:i}, y_{:j}\right) = \frac{\left\langle y_{:i}, y_{:j} \right\rangle}{\left\| y_{:i} \right\| \left\| y_{:j} \right\|} \tag{1}$$

$y_{:i} \in \mathbb{R}^{n1*1}$ denotes the sample vector and $\|y_{:i}\|$ denotes the L2-norm of $y_{:i}$. Then, the semantic similarity matrix of samples can be obtained by the following formula:

$$\left(W^s\right)_{ij} = y_i C y_j \tag{2}$$

Given thresholds T_u and T_l ($T_u \geq T_l$), samples with similarity greater than T_u are relevant, and samples with similarity less than T_l are irrelevant. Therefore, the relevant matrix W and the irrelevant matrix W^p are constructed as follows:

$$W_{ij} = \begin{cases} W_{ij}^s, & W_{ij}^s > T_u \\ 0, & W_{ij}^s \leq T_u \end{cases} \tag{3}$$

$$W_{ij}^p = \begin{cases} 1, & W_{ij}^s \leq T_l \\ 0, & W_{ij}^s > T_l \end{cases} \tag{4}$$

The corresponding Laplacian matrices are formulated as follows:

$$L = D - W \tag{5}$$

$$L^p = \mathrm{D}^p - W^p \tag{6}$$

where $D_{jj} = \sum_l W_{jl}$ and $D_{jj}^p = \sum_l W_{jl}^p$.

Having the relevant and irrelevant matrices, the following two constraint items C_1 and C_2 are incorporated to make feature representations in the new feature space consist with semantic concepts:

$$\begin{aligned} C_1 &= \sum_{i,j=1}^{n_1} \left\| v_i - v_j \right\|^2 W_{ij} = \sum_{i=1}^{n_1} v_i^T v_i D_{ii} - \sum_{i,j=1}^{n_1} v_i^T v_j W_{ij} \\ &= Tr\left(V^T D V\right) - Tr\left(V^T W V\right) = Tr\left(V^T L V\right) \end{aligned} \tag{7}$$

$$\begin{aligned} C_2 &= \sum_{i,j=1}^{n_1} \left\| v_i - v_j \right\|^2 W_{ij}^p = \sum_{i=1}^{n_1} v_i^T v_i D_{ii}^p - \sum_{i,j=1}^{n_1} v_i^T v_j W_{ij}^p \\ &= Tr\left(V^T D^p V\right) - Tr\left(V^T W^p V\right) = Tr\left(V^T L^p V\right) \end{aligned} \tag{8}$$

where n_1 denotes the number of samples in the training set and v_i and v_j represent the visual feature vectors of sample i and sample j, respectively.

3.2. An Automatic Image Annotation Model Based on Multiview Feature NMF and Graph Embedding. Let $X = \{X^{(v)}\}_{v=1}^M$ denote the data matrix, where $X^{(v)} \in \mathbb{R}^{D^{(v)} * N}$ is the feature matrix corresponding to the v-th view, $D^{(v)}$ is the dimension

of feature vectors, M is the number of views, and N is the number of samples. The objective function can be formulated as

$$O_1 = \sum_{v=1}^{M} \left\| X^{(v)} - U^{(v)} V^T \right\|^2 \quad \text{s.t.} \quad u_{ij}^{(v)} \geq 0, \quad v_{ij} \geq 0 \tag{9}$$

where $U^{(v)} \in \mathbb{R}^{D^{(v)} * K}$ and $V \in \mathbb{R}^{N * K}$ are nonnegative matrices and K denotes the dimension of the new low-dimensional feature.

Furthermore, graph embedding regularization terms (7) and (8) are combined with the above loss function, then

$$O_1 = \sum_{v=1}^{M} \left\| X^{(v)} - U^{(v)} V^T \right\|^2 + Tr\left(\left(V^l\right)^T \check{L} V^l \right) \quad \text{s.t.} \quad u_{ij} \geq 0, \quad v_{ij} \geq 0 \tag{10}$$

where $\check{L} = (\alpha L - \beta L^p)$ and α and β are two equilibrium coefficients. Equation (10) consists of two terms, where the first is the error term, and the second is the constraint term that makes semantic constrains on V by using graph embedding regularization. It implies that the semantic related sample features are closer and vice versa. It is worth noting that the model is semisupervised since that V^l refers to data with labels, and the graph embedding term is used to constrain V^l.

3.3. Update Rules Derivation. The established model is semisupervised, and only part of the data has label information. The objective function can be rewritten in the form of block matrix. The following subsection will give the derivation of update rules.

The update rule of formula (10) is derived as follows:

$$\begin{aligned} O_1 &= \sum_{v=1}^{M} Tr\left(\left(X^{(v)} - U^{(v)} V^T\right) \left(X^{(v)} - U^{(v)} V^T\right)^T \right) \\ &\quad + \alpha Tr\left(\left(V^l\right)^T L V^l \right) - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) \\ &= \sum_{v=1}^{M} \left[Tr\left(X^{(v)} X^{(v)T} \right) - 2Tr\left(X^{(v)} V U^{(v)T} \right) \right. \\ &\quad \left. + Tr\left(U^{(v)} V^T V U^{(v)T} \right) \right] + \alpha Tr\left(\left(V^l\right)^T L V^l \right) \\ &\quad - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) \end{aligned} \tag{11}$$

