IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 4, 2019, accepted June 19, 2019, date of publication July 1, 2019, date of current version July 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925891

Cograph Regularized Collective Nonnegative Matrix Factorization for Multilabel Image Annotation

JULI ZHANG[1], ZHANZHUANG HE[1], JUNYI ZHANG[2,3], AND TAO DAI[2]

[1]Department of Research and Development, Xi'an Microelectronics Technology Institute, Xi'an 710068, China
[2]School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[3]China Minsheng Bank, Beijing 110000, China

Corresponding author: Juli Zhang (juli2320@sina.com)

This work was supported in part by the China Youth Science Foundation under Grant 61702413, and in part by the Technology Innovation Funds for the Ninth Academy of China Aerospace under Grant 2016JY06.

ABSTRACT Automatic image annotation is an effective and straightforward way to facilitate many applications in computer vision. However, manually annotating images is a computation-expensive and labor-intensive task. To address these problems, this paper proposes a novel approach by using a cograph regularized collective nonnegative matrix factorization method to annotate images, which is referred to as CG-CNMF; CG-CNMF maximizes the annotation consistency for each image and minimizes the semantic gap for good annotation performance. To reduce the computation cost, this method formulates the annotation problem as a recommending issue and uses nonnegative matrix factorization (NMF) to recover the image-to-label relation for the testing images. Moreover, to find the most similar latent image features and latent label features during the matrix factorization, it exploits the image-to-image relation and label-to-label relation by utilizing the visual content information of images and the semantic cooccurrence information of labels, respectively. To reduce the semantic gap between the image visual content and semantic concepts, both the semantic features and convolutional neural networks (CNNs)-based visual features are considered. Moreover, to address the label-imbalance and incomplete-label problems, the visual-based label cooccurrence information is also considered. In this way, visually similar images are highly correlated with the true semantics of the test images. The experimental results for three multilabel image datasets demonstrate the effectiveness and the efficiency of the proposed method.

INDEX TERMS Image annotation, nonnegative matrix factorization, collective nonnegative matrix factorization, semantic gap, convolutional neural networks.

I. INTRODUCTION

With the rapid development of the internet, digital images have achieved an exponential increase. Manually annotating this huge volume of images is a rather challenging issue. Due to the capability of describing visual images with semantic concepts, automatic image annotation has been an effective and straightforward way to facilitate many applications. Automatic image annotation attracts extensive attention not only in the image retrieval [1]–[3] and image understanding fields [4], [5] but also in other domains, such as biomedical engineering [6]–[8]. Realistically, due to the high costs and time constraints associated with manual annotation, labeling all these images with semantic information by humans is impractical. An efficient and accurate automatic image annotation method is urgently needed.

To resolve this problem, researchers have devoted many efforts to solving the image annotation issue automatically. A large number of methods have been proposed. These approaches assign one or more keywords to describe the visual content of the images, which demonstrates the mapping from visual content to semantic concepts. In the literature, some studies solve this issue by extracting global image features [5], [9], [10], such as global color and texture features. Moreover, some methods are based on multiview features [10], [11]. The more features depicted

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

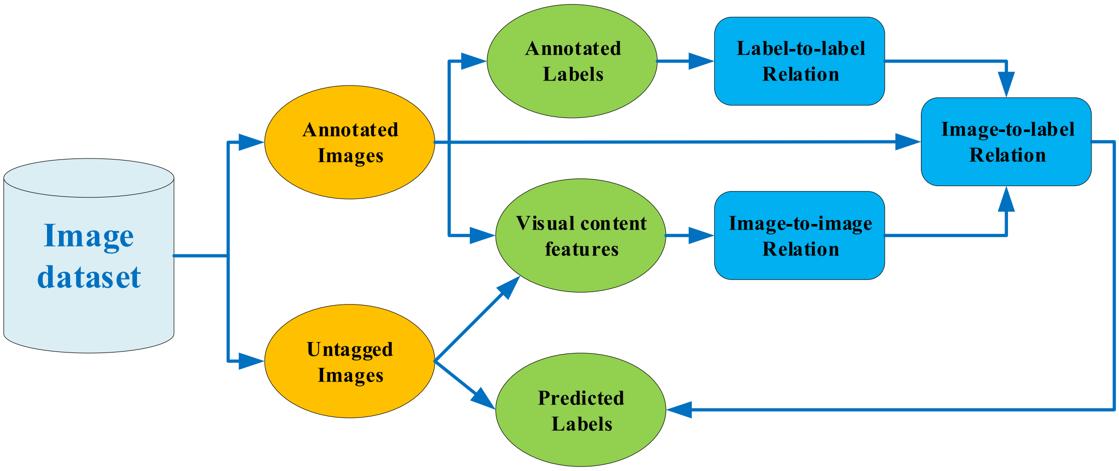

FIGURE 1. The overview of the proposed method.

from images, the more comprehensive is the information that can be attained to improve the performance. However, the fact remains that the more features there are, the more the computation will cost. This is one of the challenges for image annotation.

There is another challenge for image annotation, which is the semantic gap [12] between low-level visual features and high-level semantic features of images. With respect to the former, extracting one type of feature from an image is usually difficult and complex; this approach requires prior expert knowledge to design appropriate handcrafted features manually. With regard to the latter, semantic features cannot fully preserve the visual content of images. However, determining how to reduce the semantic gap between human languages and common visual features for images is also a rather challenging issue but is vital for automatic image annotation task. Some researchers have devoted efforts to improving the visual features extracted from images to narrow down the gap and boost the annotation performance. To this end, the state-of-the-art CNN-based feature learning methods [13]–[15] have achieved the most significant improvement. However, these feature-based approaches do not preserve the semantic features well. A few works [4], [12], [16] have achieved some improvements by exploiting the semantic information from labels to reduce the gap. For example, some of them use label cooccurrence [4], [12]]; however, this type of method does not exploit the fine-grained visual features. To make use of the information from both images and labels, some efforts have aimed to explore the possibility of obtaining more useful information not only from the images but also from labels [10], [17], [18]. For example, some methods attempt to discover correlations between visual contents and semantic concepts [17], [18]. In addition, other techniques have been employed to minimize the gap, such as the supervised dictionary learning used in [19]. The authors proposed a weakly supervised dictionary learning method that uses both the visual features and feature-to-visual word mappings to narrow down the semantic gap. Other methods such as [20] annotate images by coherent semantic concepts learned from visual contents of images. In these methods, [19], [20] consider both the image-to-image and image-to-label relations. To jointly consider the three types of relations, [21]–[23] all employ a loose joint solution for image annotation. [23] solves the image annotation problem by a graph learning method based on both the image-based graph and label-based graph, which conducts the learning of two graphs as two sequential steps of learning and does not utilize three relations simultaneously. Among these approaches, there is a common intuition that similar images share similar labels. Moreover, the similarity of images is always defined using only visual features. However, while visual similarity can deal with correlations among labels to some extent, it fails to handle the two issues of class imbalance (different labels have different frequencies in the dataset) and incomplete labels. To resolve these problems, we think the image similarity should make use of both visual similarity and semantic similarity.

To address the aforementioned problems, we propose a novel image annotation method named cograph regularized collective nonnegative matrix factorization, which simultaneously utilizes the three relations from both images and labels and employs the image graph and label graph to regularize the matrix factorization, thus enhancing the information from the three relations and narrowing down the semantic gap. We refer to this approach as CG-CNMF. In this method, we formulate the image annotation issue as a recommending problem, which uses a collective nonnegative matrix factorization [24] model to combine the three relations of images and labels. To understand the three relations in our method, we show the relations in Figure 1.