Let $\psi_{ij}^{(v)}$ and φ_{ij} be the Lagrange multipliers of constraint conditions $u_{ij}^{(v)} \geq 0$ and $v_{ij} \geq 0$, respectively, $\Psi^{(v)} = [\psi_{ij}^{(v)}]$, $\Phi = [\varphi_{ij}]$. Then the Lagrange function can be written as

$$\begin{aligned} L &= \sum_{v=1}^{M} \left[Tr\left(X^{(v)} X^{(v)T} \right) - 2Tr\left(X^{(v)} V U^{(v)T} \right) \right. \\ &\quad \left. + Tr\left(U^{(v)} V^T V U^{(v)T} \right) + Tr\left(\Psi^{(v)} U^{(v)T} \right) \right] \\ &\quad + \alpha Tr\left(\left(V^l\right)^T L V^l \right) - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) \\ &\quad + Tr\left(\Phi V^T \right) \end{aligned} \tag{12}$$

The partial derivative of L with respect to $U^{(v)}$ is as follows:

$$\frac{\partial L}{\partial U^{(v)}} = -2X^{(v)} V + 2U^{(v)} V^T V + \Psi^{(v)} \tag{13}$$

where $X^{(v)} = [X^{(v)l}, X^{(v)u}]$, $V = [(V^l)^T, (V^u)^T]^T$, and $\Phi = [(\Phi^l)^T, (\Phi^u)^T]^T$, the symbol l means labelled and the symbol u means unlabelled. Thus, $X^{(v)l}$ and V^l refer to the data with labels. Then (12) can be rewritten as

$$\begin{aligned} L &= \sum_{v=1}^{M} \left[Tr\left(X^{(v)} X^{(v)T} \right) - 2Tr\left(\left[X^{(v)l}, X^{(v)u}\right] \right. \right. \\ &\quad \left. \cdot \left[\left(V^l\right)^T, \left(V^u\right)^T \right]^T U^{(v)T} \right) \\ &\quad + Tr\left(U^{(v)} \left[\left(V^l\right)^T, \left(V^u\right)^T \right] \left[\left(V^l\right)^T, \left(V^u\right)^T \right]^T \right. \\ &\quad \left. \left. \cdot U^{(v)T} \right) + Tr\left(\Psi^{(v)} U^{(v)T} \right) \right] + \alpha Tr\left(\left(V^l\right)^T L V^l \right) \\ &\quad - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) + Tr\left(\left[\left(\Phi^l\right)^T, \left(\Phi^u\right)^T \right]^T \right. \\ &\quad \left. \cdot \left[\left(V^l\right)^T, \left(V^u\right)^T \right] \right) = \sum_{v=1}^{M} \left[Tr\left(X^{(v)} X^{(v)T} \right) \right. \\ &\quad - 2Tr\left(X^{(v)l} V^l U^{(v)T} \right) - 2Tr\left(X^{(v)u} V^u U^{(v)T} \right) \\ &\quad + Tr\left(U^{(v)} \left(V^l\right)^T V^l U^{(v)T} \right) + Tr\left(U^{(v)} \left(V^u\right)^T \right. \\ &\quad \left. \left. \cdot V^u U^{(v)T} \right) + Tr\left(\Psi^{(v)} U^{(v)T} \right) \right] + \alpha Tr\left(\left(V^l\right)^T L V^l \right) \\ &\quad - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) + Tr\left(\Phi^l \left(V^l\right)^T \right. \\ &\quad \left. + \Phi^u \left(V^u\right)^T \right). \end{aligned} \tag{14}$$

Separating the terms associated with V^l and V^u, the above equation can be written as

$$L = L\left(V^l\right) + L\left(V^u\right) \tag{15}$$

$$\begin{aligned} L\left(V^l\right) &= \sum_{v=1}^{M} \left[-2Tr\left(X^{(v)l} V^l U^{(v)T} \right) \right. \\ &\quad \left. + Tr\left(U^{(v)} \left(V^l\right)^T V^l U^{(v)T} \right) \right] + \alpha Tr\left(\left(V^l\right)^T L V^l \right) \\ &\quad - \beta Tr\left(\left(V^l\right)^T L^p V^l \right) + Tr\left(\Phi^l \left(V^l\right)^T \right) + const \end{aligned} \tag{16}$$

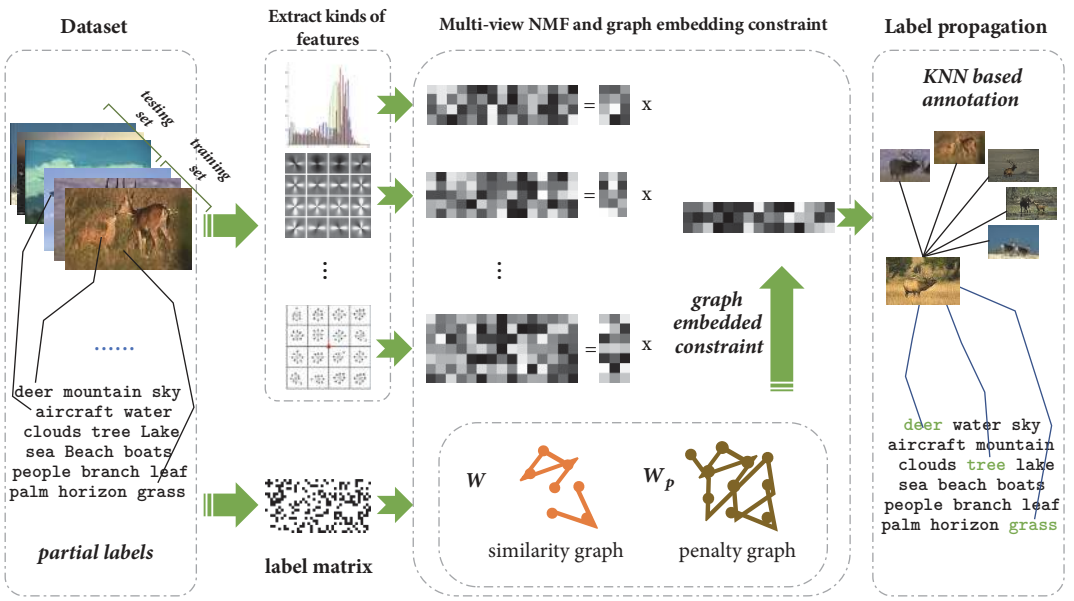


Figure 1: Schematic diagram of the GENMF model.

$$L\left(V^u\right) = \sum_{v=1}^{M}\left[-2Tr\left(X^{(v)u}V^u U^{(v)T}\right) + Tr\left(U^{(v)}\left(V^u\right)^T V^u U^{(v)T}\right)\right] + Tr\left(\Phi^u\left(V^u\right)^T\right) + const \tag{17}$$

The partial derivatives of L with respect to V^l and V^u are as follows:

$$\frac{\partial L}{\partial V^l} = \sum_{v=1}^{M}\left[-2\left(X^{(v)l}\right)^T U^{(v)} + 2V^l U^{(v)T}U^{(v)}\right] + 2\alpha LV^l - 2\beta L^p V^l + \Phi^l \tag{18}$$

$$\frac{\partial L}{\partial V^u} = \sum_{v=1}^{M}\left[-2\left(X^{(v)u}\right)^T U^{(v)} + 2V^u U^{(v)T}U^{(v)}\right] + \Phi^u \tag{19}$$

Using the KKT conditions $\psi_{ij}^{(v)}u_{ij}^{(v)} = 0$ and $\varphi_{ij}v_{ij} = 0$ (i.e., $\psi_{ij}^{(v)} = 0$ and $\varphi_{ij} = 0$), consider formulae (13), (18), and (19) and let the derivatives equal 0; the following three equations can be obtained:

$$-\left(X^{(v)}V\right)_{ij}u_{ij} + \left(U^{(v)}V^TV\right)_{ij}u_{ij} = 0 \tag{20}$$

$$\sum_{v=1}^{M}\left[-2\left(X^{(v)l}\right)^T U^{(v)} + 2V^l U^{(v)T}U^{(v)}\right]_{ij}v_{ij} + \left(\alpha LV^l - \beta L^pV^l\right)_{ij}v_{ij} = 0 \tag{21}$$

$$\sum_{v=1}^{M}\left[-2\left(X^{(v)u}\right)^T U^{(v)} + 2V^u U^{(v)T}U^{(v)}\right]_{ij}v_{ij} = 0 \tag{22}$$

The following update rules can be obtained through the above three equations:

$$u^{(v)}{}_{ik} \longleftarrow u^{(v)}{}_{ik}\frac{\left(X^{(v)}V\right)_{ik}}{\left(U^{(v)}V^TV\right)_{ik}} \tag{23}$$

$$v^l{}_{jk} \longleftarrow v^l{}_{jk}\frac{\left(\sum_{v=1}^{M}\left(X^{(v)l}\right)^T U^{(v)} + \alpha WV^l + \beta D^pV^l\right)_{jk}}{\left(\sum_{v=1}^{M}V^l\left(U^{(v)}\right)^T U^{(v)} + \alpha DV^l + \beta W^pV^l\right)_{jk}} \tag{24}$$

$$v^u{}_{jk} \longleftarrow v^u{}_{jk}\frac{\left(\sum_{v=1}^{M}\left(X^{(v)u}\right)^T U^{(v)}\right)_{jk}}{\left(\sum_{v=1}^{M}V^u\left(U^{(v)}\right)^T U^{(v)}\right)_{jk}} \tag{25}$$

It is mentioned in [34] that in order to ensure the convexity of the loss function, β needs to be taken as an appropriately small value, which is suggested by $\beta = 10^{-4}$. Besides, [35] gives a modified strategy to the original update rules to ensure convergence. The same strategy can be applied to the derived update rules.

3.4. Framework of the GENMF. The schematic diagram of the proposed GENMF model can be illustrated as in Figure 1. First, multiview features are extracted from images as the input matrix X in (10). Equations (1)-(8) are utilized to build graph embedding regularization terms as the input matrices L and L^p in (10). Then, $U^{(v)}$ and V are updated iteratively by using updated equations (23) to (25) until the maximum number of iterations is reached or the loss value is within the permissible range. Finally, the new features V^u of the test set and the training set features V^l are input to the KNN-based labelling algorithm to obtain the predicted labels. The flowchart of the algorithm is shown in Figure 2.

Input: Image set $\boldsymbol{I} = [\boldsymbol{I}_{train}, \boldsymbol{I}_{test}]$ and label matrix $\boldsymbol{Y}^{\boldsymbol{I}}$ of the training set.
Output: Predicted label matrix $\boldsymbol{Y}^{\boldsymbol{pre}}$ for the test set $\boldsymbol{I}_{test}$.
(1) Extract different feature $\boldsymbol{X}^{(v)} \in \mathbb{R}_{+}^{D^{(v)} * N}$ for image set $\boldsymbol{I}$;
(2) Construct Laplacian graph $\boldsymbol{L}$ and $\boldsymbol{L}^{\boldsymbol{p}}$;
(3) Initialize $\boldsymbol{U}^{(v)} \in \mathbb{R}_{+}^{D^{(v)} * K}$ and $\boldsymbol{V} \in \mathbb{R}_{+}^{N * K}$ randomly;
(4) do
(5) Update $\boldsymbol{U}^{(v)}$ and $\boldsymbol{V}$ based on equation (23)-(25);
(6) while the terminating condition is not satisfied
(7) Input $\boldsymbol{V}$ into 2PKNN [21] image annotation algorithm;
(8) Output predicted labels $\boldsymbol{Y}^{\boldsymbol{pre}}$ for the test set.

Algorithm 1: Multiview NMF with graph embedding for image annotation.