In this method, we first construct the image-label matrix, image similarity matrix and label cooccurrence matrix by the three relations. Then, we factorize each matrix into two factors simultaneously. This process is similar to the

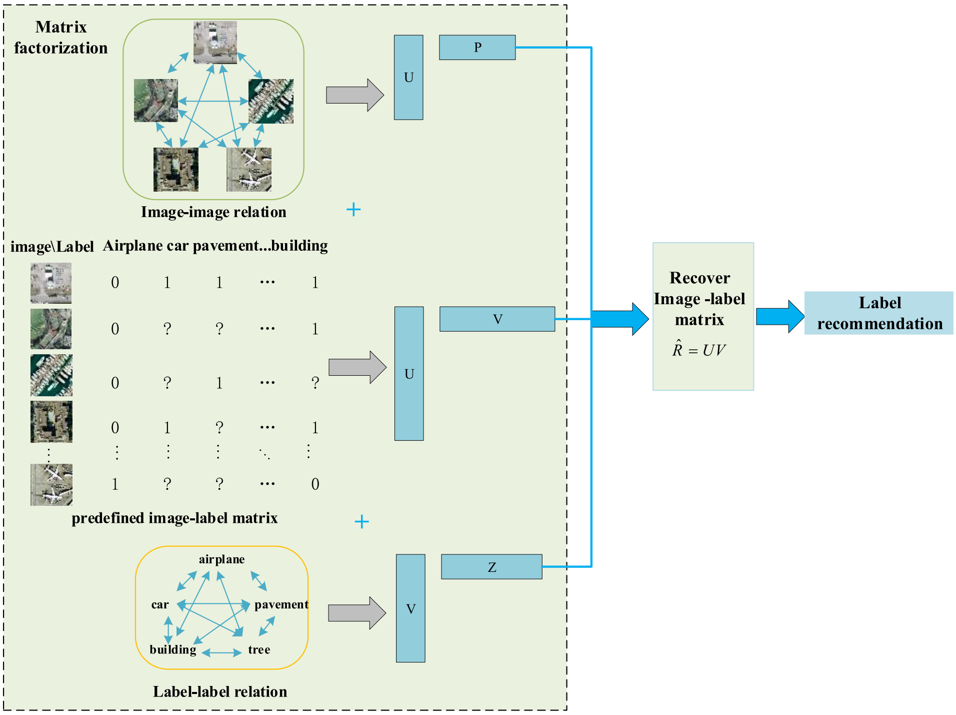

FIGURE 2. Overview of the collective nonnegative matrix factorization for image annotation.

collective matrix factorization. To obtain a comprehensive interpretation of the proposed method, we display the matrix factorization in Figure 2.

During this process, to reduce the computation cost, we try to find a new low-dimensional space that bridges the semantic gap by extracting latent factors from three relation matrices. These latent features can make more meaningful features and affect each other, which can help to force the factorizations to find the most useful latent factors. Finally, we recover the image-label matrix (R) by the product of these matrices and recommend the labels for each testing image. To formulate the image-to-image similarity, we consider two types of similarities: visual-based and semantic-based. For visual-based image similarity, visual feature learning is the key step. Thus, we employ the CNN feature to construct the similarity, which avoids manual feature designing and has recently shown the best performance in computer vision tasks [15], [25], [26]. The visual-based similarity matrix S is factorized into two latent features U and P. Semantic-based similarity is used in the terms of the Laplace matrix [27] that is built by pairwise image similarity according to their labels. Consequently, the two image similarities can affect the factorizations of R and S, and then affect the shared latent feature matrix U in turn. With respect to label-to-label cooccurrence, we take two kinds of information into account, pure semantic-based and visual-based information. Pure semantic-based cooccurrence mainly depends on the frequencies of labels in the datasets. We decompose the related matrix C into two matrices P and Z. The visual-based label cooccurrence is calculated by the frequencies of visual contents that are related to some labels. The visual-based label cooccurrence is used as a Laplace matrix regularization term in the matrix factorization. This term affects the factorizations R and C simultaneously, then affects the shared latent feature matrix V in turn.

To conduct latent factor analysis, we learn the low-rank latent feature spaces by employing the image-label matrix, image-image matrix and label-label matrix. We connect these matrices by the label latent feature space and image latent feature space. That is, the label latent factors in image-label space are connected to the ones in label-label space, and the image latent factors in image-label space are tied to the ones in image-image space. Finally, the learned image latent factors and label latent factors are used to recover the image-label matrix, which can be utilized to recommend the image labels.

Given all of the above, we can summarize the main novelty and technique contributions of this method as follows:

- We formulate the image annotation problem as a label recommendation problem, which can simplify the process of image annotation.
- To make full use of the three relations, we use a collective matrix factorization model to factorize three relation matrices simultaneously.
- To learn a more precise similarity for images, we use a CNN feature learning method to learn the visual features from images offline, which reduces the running time and improves the overall performance of this method.
- To narrow down the semantic gap between images and labels, we build the image graph by semantic

information and the label graph by visual-based cooccurrence information simultaneously.

Experimental results demonstrate that the proposed method achieves promising annotation performance by using three relations on both images and labels. To further improve the performance, this method utilizes both the visual content-based similarity and semantic-based similarity for images, and it also explores the visual-based label similarity and semantic-based label cooccurrence for labels. It makes full use of both images and labels information for image annotation.

The remainder of this paper is organized as follows. In Section II, we review some of the recent works in this domain. Section III describes the proposed method of this paper. Section IV analyzes the optimization process of the method. In Section V, we conduct a set of experiments to evaluate the performance of our method. Finally, we analyze the experimental results and state the conclusion for this paper.

II. RELATED WORK

A. AUTOMATIC IMAGE ANNOTATION

Automatic image annotation plays an important role in computer vision, multimedia and information retrieval domains. Early annotation works usually can be categorized into four types: mixture models, generative models, discriminative models and nearest neighbor-based methods. Mixture models such as [28]–[30] usually define a joint distribution between images and labels then estimate the labels over the cooccurrence of labels and images from training images. Generative models often utilize topic models to represent image-label relationships, such as probabilistic latent semantic analysis (PLSA) [31], [32] or latent Dirichlet allocation (LDA) [33], [34], and nonnegative matrix factorization (NMF) [11], [35], [36]. Discriminative models often pose annotation as a classification problem, such as an SVM [9] and multiple instance learning [37]. These models learn a separate classifier for each label based on low-level visual features. Both generative and discriminative models require clean and large-scale image datasets for training process.

Due to the simplicity and efficiency, nearest neighbor-based approaches [16], [38]–[40] have primarily been the most important and popular method for the image annotation domain. It predicts labels for a test image by calculating the similarity with the training images. However, these methods tend to overfit to local distributions of samples. To address this issue, some techniques have been added, such as metric learning [41], [42] and weighted KNN [35], [43]. Representative examples of these methods are TagProp [41] and 2PKNN [42]. TagProp transfers labels to a test image by wrapping a logistic discriminant model over a weighted KNN method, which resolves the class imbalance problem and boosts the importance of the infrequent labels by suppressing the importance of the frequent labels. This method directly maximizes the log-likelihood of the tag predictions in the training data. 2PKNN is a two-step variant of the KNN method. This method utilizes image-to-tag and image-to-image similarities and learns weights for multiple features.

Recently, the graph-based methods have achieved huge successes in image annotation [3], [17], [18], [44]. These methods usually exploit the image feature distance to establish relevant graphs of samples. They connect both annotated images and unannotated images according to their visual similarities. There is an assumption that neighboring images in the relevant graph have similar labels. Based on this assumption, these methods propagate the labels from labeled images to unlabeled images by considering the visual similarity between nodes. One weakness of these methods is the complexity. The graph-based methods construct a k-NN similarity graph with pairwise relations over images but do not consider the correlations between labels. Another issue is the high computational time required in the testing phase. Usually, these methods need to search the nearest neighbors in the entire dataset with high-dimensional feature vectors for each testing image. Using dimensional reduction techniques can reduce the testing time. To this end, NMF-based approaches [11], [18], [45] are proposed to solve image annotation problems. In [45], single-view features are used. References [11] and [18] extend this method to multiview by simply concatenating multiple feature vectors into one vector before dimension reduction. However, this approach causes the dimension disaster problem.

To address these issues, motivated by their advantages and weaknesses, we do not utilize the multiview features for image annotation. To resolve this problem in a different way, we formulate it as a label recommending problem. Based on this consideration, we factorize the image-label matrix into an image feature matrix and a label feature matrix. Simultaneously, an image visual-based similarity matrix and label cooccurrence matrix are factorized into two matrices. The latter two factorizations share the latent image feature matrix and label feature matrix with the first factorization.

We refer to the proposed method as the cograph nonnegative matrix factorization method. In this method, to reduce the feature dimension and obtain high efficiency in computation, we use the NMF-based method. To make full use of the image and label information, we combined the cograph regularization terms in this method and consider both the visual-based and semantic-based information for both images and labels. To further reduce the semantic gap, we use the CNN features to build the visual-based image similarity matrix. Furthermore, we employ the three relations simultaneously in the image annotation, allowing the relations to affect each other in the matrix factorizations.

B. CNN-BASED FEATURE LEARNING

CNNs have the advantages of low complexity, by sharing weights, and high performance in vision tasks when compared with traditional handcrafted feature learning methods; further, CNN-based feature learning methods have been shown to be the most powerful feature learning approaches in computer vision [13]–[15] and have been applied in

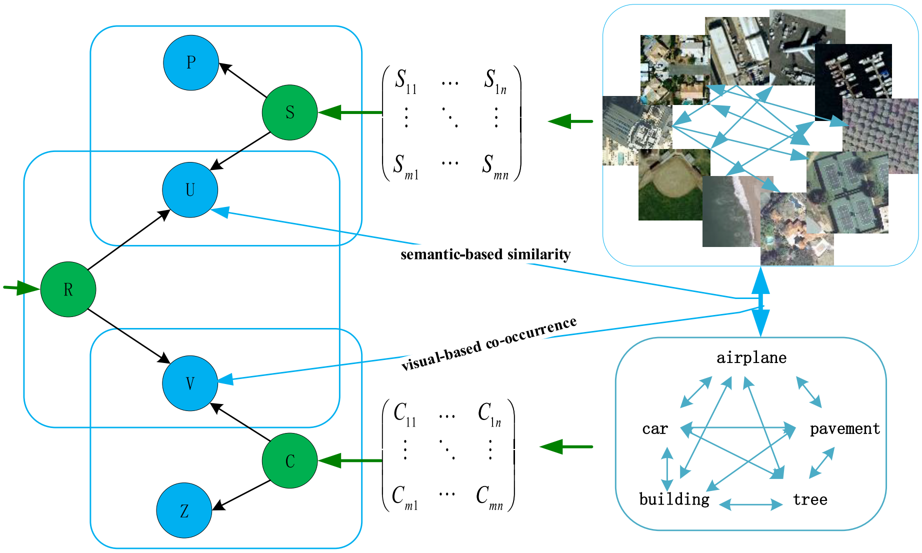

FIGURE 3. **Overview of the proposed method for image annotation.**

many tasks, such as image classification [13], object recognition [14], image parsing [46] and image retrieval [25]. There are many CNN models that have achieved good performances in image feature extraction. Alexnet [47], a 5-layer network, was proposed for image classification and won the ILSVRC-2012 competition. In this work, significant improvement in large-scale image classification on ImageNet [48] was achieved by using CNN. In [14], a deeper architecture VGG-net was proposed and achieved better performance in image classification accuracy. Other CNN models, such as GoogLeNet [49] and ResNet [50], have refreshed the accuracy record of recognition. It is known that deeper CNN can achieve better performance in extracting features. However, deeper networks have much more computational cost. Moreover, in [26], VGG-net and ResNet were shown to perform better than several other famous models. Moreover, VGG-net has simpler architecture compared with ResNet; as such, we choose VGG-net as the feature extraction method to build the visual-based image similarity matrix.

III. COGRAPH REGULARIZED COLLECTIVE NONNEGATIVE MATRIX FACTORIZATION

In this paper, we focus on the annotation problem in which an untagged image can be assigned multiple labels. Let $X = \{x_1, x_2, \ldots, x_m\}$ denote the image set, which has m images in it. For each image, CNN features have been extracted.

A. PROBLEM FORMULATION

For the multilabel image annotation task, we suppose there are m images and n semantic labels $L = \{l_1, \ldots l_n\}$. The aim of this paper is to annotate the unlabeled images efficiently. To achieve this goal, we exploit three relationships: image-to-image, label-to-label and image-to-label. To utilize the three relations, we construct three matrices as image-image similarity matrix S, label-label cooccurrence matrix C and image-label matrix R. As the image-to-label matrix R is incomplete with many missing entries, our objective goal is to fill these missing entries to obtain a label set for each test image. We use the nonnegative matrix factorization model to factorize these three matrices to find the inherent relationships of images and labels. We show the overall view of the proposed method in Figure 3.

Figure 3 shows the main idea of the proposed method based on collective matrix factorization. Given the image-label matrix $R \in \mathbb{R}^{m \times n}$, we decompose it into two low-rank matrices $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$, where $k \leq n$. The latent image information is shared through sharing matrix U with the image-image similarity matrix S and Laplace matrix of the image semantic graph. Simultaneously, the image-label matrix shares the label information through sharing matrix V with the label-label cooccurrence matrix C and the Laplace matrix of the label visual-based graph.

B. MODELING THE IMAGE-TO-IMAGE RELATION

For humans, we consider the similarity between two images usually according to two reasons. One is their low-level features, such as color, texture and so on. The other is the annotated labels of them, which denotes the high-level semantic features of images. Therefore, we design two types of image similarities: visual-content similarity and semantic similarity. To calculate the two similarities for images, we construct two image similarity matrices according to different goals.

1) VISUAL-CONTENT IMAGE SIMILARITY

While calculating the visual-content-based images similarity, feature extraction plays an important role in this process. The deep convolutional neural network as an end-to-end feature learning method leads this trend. To achieve the optimal performance, we exploit the CNN as the visual-based feature learning method.

To avoid designing a new network, we employ the 16-layer deep CNN architecture VGG-net [14] and recommend readers to refer to the details from the original paper. We utilize this architecture to extract a 4096-dimensional feature vector for each image. To be compatible with VGG-net, we resize each image to 224 × 224, and extract visual features through 8 convolutional layers and three fully connected layers. The activations of the last fully connected layer are the visual features. To reduce the running time, we extract CNN features offline.

Let sx_i and sx_j indicate the feature vector of the i^{th} image and j^{th} image respectively. We define the pairwise similarity based on CNN features as follows:

$$S^{VS}_{x_i,x_j} = \frac{< sx_i, sx_j >}{\|sx_i\| \, \|sy_i\|} \tag{1}$$

where $< sx_i, sx_j >$ calculates the inner product of the two feature vectors. According to Eq. (1), we can construct the similarity matrix S for pairwise images without considering the semantic meaning of each image.

2) SEMANTIC IMAGE SIMILARITY

As discussed above, the visual-content-based similarity does not take the multilabel information into account. In addition, this approach cannot employ the label information of training images. To make full use of the label information collected from training data, we use the label cooccurrence to build a semantic similarity graph G^U.

In the graph, nodes represent label sets of images in the training dataset, and edges represent the affinity between the label sets. The affinity matrix $W^U \in \mathbb{R}^{m\times m}$ of the graph is defined as

$$W^U_{i,j} = \begin{cases} sim(l(x_i), l(x_j)) & \textit{if } x_i \textit{ and } x_j \textit{ share some classes}, \\ 0, & \textit{otherwise}. \end{cases} \tag{2}$$

where $sim(l(x_i), l(x_j))$ denotes the similarity of the pairwise label vectors, which is calculated as Eq. (3),

$$sim(l(x_i), l(x_j)) = \frac{< l(x_i), l(x_j) >}{\|l(x_i)\| \, \|l(x_j)\|}, \tag{3}$$

where $l(x_i)$ and $l(x_j)$ denote the semantic label vector of image x_i and x_j, respectively. $< l(x_i), l(x_j) >$ computes the inner product of image-label vectors. Since $W^U_{i,j}$ in our paper is only measuring the closeness of the two label sets for the two images, we only use the simple semantic similarity. Preserving the geometric structure in the image space is reduced to minimizing the following loss function:

$$O_1 = \frac{1}{2}\sum_{i,j=1}^{m} \|u_i - u_j\|^2 W^U_{ij} = Tr(U^T L_U U) \tag{4}$$

where $D^U \in \mathbb{R}^{m\times m}$ is a diagonal matrix whose entries are column sums of W^U, $D^U_{ii} = \sum_{j=1}^{m} W^U_{ij}$, and $L_U = D^U - W^U$ is the Laplacian matrix of the graph G^U.