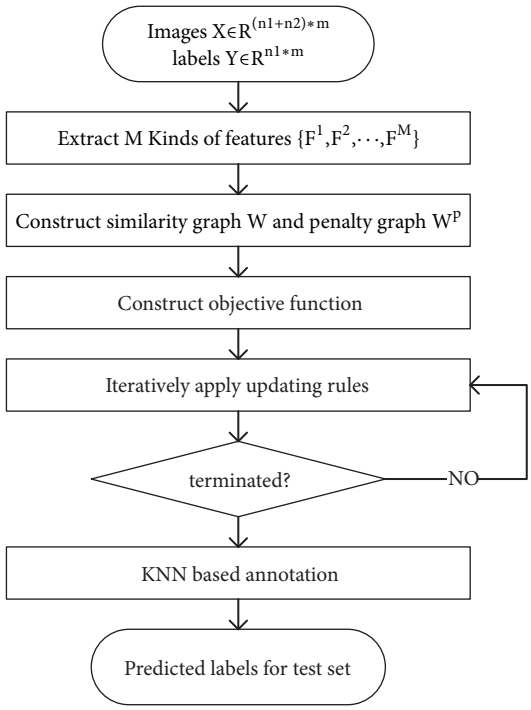


Figure 2: Flowchart of the GENMF.

Algorithm 1 gives the pseudocode of the GENMF.

4. Experimental Studies

4.1. Dataset and Experiment Design. The main purpose of the proposed algorithm is to improve the performance of automatic image annotation by fusing the multiview features and reducing the feature dimension, which makes it better to represent semantic concepts under semantic constraints in new low-dimensional feature spaces. So this paper selects the dataset Corel5k with 15 different features, and Corel5k consists of 4500 images for training and 499 images for test, which is available on http://lear.inrialpes.fr. The 15 features are all low-level image features including Gist, DenseSift, DenseSiftV3H1, HarrisSift, HarrisSiftV3H1, DenseHue, DenseHueV3H1, HarrisHue, HarrisHueV3H1, Rgb, RgbV3H1, Lab, LabV3H1, Hsv, and HsvV3H1. In the experiment, we select a local feature DenseSiftV3H1, a global feature Gist, and a color feature Hsv.

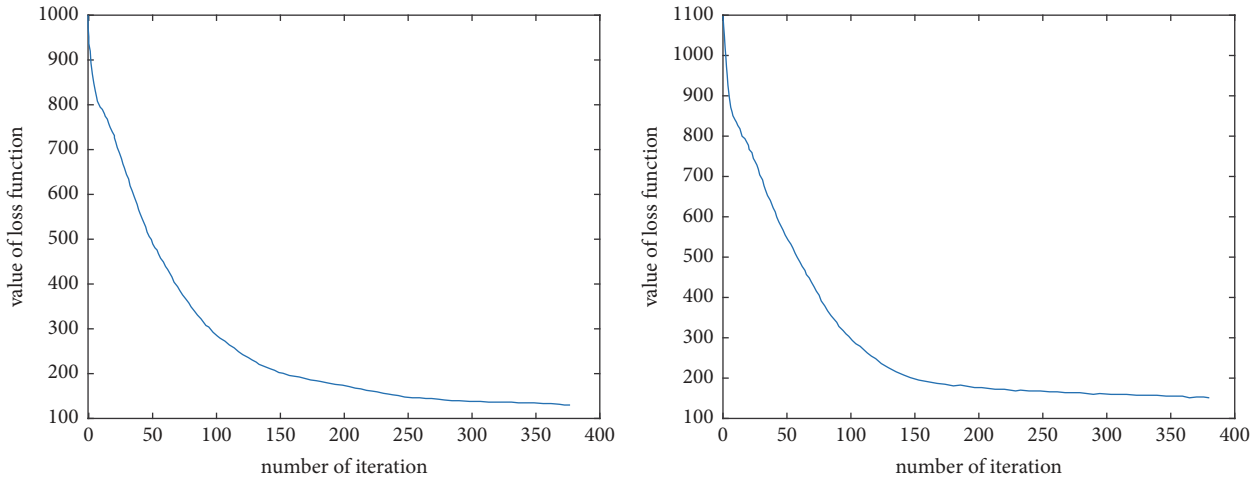
In the experiments, the multiple features except Gist are regularized through L2-normalization, and the normalized features are input into the GENMF to obtain low-dimensional representations. Then the low-dimensional feature vectors are input into the 2PKNN annotation algorithm to obtain the predicted labels for the test set. The performance of the algorithm is evaluated in terms of four metrics Pre, Rec, F1, and N+. Table 1 lists the parameters used in the experiments.

4.2. Experimental Results

4.2.1. Convergence Curve of Loss Function. Figure 3 shows the convergence curves of loss function with different parameters. It can be observed that, after about 300 iterations, the trend of the loss curve tends to be stable.

4.2.2. The Influence of Different T_u and T_l. The relation matrix $W^s \in \mathbb{R}^{4500*4500}$ can be established according to formula (2). Observed by experimental methods, the maximum value of W^s is 12.9554 and the minimum value of W^s is 0. The values of $T_u = \{1, 2, \ldots, 10\}$ and $T_l = \{0, \ldots, T_u\}$ are traversed, where $T_u \geq T_l$. Figure 4 shows the changes in the performance of the annotation when the different values of parameters are selected. On the whole, when $T_u = 2$ and $T_l = 1$, the algorithm obtains the highest F1 value. Thus, in the following experiments T_u is taken as 2 and T_l is taken as 1.