C. MODELING THE LABEL-TO-LABEL RELATION

In the multilabel image dataset, one label is usually assigned to many images if the visual contents of these images are related to this label. Scanned over the whole dataset, we can find that the relationships among these labels are not independent of the visual contents of the images. Therefore, in this paper, we consider the label-to-label relation from two perspectives. First, we consider the semantic-based label relation, which we named as label cooccurrence. For example, if two labels are always assigned to the same image in the dataset, we can calculate the cooccurrence percentage of the two labels by counting the label pairs annotated jointly in the whole dataset. Second, if two labels are always shared by the same image, then these images always have some similar visual-based characteristics. In the following, we will design the two cooccurrences.

1) VISUAL-BASED LABEL COOCCURRENCE

Since the visual content is the direct representation of one image, it should contribute to the label cooccurrence. We consider the visual-based similarity for two labels in the following measures: if two labels always occur as candidates for the same image x_i and never with any other labels, then they are considered as highly visually similar, and we use $vsim(l_a, l_b) = 1$ to denote the similarity. If labels l_a and l_b never occur together, we consider they are not visually similar, and $vsim(l_a, l_b) = 0$. These two cases are the special cases. In other cases, if two labels occur together with other labels for the same image, we define the visual-based label cooccurrence as follows:

$$\begin{aligned} vsim(l_a, l_b) &= \frac{1}{T_P} K_S(I(l_a), I(l_b)) \\ &= \frac{1}{T_P} \sum\nolimits_{i=1,j=1}^{T_P} S_I(I_i(l_a), I_j(l_b)) \end{aligned} \tag{5}$$

where $I(l_a)$ and $I(l_b)$ indicate the image sets related by label l_a and label l_b, respectively. $I_i(l_a)$ is the i^{th} image in image set $I(l_a)$. $K_S(.)$ denotes the similarity of the two image sets, and S_I means the similarity function between the two images. T_P is the number of the most similar images from the labeled images. We set $T_p = 10$ in our paper. That means we choose 10 images for calculating the visual-based similarity of labels and then obtain the average value as the final similarity.

Similar to the semantic similarity graph G^U, we construct the visual-based label graph G^V. We use $W^V \in \mathbb{R}^{n\times n}$ to indicate the affinity matrix and define W^V as

$$W^V_{ij} = \begin{cases} vsim(l_a, l_b); & l_a \textit{ and } l_b \textit{ occur together}, \\ 0 & \textit{otherwise} \end{cases}$$

Then, we get the following loss function:

$$O_2 = \frac{1}{2}\sum_{i,j=1}^{n} \left\| v_i - v_j \right\|^2 W_{ij}^v = Tr(V^T L_V V)$$

where $L_V = D^V - W^V$ is the Laplacian matrix, D^V is a diagonal matrix and $D_{ii}^V = \sum_{j=1}^n W_{ij}^V$.

2) SEMANTIC-BASED LABEL COOCCURRENCE

Generally, if there are two labels with high cooccurrence in the training dataset, then there will be a high probability to annotate other images simultaneously. In this paper, to calculate the cooccurrence percentage of the two labels, we construct an image-label matrix T for the training data, where the rows denote the images and the columns denote the labels. If image x_i is assigned the label l_j, then $t_{ij} = 1$ and $t_{ij} = 0$ otherwise. We use $t_{:i}$ and $t_{:j}$ to denote the i^{th} and j^{th} column of matrix T, respectively. Then, we define the label cooccurrence of the two labels as follows:

$$sim(l_i, l_j) = \frac{< t_{:i}, t_{:j} >}{\|t_{:i}\| \, \|t_{:j}\|} \tag{6}$$

where $< t_{:i}, t_{:j} >$ calculates the inner product of the two label vectors in the training data. According to Eq. (6), we can construct the label cooccurrence matrix C.

D. OBJECTIVE FUNCTION OF CG-CNMF

This multilabel annotation problem can be considered as an optimization problem. We solve the problem by the following objective function:

$$\begin{aligned}
&\underset{s.t. U\geq 0, V\geq 0, P\geq 0, Z\geq 0,}{L_{JWNMF}(U, V, P, Z)} \\
&= \frac{1}{2}\left\|Y \odot (R - UV^T)\right\|_F^2 + \frac{\alpha}{2}\left\|S - UP^T\right\|_F^2 \\
&\quad + \frac{\beta}{2}\left\|C - VZ^T\right\|_F^2 \\
&\quad + \frac{\lambda_U}{2} Tr(U^T L_U U) + \frac{\lambda_V}{2} Tr(V^T L_V V) \\
&\quad + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2 + \|P\|_F^2 + \|Z\|_F^2)
\end{aligned} \tag{7}$$

where Y is the indicator matrix of the missing "rating", of which missing values are addressed by binary weights Y_{ij}

$$Y_{ij} = \begin{cases} 1, & if\ R_{ij}\ is\ observed; \\ 0, & if\ R_{ij}\ is\ unobserved. \end{cases}$$

where $\|.\|$ denotes the Frobenius norm, $\odot$ is the Hadamard product operator. $\alpha, \beta, \lambda_U, \lambda_V$ are the regularization parameters that balance the reconstruction error of CG-CNMF in the first three terms and the rest of the terms. Moreover, the last four terms with λ help the objective function avoid overfitting.

As shown in Figure 3 and the objective function in Eq. (7), we aim to propagate the information among the image-label matrix R, image-to-image matrix S and label-to-label matrix C by sharing some low-rank matrices U and V. The first three terms in Eq. (7) control the loss in matrix factorization, the fourth and fifth term control the information from semantic-based image similarity and visual-based label cooccurrence to help the first term find more interpretable representations, and the last 4 terms help the regularization over the factorization matrices to prevent overfitting.

IV. OPTIMIZATION PROCESS AND IMAGE ANNOTATION VIA CG-CNMF

In this section, we will investigate the solution for Eq. (1). In general, the objective function in Eq. (7) is not jointly convex to all the variables, and we cannot obtain a closed-form solution by minimizing this equation with respect to U, V, P and Z. Therefore, we will optimize the objective function by an alternating scheme, in which optimizing one variable can be achieved by fixing the others and repeating this procedure until convergence.

A. THE OPTIMIZATION PROCESS OF CG-CNMF

In this subsection, we iteratively solve one variable while fixing all others.

1) UPDATE FOR U

If we fix V, P and Z at the current iteration step, the objective function in Eq. (7) with respect to U can be written as

$$\begin{aligned}
L(U) &= \frac{1}{2}\left\|Y \odot (R - UV^T)\right\|_F^2 + \frac{\alpha}{2}\left\|S - UP^T\right\|_F^2 \\
&\quad + \frac{\lambda_U}{2} Tr(U^T L_U U) + \frac{\lambda}{2}\|U\|_F^2 \\
&\quad s.t.\ U \geq 0
\end{aligned}$$

By taking the first derivative of $L(U)$, we have

$$\begin{aligned}
\frac{\partial L(U)}{\partial U} &= -Y \odot RV + Y \odot (UV^T)V - \alpha SP \\
&\quad + \alpha UP^T P + \lambda_U L_U U + \lambda U
\end{aligned}$$

Considering L_U may take any signs, following [51], we introduce $L_U = L_U^+ + L_U^-$, and $M_{ij}^+ = (|M_{ij}| + M_{ij})/2$, $M_{ij}^- = (|M_{ij}| - M_{ij})/2$. Then, we set the above partial derivation to zero by using the Karush-Kuhn-Tucker (KKT) complementary condition [52] and attain the following multiplicative updating rule:

$$U_{ij} \leftarrow U_{ij}\sqrt{\frac{[Y \odot RV + \alpha SP + \lambda_U L_U^- U]_{ij}}{[Y \odot (UV^T)V + \alpha UP^T P + \lambda_U L_U^+ U + \lambda U]_{ij}}} \tag{8}$$

2) UPDATE FOR V

Considering V only, we need to solve the following problem:

$$\begin{aligned}
L(V) &= \frac{1}{2}\left\|Y \odot (R - UV^T)\right\|_F^2 + \frac{\beta}{2}\left\|C - VZ^T\right\|_F^2 \\
&\quad + \frac{\lambda_V}{2} Tr(V^T L_V V) + \frac{\lambda}{2}\|V\|_F^2 \\
&\quad s.t.\ V \geq 0
\end{aligned}$$

Considering the symmetry of U and V in Eq. (7), the solution of the updating rule with respect to V is analogous to that of U. We can obtain the following updating rule

$$V_{ij} \leftarrow V_{ij}\sqrt{\frac{[(Y \odot R)^T U + \beta C^T V + \lambda_V L_V^- V]_{ij}}{[Y \odot (UV^T)^T U + \beta V Z^T Z + \lambda_V L_V^+ V + \lambda V]_{ij}}} \tag{9}$$

3) UPDATE FOR P AND Z

Similarly, we fix other variables for P and Z, and we can obtain the following equations

$$L(P) = \frac{1}{2}\left\|S - UP^T\right\|_F^2 + \frac{\lambda}{2}\|P\|_F^2$$
$$L(Z) = \frac{1}{2}\left\|C - VZ^T\right\|_F^2 + \frac{\lambda}{2}\|Z\|_F^2$$

Then, we calculate the first derivatives of P and Z,

$$\frac{\partial L(P)}{\partial P} = -S^T U + PU^T U + \lambda P$$
$$\frac{\partial L(Z)}{\partial Z} = -C^T V + ZV^T V + \lambda Z$$

By setting the first derivatives to zero, we will obtain the following updating rules for P and Z

$$P_{ij} \leftarrow P_{ij}\sqrt{\frac{\left[S^T U\right]_{ij}}{\left[PU^T U + \lambda P\right]_{ij}}} \tag{10}$$

$$Z_{ij} \leftarrow Z_{ij}\sqrt{\frac{\left[C^T V\right]_{j}}{\left[ZV^T V + \lambda Z\right]_{jj}}} \tag{11}$$

These updating rules are derived one-by-one by fixing the other three variables. They are analogous to those of NMF [53]. The difference is how to update the image feature factors U and the label feature factor V. In Eq. (8), the update of U mainly depends on two sources of data: the image-label matrix R and the image-to-image similarity matrix S. Similarly, the update of V depends on two sources of data: the image-label matrix R and the label cooccurrence matrix C.

The successive iterations will lead the objective function to converge. After convergence, we can easily recover the image-label matrix by the learned matrices and recommend the labels for the unlabeled images.

B. THE CONVERGENCE OF CG-CNMF

The objective function in Eq. (7) is not a strict convex function with respect to U, V, P and Z together; however, it is a convex function with respect to U, V, P and Z separately. Therefore, we can solve the optimization problem by the alternative multiplicative updating rules. To prove the convergence of Eq. (7), we can obtain the following theorem.

Theorem 1: The objective function in Eq. (7) is nonincreasing under the updating rules of Eq. (8)-(11); hence, it converges to a local minimum.

For a better flow of this paper, we provide the proof of Theorem 1 in the Appendix. This theorem guarantees the objective function in Eq. (7) always decreases and hence converges.

C. IMAGE ANNOTATION VIA CG-CNMF

After learning all the feature matrices U, V, P and Z, we perform the annotation by reconstructing the image-label matrix R. We use $\hat{R}$ to denote the recovered matrix, which is the product of the image feature matrix U and label feature matrix V. Each image is labeled with the top 5 labels by matrix $\hat{R}$. The process is summarized as follows:

Algorithm 1 Image Annotation via CG-CNMF

```
Input: image-label matrix $R$ with labeled and unlabeled images, image-to-image semantic similarity matrix $W^U$ and visual-based similarity matrix $S$, label-to-label semantic cooccurrence matrix $C$ and visual-based cooccurrence matrix $W^V$, loss error $\varepsilon$, regularization parameters $\alpha, \beta, \lambda_U, \lambda_V, \lambda > 0$, number of images m, number of total labels n, and number of latent features k;
Output: $U \geq 0, V \geq 0, Z \geq 0, P \geq 0, \hat{R} \geq 0$
Initialize: $U_0 \geq 0, V_0 \geq 0, P_0 \geq 0, Z_0 \geq 0$
1: Construct weight matrix $Y$ according $R$, $Y_{ij} = 1$ if $R_{ij}$ can be observed; otherwise, $Y_{ij} = 0$;
2: Construct image visual-based similarity matrix S;
3: Construct image semantic-based similarity matrix $W^U$;
4: Construct label semantic cooccurrence matrix C;
5: Construct label visual-based cooccurrence matrix $W^V$;
6: while the loss error of Eq. (7) $> \varepsilon$ do
7:    $t := t + 1$;
8:    update $U^{t+1}$ according to Eq. (8);
9:    update $V^{t+1}$ according to Eq. (9);
10:   update $P^{t+1}$ according to Eq. (10);
11:   update $Z^{t+1}$ according to Eq. (11);
12: end while
13: Take $\hat{R}$ as the approximation of $R$;
14: Return a tag recommendation list of top 5 tags with the largest 5 values in the recovered matrix $\hat{R}$ for each test image.
```

The image annotation process via CG-CNMF is summarized in Algorithm 1. In steps 8–11, the algorithm updates U, V, P and Z iteratively until convergence. The optimal solution of the objective function in Eq. (7) can be obtained simultaneously. After the optimization process, we can obtain the approximate image-label matrix $\hat{R}$ from the learned feature matrices U and V according to $\hat{R} = UV$.

Then, we take the top *5* entries in a row of the image-label matrix $\hat{R}$ as the recommended labels for an image.

D. TIME COMPLEXITY OF CG-CNMF

In this subsection, we discuss the time complexity of the proposed method. We use big O to express the complexity. The time complexity of Algorithm 1 dominates two parts: matrix factorization and reconstruction. In the first part, the main cost is the multiplicative updating rules and the constructions

TABLE 1. Statistics of datasets.

Name	Instances	Labels	TPI	Train	Test
Corel5k	5,000	260	3.522	4,500	500
IAPR TC12	19,627	291	5.723	17,665	1,962
ESP	20,700	268	4.762	18,689	2,081

of the following matrices: image-image visual-based and sematic-based similarity matrices, label-label visual-based matrix and semantic-based cooccurrence matrix. We suppose the multiplicative updates stop after t_{in} iterations. The time cost of the multiplicative updates is $O(t_{in}F_k(mkn) + t_{in}mk + t_{in}kn)$, where F_k denotes the number of the observed entries in the image-label matrix R. Because k and n are much smaller than m, the time complexity of updates is approximate to $O(t_{in}F_k m)$. The construction of the image semantic-based and label visual-based cooccurrence graphs spend $O(2m^2 + 2n^2)$. In the recovering step, the time complexity is $O(mkn)$, which is approximate to $O(m)$. Therefore, the overall time complexity of Algorithm 1 is approximate to $O(t_{in}F_k m + 2m^2 + +2\mathrm{n}^2 + m)$.