4.2.3. The Influence of Different α. Figure 5 shows the varying curve of Pre, Rec, F1, and N+ in the case of $K = 300$ with different α values. Figure 5-1 shows that the annotation accuracy increases first and then decreases with the increase of α. When $\alpha = 1000$, the accuracy reaches the highest value. Figure 5-2 shows that the recall rate generally increases first and then decreases. When $\alpha = 2000$, the recall rate reaches the highest value. From Figure 5-3, it can be seen that the F1 value also increases first and then decreases with the increase of α, but a concave point appears at $\alpha = 1500$. When $\alpha = 1000$, the F1 value reaches the highest value. In Figure 5-4, the N+ value fluctuates in the interval $[0, 1500]$, and its value reaches the highest value at $\alpha = 2000$ and decreases afterwards.



(a) $\alpha = 1000, K = 300$ (b) $\alpha = 1500, K = 400$

Figure 3: Convergence curves of loss function.

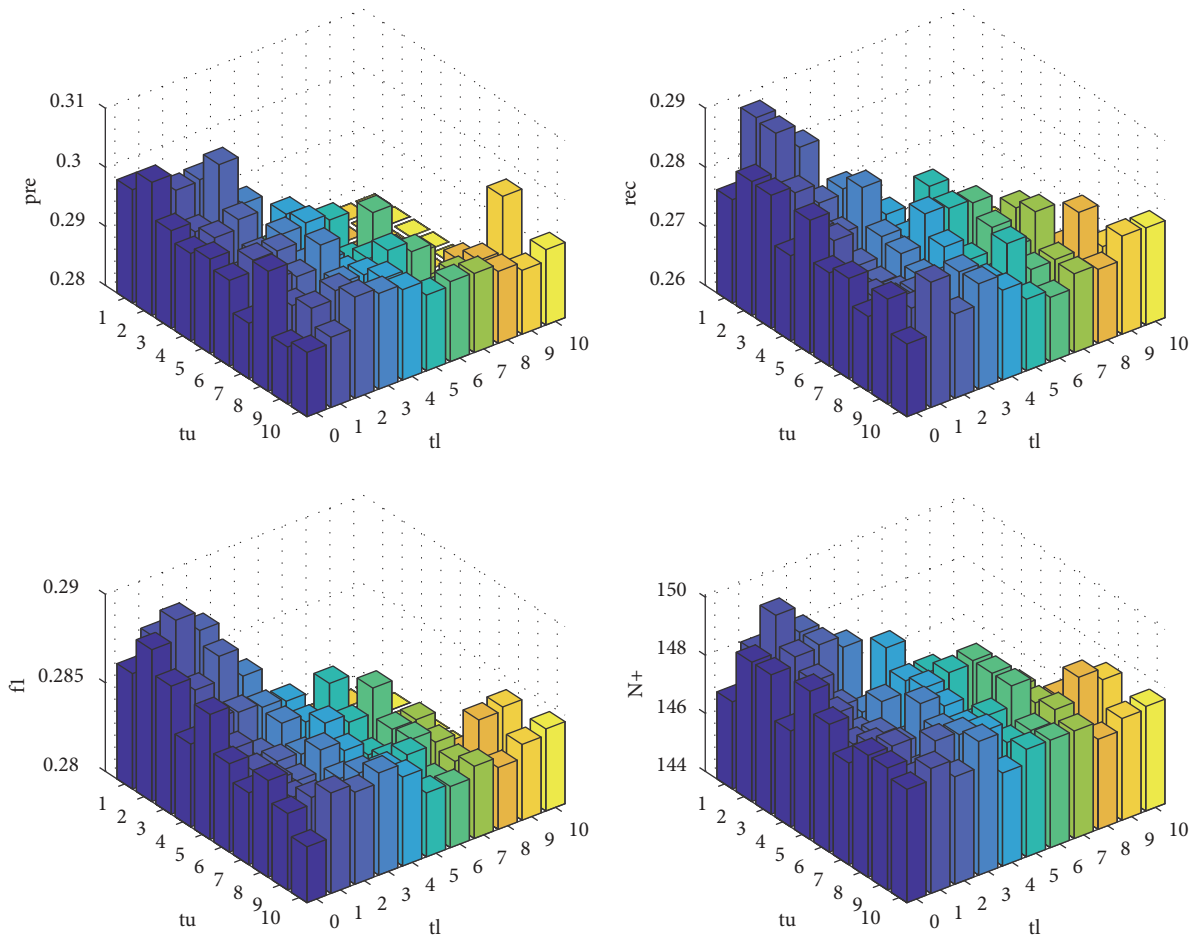


Figure 4: Impact of different values for T_u and T_l.