V. EXPERIMENTS AND EVALUATIONS

In this section, we will investigate the effectiveness of the proposed method by comparing it with other multilabel approaches. Furthermore, we will analyze the results and show the influence of related parameters used in this paper.

A. DATASETS AND PREPROCESSING

To evaluate the performance of the proposed method and make it easy to compare with other annotating methods, we choose three popular and publicly available multilabel datasets: Corel5K [54], IAPR TC12 [55] and ESP-GAME [56]. Corel5k is the standard multilabel dataset and has been the most common dataset employed for tag-based image annotation. It has 4,500 training sets and 500 testing sets. The tag per image (TPI) is 3.4. IAPR TC12 has 19,627 images covering several scenes such as landscape shots, animals, and city pictures, and the TPI is 5.7. The last dataset is constructed from an online game. It consists of 18, 689 training images and 2,081 testing images, and the TPI is 4.7. We summarize the statistics of these datasets in Table 1.

B. EVALUATION METHODOLOGY

For performance evaluation, we adopt the widely used performance metrics, mean precision (P%), mean recall (R%), F1 score and N plus (N^+). The precision measures the percentage of images correctly annotated in the total images. The recall rate refers to images that are correctly annotated relevant to the ground-truth annotations. It is a commonly used metric in the image annotation field. The F1 score is the harmonic mean of precision and recall. N plus reports the number of tags with nonzero recall. Similar to other works, we first automatically annotate each image with 5 tags and then compute precision and recall for each tag. After that, we calculate the F1 score and N plus measures. The precision, recall and F1 score are defined as follows:

$$precision(li) = \frac{N_{correct}}{N_{labeled}}, \quad recall(li) = \frac{N_{correct}}{N_{all}}$$
$$F_1 - score(l_i) = 2\frac{\Pr ecision(l_i) \times \operatorname{Re} call(l_i)}{\Pr ecision(l_i) + \operatorname{Re} call(l_i)}$$

where $N_{correct}$ denotes the number of images that are correctly annotated, $N_{labeled}$ is the number of correct images relevant to the ground-truth annotations and N_{all} is the total number of images to be automatically annotated. To reduce the errors caused by inappropriate sampling, the experiments were cross-validated on 10 sets of randomly chosen samples.

C. COMPARISON WITH OTHER APPROACHES

To evaluate the annotation performance of CG-CNMF, we compare it against several other annotation approaches. The compared methods are summarized as follows:

- TagProp [41]: This is a KNN-based method that uses the tag propagation to learn a weighted nearest-neighbor model. It integrates the metric learning by directly maximizing the log-likelihood of the tag predictions in the training set.
- NMF-KNN [35]: NMF-KNN represents a query-specific generative model, which learns the features of nearest-neighbors and tags using a weighted extension of the multiview nonnegative matrix factorization method.
- 2PKNN [42]: 2PKNN is a two-phase method, in which the first pass is to address the class imbalance by constructing a balanced neighborhood for each test image, and the second pass is to assign the actual tag importance based on image similarity. This method uses "image-to-label" similarities in the first step, while it uses "image-to-image" similarities in the second step, thus combining the benefits of both. Our method simultaneously utilizes three relations including these two similarities, which is very helpful to the performance of image annotation.
- FastTag [57]: FastTag recasts a supervised multilabel classification problem as unlabeled multiview learning. It jointly learns two classifiers for images and text. To trade off complexity in the classifiers, it utilizes a nonlinear mapping for features, which can efficiently deal with sparsely tagged training data and rare tags.
- CCA-KNN [58]: CCA-KNN is a canonical correlation analysis (CCA) framework with k-nearest neighbor (CCA-KNN) clustering for image annotation. This method makes use of convolutional neural network (CNN) features and word embedding vectors to represent the associated tags of images. It extracts CNN features for images using a pretrained VGG-16 [59] network. Meanwhile, the word embedding vectors are extracted using a pretrained ship-gram architecture word2vec. Both networks are publicly available.

- MLDL [60]: MLDL describes a multilabel learning method by using label consistency regularization and a partial-identical label embedding method for image annotation. It incorporates the dictionary learning technique into multilabel learning in the input feature space. Moreover, in the output label space, it uses the label embedding to cluster the samples with the same label set and collaboratively represents the label set for the partial-identical samples.
- JEC [61]: This method treats image annotation as a retrieval issue. It uses a greedy algorithm to transfer a label from neighbors by using multiple global features. In Joint Equal Contribution (JEC), each feature contributes equally toward the image distance. It scales the distances for each feature such that they are bounded by 0 and 1.
- RMLF [62]: RMLF is a method of late fusion for image annotation based on rank minimization. It obtains an optimal matrix by solving a minimization optimization problem and gives the final prediction of tags with this matrix.

These eight state-of-the-art algorithms are employed as benchmark baselines. Considering the three matrices used in the proposed method, we first construct the image-label matrix, image-image visual-based similarity and label-label semantic-based cooccurrence matrix for each dataset. The image-label matrix shows the relationship between images and labels. We use the rows as the different images, and the columns as the different labels. The image-image visual-based similarity matrix demonstrates the interrelationships among images. We construct this matrix by CNN features offline. With respect to the label-label semantic-based cooccurrence matrix, we use the label frequency for each dataset to construct the matrices.

To evaluate the effectiveness of the proposed method, we construct a set of experiments with the three multilabel datasets to compare with the 8 state-of-the-art algorithms. The parameters of these compared methods are set according to their papers or their codes. For fairness, we perform parameter tuning in advance for the proposed method and use the best setting to compare with other methods. Table 2 lists the parameters used in the experiments.

All these approaches are executed on a desktop computer with an Intel Core7 2.4 G CPU and 16 GB memory.

TABLE 2. Parameters used for experiments.

Parameters Notation	Description	Parameter Setting
α	Weight of image-image visual-based similarity information	200
β	Weight of label-label semantic cooccurrence information	100
λ_U	Weight of image-image semantic-based similarity information	100
λ_V	Weight of label-label visual-based cooccurrence information	90
K	Dimension of latent features	40
λ	Regularization Parameter	10

D. EXPERIMENTAL RESULTS AND DISCUSSION

In this subsection, we report the image annotation performance of the proposed method by comparing it with the existing image annotation approaches for three datasets. To evaluate the robustness of the method, we conduct the experiments on different ratios of labeled images. We randomly select 20%, 50%, and 80% of the ratings as the training data, and the rest of the data is used as the test data to evaluate the performance of these methods. Table 3, 4 and 5 exhibit the precision (P), recall (R), F1 score (F1) and N^+ for the three datasets. Because F1 is the harmonic mean of recall and precision, it is more reliable than the analysis of precision or recall performed separately. Thus, we just analyze the F1 score for these methods.

First, we compare the performance among TagProp, NMF-KNN, 2PKNN, CCA-KNN and the proposed CG-CNMF on the Corel5K dataset, of which the training dataset consists of 20%. In these methods, the former four approaches are all KNN-based methods and show promising results. Among these methods, the proposed method completely and significantly outperforms the other methods. It attains 3.4% achievement under F1 score for Corel5k, 1.6% achievement for IAPR TC12 and 0.7% for ESP when compared with CCA-KNN. We believe the reasons are the CNN features that we used to construct the visual-based image similarity and the three relations we employed. These techniques help to reduce the semantic gap and enhance the related information. Moreover, CCA-KNN achieves the best results compared with the other three KNN-based methods. This method utilizes visual features extracted by a convolutional neural network (CNN) from images along with word embedding vectors for semantic concepts. It incorporates both CNN features and text features. The significant performance proves the efficiency of CNN features used in this method. NMF-KNN is the second best KNN-based method, which is better than the traditional weighted nearest neighbor-based approaches such as TagProp. TagProp addresses the class imbalance problem by wrapping a logistic discriminant model over the weighted KNN method. This improves the performance of the image annotation by boosting the importance of infrequent labels and suppressing frequent labels among neighbors. A two-step k-nearest neighbor method 2PKNN works slightly worse than NMF-KNN and a slightly better than TagProp on Corel5k. This method does not need to choose the parameter of the neighborhood dimension, but it implicitly defines this by exploring the most similar images per label. This advantage is obvious in the ESP dataset, which has more labels per image, but it fails to achieve promising results for the IAPR TC12 dataset. We believe that this is due to the TPI of this dataset, which is higher than that of Corel5k. The recall is hard to improve on this dataset. Among these KNN-based methods, FastTag performs the worst. We think the reason is that it focuses more on the speed of

TABLE 3. Experimental results for the three datasets with 20% training data.

Method	Corel5k				IAPR TC12				ESP			
	R	P	F1	N^+	R	P	F1	N^+	R	P	F1	N^+
TagProp	36	28	31.5	136	29	41	33.9	181	21	33	25.6	176
NMF-KNN	42	31	35.6	149	31	42	35.6	192	22	29	25.0	182
2PKNN	40	30	34.2	142	28	43	33.9	196	25	35	29.1	193
FastTag	36	27	30.8	141	27	38	31.5	187	23	39	28.9	209
CCA-KNN	39	34	36.3	169	31	**44**	36.3	206	**31**	40	34.9	211
MLDL	41	**38**	39.4	168	30	42	35.0	202	25	**47**	32.6	195
JEC	23	22	22.4	123	24	27	25.4	165	17	22	19.2	165
RMLF	34	26	29.4	127	27	31	28.8	189	20	25	22.2	173
Proposed	**43**	37	**39.7**	**172**	**34**	43	**37.9**	**211**	**31**	42	**35.6**	**213**

TABLE 4. Experimental results for the three datasets with 50% training data.

Method	Corel5k				IAPR TC12				ESP			
	R	P	F1	N^+	R	P	F1	N^+	R	P	F1	N^+
TagProp	42	35	38.2	160	31	45	36.7	229	26	38	30.8	193
NMF-KNN	47	40	43.2	156	33	44	37.1	233	28	34	30.7	195
2PKNN	47	36	40.8	168	30	47	36.6	228	31	42	35.6	203
FastTag	45	32	37.4	157	31	43	36.0	216	28	43	33.9	212
CCA-KNN	46	41	43.3	178	37	**48**	41.7	**258**	35	46	39.7	233
MLDL	48	**46**	**46.9**	**188**	34	47	39.4	237	29	**51**	36.9	226
JEC	29	27	27.9	137	27	29	27.9	206	22	27	24.2	191
RMLF	47	34	39.4	156	29	35	31.7	198	26	31	28.2	207
Proposed	**51**	43	46.6	185	**39**	**48**	**43.0**	249	**36**	46	**40.3**	**239**

TABLE 5. Experimental results for the three datasets with 80% training data.

Method	Corel5k				IAPR TC12				ESP			
	R	P	F1	N^+	R	P	F1	N^+	R	P	F1	N^+
TagProp	44	37	40.1	168	34	47	39.4	253	29	41	33.9	215
NMF-KNN	49	44	46.3	161	36	46	40.4	256	30	36	32.7	221
2PKNN	**52**	39	44.5	189	35	49	40.8	245	34	46	39.1	234
FastTag	47	35	40.1	163	33	46	38.4	235	33	45	38.0	228
CCA-KNN	50	43	46.2	**206**	39	**52**	44.5	**267**	37	48	41.7	255
MLDL	51	**46**	48.3	196	38	49	42.8	252	32	**53**	39.9	243
JEC	34	29	31.3	144	30	31	30.5	232	27	31	28.8	216
RMLF	49	38	42.8	169	31	37	33.7	249	29	36	32.1	223
Proposed	**52**	46	**48.8**	201	**41**	51	**45.4**	263	**38**	49	**42.8**	**257**

the tagging rather than the accuracy. Fortunately, CG-CNMF achieves a 3.4% gain compared with CCA-KNN in the F1 score, and CCA-KNN does better than other KNN-based methods. This improved performance is due to the advantage of employing all the word2vec vectors as text features used in CCA-KNN, which has been proven to be better than a binary vector of labels [58]. However, our method makes full use of three relationships among images and labels, which not only considers the visual-based and semantic-based image similarities but also utilizes the visual-based label cooccurrence and semantic-based label cooccurrence. Furthermore, we combined CNN features to calculate the visual-based image similarity, and the results have proven that CNN features are better than handcrafted features. All these techniques help us to achieve a good performance in annotation.

Second, we compare CG-CNMF with FastTag, MLDL, JEC and RMLF. Among these methods, JEC performs the worst. Even more, it is also the worst among all the methods. To our knowledge, JEC depends heavily on the features of images and weights these features equally. It could not perform any better than using equal weights. This is due to the limitations of the classification-based metric learning that they used for annotation. It is interesting that FastTag almost aligns with TagProp, which is similar in [57]. Moreover, this method achieves 8.4% gains compared with JEC in F1 on Corel5k. This improvement occurs because of using two co-regularized linear mappings in a joint convex function. RMLF achieves a large margin (7%) in F1 compared with JEC on Corel5k. Even more, it performs slightly better than TagProp and FastTag, which is likely because of the rank minimization-based late fusion method that provides more useful information for annotation. Due to the use of the label consistency regularization and partial-identical label embedding method, MLDL achieves the second-best performance in the F1 score measure, which is slightly worse (0.3%) than the proposed approach. This proves that label consistency and label embedding are helpful to the image annotation in MLDL. Even more, in the output space, the MLDL method employs the label embedding and collaboratively predicting of the labels for the partial-identical samples, which further improves the performance of image annotation. CG-CNMF performs slightly better than MLDL on the Corel5K dataset when the training set is set at 20%. Moreover, it is also better than CCA-KNN on the ESP dataset and IAPR TC12 datasets. We determine there are two reasons for this difference in performance. On the one hand, CNN feature-based image similarity is more accurate than handcrafted features-based similarity. On the other hand, the use of multiple relationships provides much useful information for the matrix factorization and affects the annotation performance in turn. Moreover, the proposed method is much more efficient in computation, which is due to the low-rank representation solution of NMF.

Third, we deploy the experiments on the 50% and 80% training datasets. It is not surprising that the performance of each method is improved as the ratios of training datasets increase. When the ratio of the training dataset size increases to 50%, MLDL also performs the best among all the methods on the Corel5k dataset, and the proposed method is the second best approach. Different from this, considering the ESP dataset, the proposed method achieves the best performance for the measures of FI and N+. This performance difference occurs because ESP has a higher TPI than Corel5K. It can achieve better precision in ESP. We also believe that more images and more labels provide more useful information for image annotation. Moreover, the proposed method can scale well on larger datasets than other methods due to the neural networks. However, the high computation cost is its shortcoming, which needs not only the hardware support but also the algorithm support. Thus, we calculate the CNN features offline. With the ratio of the training dataset increasing, more useful information can be provided. As the training set continues to increase, the performances of most of these methods will improve in turn, but due to their limitations, the performances will be stable to some extent. As the training set increases to 80% for the three datasets, our method can achieve the best results under most measures among these methods. On the one hand, more labeled images can provide more useful information for our method which can make full use of such information to improve the performance. More labeled images not only make the semantic-based image similarity more accurate but also provide more meaningful information to construct the semantic-based label cooccurrence, although the visual-based image similarity and visual-based label cooccurrence is stable. On the other hand, the CNN features help the method to find the most similar images from the view of visual features. All of these techniques can boost the annotation performance.

By summarizing the performances of the above approaches, we find that nearest-neighbor-based methods usually have promising results in annotation. However, they depend heavily on how visual features are compared, which is a truly time-consuming issue. Although MLDL and CCA-KNN methods achieve better performance than other methods, they have their own shortcomings. The MLDL model explores the underlying correlation among labels by using a multilabel dictionary learning algorithm, which puts the label correlation in the input space rather than in the output space. It depends too much on the labeled images. Moreover, a dictionary learning method is a practical time-consuming method. Additionally, the proposed method and CCA-KNN model both use the CNN features. The experimental results of the proposed method are slightly better than that of CCA-KNN. We think there are three reasons for this. First, the CNN features we used to construct the image similarity matrix can accurately display the similarity of images. Second, the three relationships we used to factorize make full use of the relations of images and semantic concepts, further reducing the semantic gap. Third, we also consider the visual-based label cooccurrence and semantic-based image similarity, which can help to find some latent features between images and labels. These promising results have proven the efficiency of the combination of these techniques.