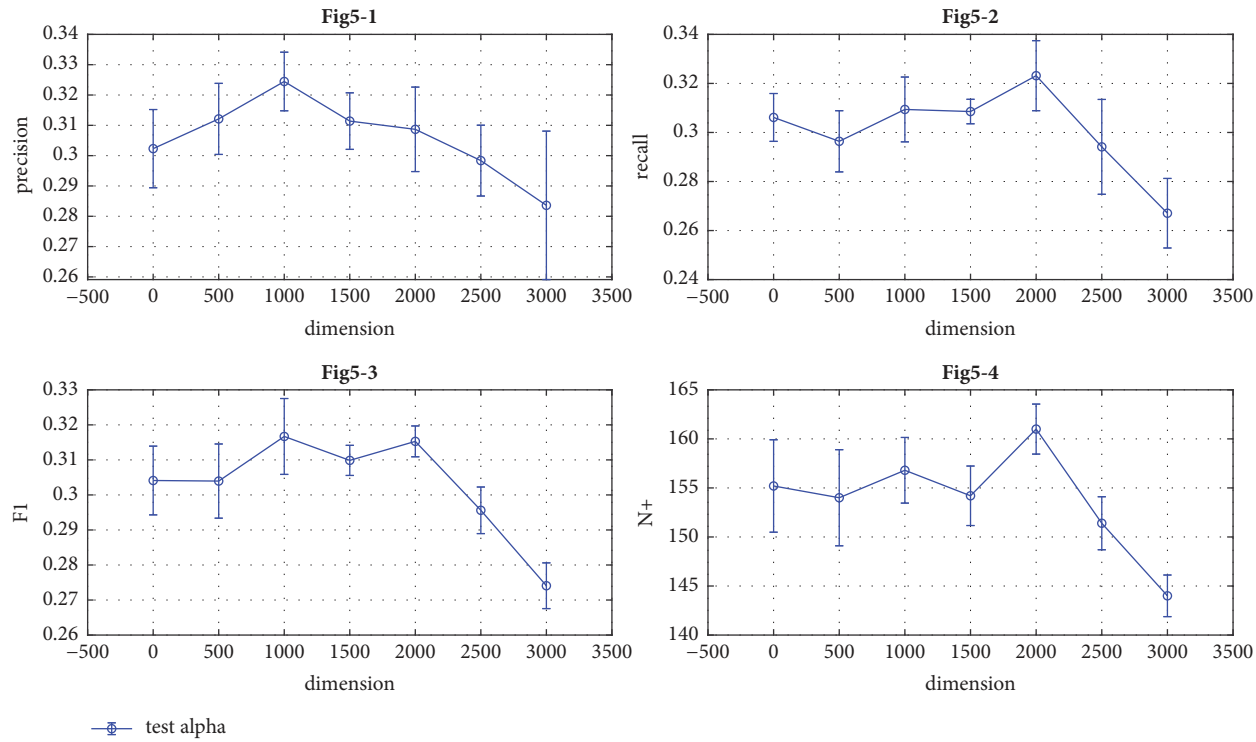


Figure 5: Curve of Pre, Rec, F1, and N+ with different α values.

Table 1: Parameters required in the algorithm and their ranges of values.

Notation	Description	Range of values
$\boldsymbol{\alpha}$	Weight for graph embedding terms	{0, 500, 1000, 1500, 2000, 2500, 3000}
$\boldsymbol{K}$	Dimension of the new features	{100, 200, 300, 400, 500, 600, 700, 800}
$\boldsymbol{T_u}$	Label-relevant coefficient	$\{1, 2, \ldots, 10\}$
$\boldsymbol{T_l}$	Label-irrelevant coefficient	$\{0, 1, \ldots, T_u\}$

4.2.4. The Influence of Different Feature Dimensions K. Figure 6 shows the annotation performance curves when α is taken as 0, 1000, and 2000, respectively, and the value of K changes from 100 to 800 with an increase of 100 each time. The three curves with different values of parameter α show the consistent trend of change. In Figure 6-1, the accuracy increases with the increase of dimension because more information can be retained, and the curve becomes stable until α reaches 2000. The worst performance is at α = 0. Figure 6-2 shows that the recall rate decreases slightly with the increase of dimension because the requirement for retrieval is higher with the increase of dimension. In Figure 6-3, F1 is reflecting the comprehensive effect of the accuracy and recall rate. It can be observed that the F1 increases in the interval [100, 300] with the increase of dimension and then tends to be stable except for α = 0. Figure 6-4 shows that N+ value fluctuates but the overall trend is stable. In general, the performance of proposed algorithm on four metrics outperforms using the original features when α = 1000 or α = 2000 with dimension in the range of [200-800].

4.2.5. Comparison with Existing Annotation Algorithms. Table 2 presents the comparison results with existing annotation algorithms. RMLF [36] optimizes the final prediction tag score by fusing prediction tag scores of 15 different features. LDMKL [14] and SDMKL [14] use the different classifiers based on the nonlinear kernel of three-layer network to annotate images. 2PKNN [22] uses two steps for annotation: after dealing with data imbalance, images are annotated through a KNN-based method in data-balanced dataset. LJNMF [31], merging features [31], and Scoefficients [31] consider different kinds of NMF modeling, extract new features, and annotate images through a KNN-based method. TagProp (ML) [21] and TagProp (σML) [21] acquire discriminative feature fusion on the training set by designing a metric learning model and annotate images using weighted KNN method. JEC [37] is a KNN-based algorithm based on the average distance of multiple features, which is a benchmark algorithm for image annotation. MRFA [38] proposes a new semantic context modeling and learning method based on multimarkov random fields. SML [39] is a discriminative model that treats each label as one class in multiclass classification problems;

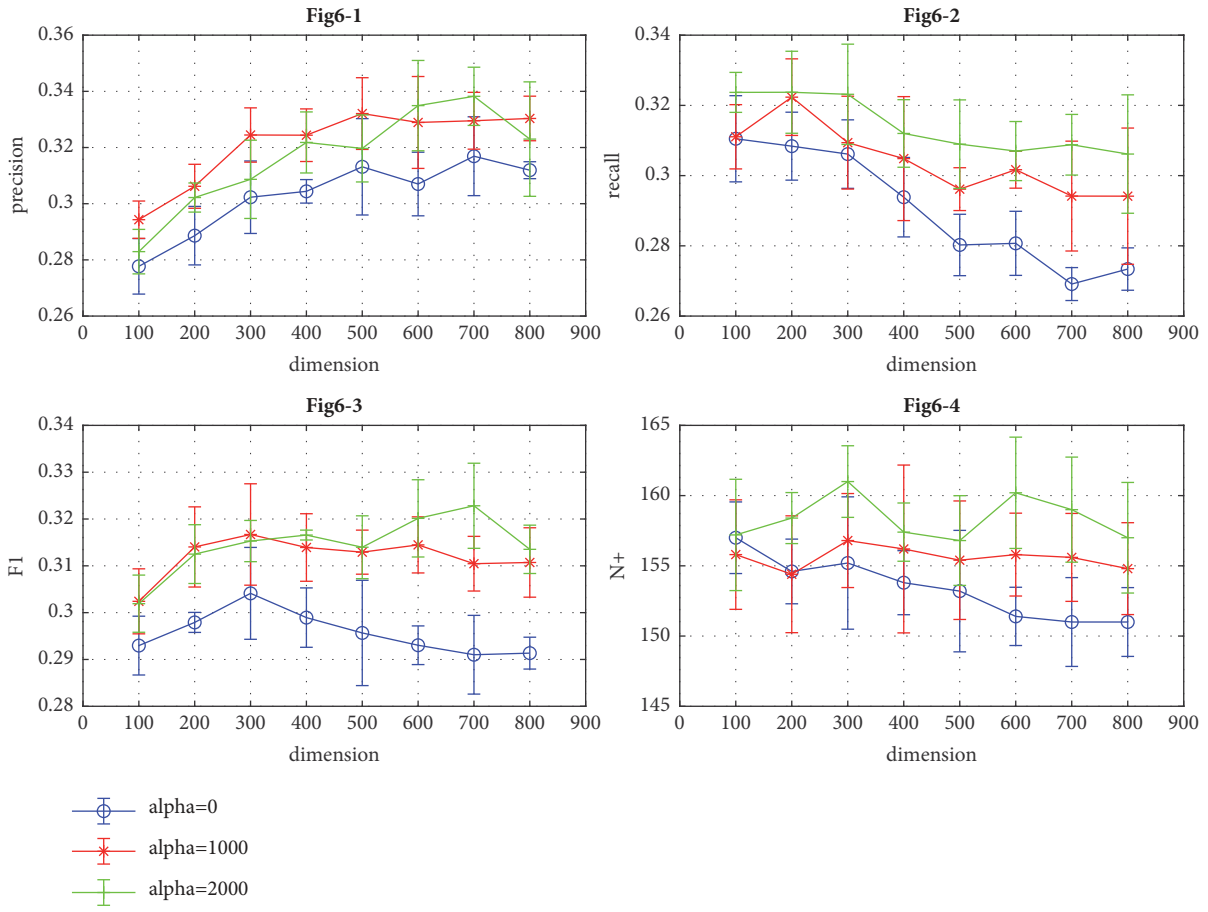


Figure 6: Annotation performance curves for different values of K.

Table 2: Comparison results with other annotation algorithms.

Methods	Pre	Rec	F1	N+
SML	23	29	25.7	137
JEC	27	32	29.3	139
GS	30	33	31.4	146
MRFA	31	36	33.3	172
TagProp(ML)	31	37	33.7	146
TagProp(σML)	33	42	37.0	160
RMLF	29.7	32.6	31.1	-
Merging features	33	40	36.5	-
Scoefficients	30	39	34.6	-
LJNMF(3f')	35	43	39.1	-
2PKNN(3f)	32	28	30.6	177
SDMKL	38	25	30	158
LDMKL	44	29	34.9	179
GENMF (3f)	38	39	39.2	168

GS [38] introduces the regularization-based feature selection algorithm to exploit the sparsity and clustering properties of features.

In Table 2, the note (3f) denotes using the three features selected in this paper, and the note (3f') indicates using three features that are not the same as in this paper. The results of other algorithms are directly taken from respective literatures and all the 15 features are utilized. Our algorithm uses only three features, and it can be seen in Table 2 that the proposed GENMF achieves the competitive performance.

Table 3: The maximum, mean, and standard deviation of results using 10 independent runs.

metrics	Precision	Recall	F1	N+
mean	0.38	0.39	0.392	168
SD	0.017	0.010	0.012	4.50
maximum	0.41	0.40	0.398	175

4.2.6. The Best, Average, and Standard Deviation of the Results. Table 3 shows the best, average, and standard deviation of the results using 10 independent runs. The NMF-based algorithms have a certain randomness, and different initial values may produce different results. Table 3 shows that the influence of different initialization values is limited, but better performance could be expected if a better initialization strategy is chosen. Besides, the average time consumption of the proposed GENMF with the new low-dimensional features is 13.945 seconds to label all 499 test images, whereas utilizing the original features to label takes 34.652 seconds, which is about 2.5 times that of GENMF.

5. Conclusions

In this paper, we propose a semisupervised framework based on graph embedding and multiview nonnegative matrix factorization for automatic image annotation with multilabel images. The main purpose of the proposed algorithm is to improve the performance of automatic image annotation by fusing multiview features and reducing feature dimension, which makes it better to represent semantic concepts under semantic constraints in new low-dimensional feature spaces. For feature fusion and dimension deduction, a novel graph embedding term is constructed based on the relevant graph and the irrelevant graph. Then, the fusion of multiview features and the reduction of dimensionality are realized based on multiview NMF model. Moreover, the updated rules of the model are derived. Finally, images are annotated by using a KNN-based approach. Experimental results validate that the proposed algorithm can achieve competitive performance in terms of accuracy and efficiency.

Data Availability

The code used in this paper is released, which is written in Matlab and available at https://github.com/MenSanYan/image-annotation.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are grateful to the support of the National Natural Science Foundation of China (61572104, 61103146, 61425002, and 61751203), the Fundamental Research Funds for the Central Universities (DUT17JC04), and the Project of the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (93K172017K03).

References

[1] L. Mai, H. Jin, Z. Lin, C. Fang, J. Brandt, and F. Liu, "Spatial-Semantic Image Search by Visual Feature Synthesis," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1121–1130, Honolulu, HI, USA, July 2017.

[2] H. Guan and W. A. Smith, "BRISKS: Binary Features for Spherical Images on a Geodesic Grid," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4516–4524, Honolulu, HI, USA, July 2017.

[3] Y. Zhang, W. Lin, Q. Li, W. Cheng, and X. Zhang, "Multiple-level feature-based measure for retargeted image quality," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 451–463, 2018.

[4] F. Zhang and B. W. Wah, "Fundamental principles on learning new features for effective dense matching," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 822–836, 2018.

[5] H. Zhang, V. M. Patel, and R. Chellappa, "Hierarchical Multimodal Metric Learning for Multimodal Classification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3057–3065, Honolulu, HI, USA, July 2017.

[6] P. Li, Q. Wang, H. Zeng, and L. Zhang, "Local Log-Euclidean Multivariate Gaussian Descriptor and Its Application to Image Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 803–817, 2017.

[7] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2355–2368, 2015.

[8] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 323–330, 2004.

[9] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Proceedings of the 2005 Annual Conference on Neural Information Processing Systems, NIPS 2005*, pp. 355–362, December 2005.

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

[11] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information Fusion in Multimedia Information Retrieval," in *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, vol. 4918 of *Lecture Notes in Computer Science*, pp. 147–159, Springer Berlin Heidelberg, Berlin, Germany, 2008.

[12] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*, pp. 399–402, ACM, November 2005.

[13] Y. Gu, X. Qian, Q. Li, M. Wang, R. Hong, and Q. Tian, "Image Annotation by Latent Community Detection and Multikernel Learning," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3450–3463, 2015.

[14] M. Jiu and H. Sahbi, "Nonlinear Deep Kernel Learning for Image Annotation," in *Proceedings of the IEEE International*

Conference on Acoustics, Speech and Signal Processing, pp. 1551–1555, 2017.

[15] L. Sun, H. Ge, S. Yoshida, Y. Liang, and G. Tan, "Support vector description of clusters for content-based image annotation," *Pattern Recognition*, vol. 47, no. 3, pp. 1361–1374, 2014.

[16] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011*, pp. 1609–1614, July 2011.

[17] M. Zand, S. Doraisamy, A. Abdul Halin, and M. R. Mustaffa, "Visual and semantic context modeling for scene-centric image annotation," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8547–8571, 2017.

[18] D. Tian and Z. Shi, "Automatic image annotation based on Gaussian mixture model considering cross-modal correlations," *Journal of Visual Communication and Image Representation*, vol. 44, pp. 50–60, 2017.

[19] J. Tian, Y. Huang, Z. Guo, X. Qi, Z. Chen, and T. Huang, "A multi-modal topic model for image annotation using text analysis," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 886–890, 2015.

[20] K. Pliakos and C. Kotropoulos, "PLSA driven image annotation, classification, and tourism recommendation," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 3003–3007, 2014.

[21] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 309–316, IEEE, Kyoto, Japan, September-October 2009.

[22] Y. Verma and V. Jawahar C, "Image Annotation Using Metric Learning in Semantic Neighbourhoods," in *Proceedings of the European Conference on Computer Vision*, pp. 836–849, 2012.

[23] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 184–191, IEEE, Columbus, OH, USA, June 2014.

[24] Z. Chen, M. Chen, and K. Q. Weinberger, "Marginalized denoising for link prediction and multi-label learning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1707–1713, 2015.

[25] S. Hamid Amiri and M. Jamzad, "Efficient multi-modal fusion on supergraph for scalable image annotation," *Pattern Recognition*, vol. 48, no. 7, pp. 2241–2253, 2015.

[26] Z. He, C. Chen, J. Bu, P. Li, and D. Cai, "Multi-view based multi-label propagation for image annotation," *Neurocomputing*, vol. 168, no. C, pp. 853–860, 2015.

[27] F. Su and L. Xue, "Graph learning on K nearest neighbours for automatic image annotation," in *Proceedings of the 5th ACM International Conference on Multimedia Retrieval, ICMR 2015*, pp. 403–410, June 2015.

[28] S. Hasija, M. J. Buragohain, and S. Indu, "Fish species classification using graph embedding discriminant analysis," in *Proceedings of the 2017 International Conference on Machine Vision and Information Technology, CMVIT 2017*, pp. 81–86, February 2017.

[29] X. Liu, R. Liu, F. Li, and Q. Cao, "Graph-based dimensionality reduction for KNN-based image annotation," in *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012*, pp. 1253–1256, 2013.

[30] J. BenAbdallah, J. C. Caicedo, F. A. Gonzalez, and O. Nasraoui, "Multimodal image annotation using non-negative matrix factorization," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, pp. 128–135, September 2010.

[31] R. Rad and M. Jamzad, "Automatic image annotation by a loosely joint non-negative matrix factorisation," *IET Computer Vision*, vol. 9, no. 6, pp. 806–813, 2015.

[32] Z. Guan, L. Zhang, J. Peng, and J. Fan, "Multi-View Concept Learning for Data Representation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3016–3028, 2015.

[33] H. Wang, C. Ding, and H. Huang, "Multi-label Linear Discriminant Analysis," in *Proceedings of the European Conference on Computer Vision*, pp. 126–139, 2010.

[34] N. Guan, X. Huang, L. Lan, Z. Luo, and X. Zhang, "Graph based semi-supervised non-negative matrix factorization for document clustering," in *Proceedings of the 11th IEEE International Conference on Machine Learning and Applications, ICMLA 2012*, pp. 404–408, 2013.

[35] C.-J. Lin, "On the convergence of multiplicative update algorithms for non-negative matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 18, no. 6, pp. 1589–1596, 2007.

[36] Y. Yao, X. Xin, and P. Guo, "A rank minimization-based late fusion method for multi-label image annotation," in *Proceedings of the 23rd International Conference on Pattern Recognition, ICPR 2016*, pp. 847–852, 2017.

[37] A. Makadia, V. Pavlovic, and S. Kumar, "A New Baseline for Image Annotation," in *Proceedings of the European Conference on Computer Vision*, pp. 316–329, 2008.

[38] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo, "A revisit of generative model for automatic image annotation using markov random fields," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 1153–1160, June 2009.

[39] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.