E. PARAMETER TUNING

In this method, there are 6 important hyperparameters. These parameters include 5 regularization parameters, α, β, λ_U, λ_V, λ, and the latent rank of the factor matrices K. In this subsection, the effects of these parameters will be studied and evaluated.

1) IMPACT OF THE IMAGE-TO-IMAGE VISUAL INFORMATION

The parameter α controls the contribution of image-to-image visual information to the objective function in Eq. (7). To study the impact of this information, we vary the value of α by fixing other parameters. In this study, we fix $\beta = 100$, $\lambda_U = 100$, $\lambda_V = 100$, $\lambda = 10$ and $K = 20$. We also set the ratio of the training set to 20%. Moreover, we implement these experiments on the three datasets. In this way, we make

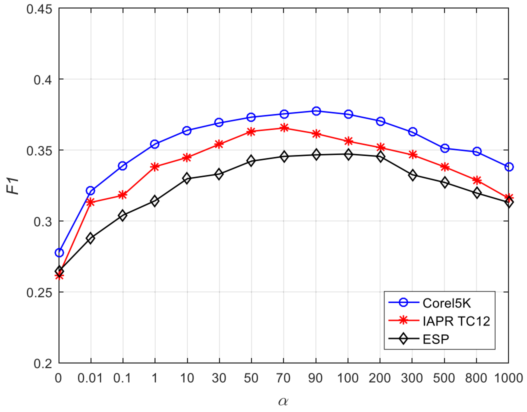

FIGURE 4. Impact of α in different datasets.

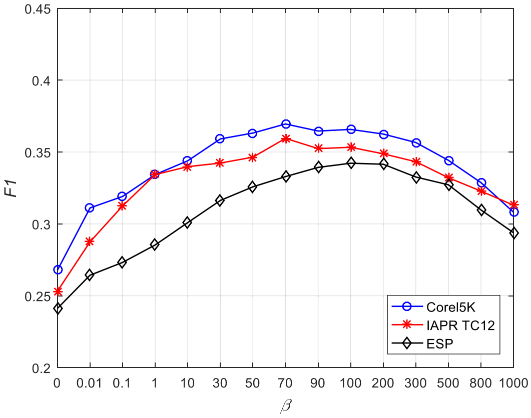

FIGURE 5. Impact of β in different datasets.

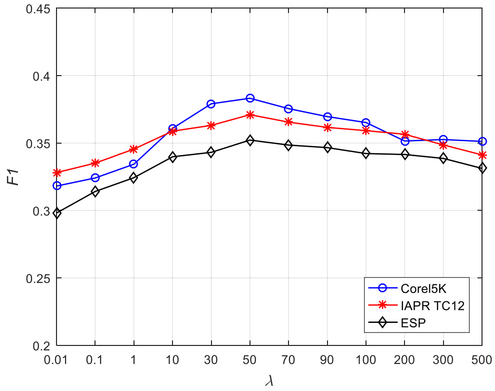

FIGURE 6. Impact of parameter λ.

sure that both image-label and visual-based image similarity can contribute to the objective function.

As shown in Figure 4, the F1 score of the proposed method first increases and later decreases as α increases. This occurs because when α is too small, the model cannot fully utilize the information from the image-image visual-based similarity to find the most visually similar image features. However, when α is too large, the image-image visual-based information will dominate the objective function in Eq. (7), thus overwhelming the label information from image-label matrix R and label-label cooccurrence matrix C. Additionally, matrix R is the relation matrix between images and labels, which mainly describes the semantic relation of image-to-label. Therefore, visual-based image information is helpful and necessary for the factorization. Note that, when $\alpha = 0$, the method is equal to only exploiting other additional information sources, i.e., the label-label information. Thus, the performance at $\alpha = 0$ is lower than the performance at $\alpha > 0$. From Figure 4, we can see that CG-CNMF achieves the best performance at the range of $\alpha = 90 \sim 300$.

2) IMPACT OF THE LABEL-LABEL COOCCURRENCE INFORMATION

In this part, we will study the impact of parameter β, which controls the contribution of the label cooccurrence information to the objective function. In this study, we fix $\alpha = 200, \lambda_U = \lambda_V = 100, \lambda = 10$ and $K = 40$ according to the previous study. Then, we search the parameter β within the set $\{0,0.01,0.1,1,10,30,50,70,90,100,200,300,500,800,1000\}$ and show the results in Figure 5.

As shown in Figure 5, we similarly observe the method's performance, which first increases and later decreases as β increases. As we know, with β increasing, increasingly more label-label semantic-based cooccurrence information can be added to help the matrix factorization to find more interpretable latent feature factors. However, to some extent, the information will be saturated; thus, the performance will be hard to improve. Even more, after saturated, more information becomes noisy information that causes a decrease in performance. In Figure 5, when $\alpha \geq 100$, the performance will decrease. When $\beta = 90 \sim 100$, the method will achieve a better performance.

3) IMPACT OF REGULARIZATION PARAMETER λ

In this method, the parameter λ has two effects. On the one hand, it increases the robustness of the method. On the other hand, it prevents the method from overfitting when the other parameters are too small. We conduct the experiment considering the 20% training set and evaluate the impact of λ. The other parameters are $\alpha = 200, \beta = \lambda_U = \lambda_V = 100$ and $K = 40$. The F1 results are shown in Figure 6.

We can see from the figure that when λ increases, the performance can be improved slightly. When $10 \leq \lambda \leq 50$, the method can achieve the best performance on the three datasets. However, a too-large λ cannot improve the performance significantly but can introduce much computation and cause the model to converge slowly. Thus, we take $\lambda = 10$ for a tradeoff.

4) IMPACT OF NUMBER OF LATENT FEATURES

This method is based on the low-rank matrix factorization. Thus, the number of latent features K is an important parameter for the performance. Here, we conduct a set of experiments on the 20% training sets to evaluate the effect of K. We set other parameters as $\alpha = 200, \beta = 100, \lambda_U = \lambda_V = 100$

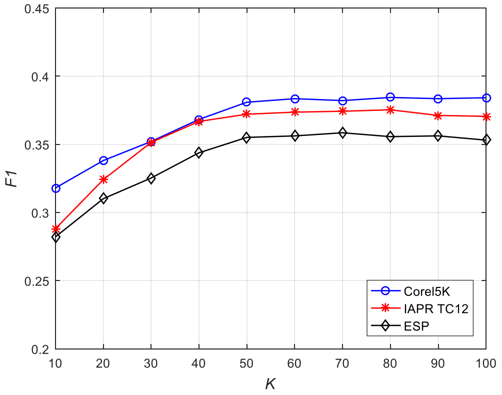

FIGURE 7. Impact of number of latent features K.

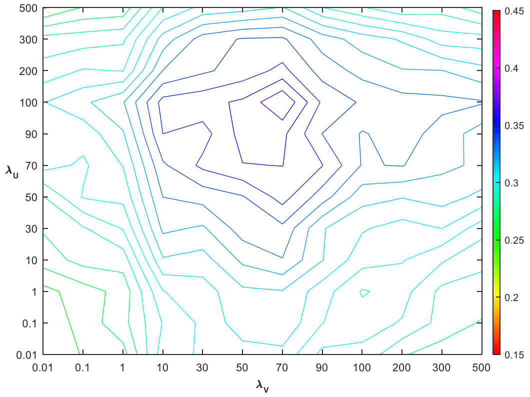

FIGURE 8. Impact of λ_U and λ_V on the Corel5K dataset.

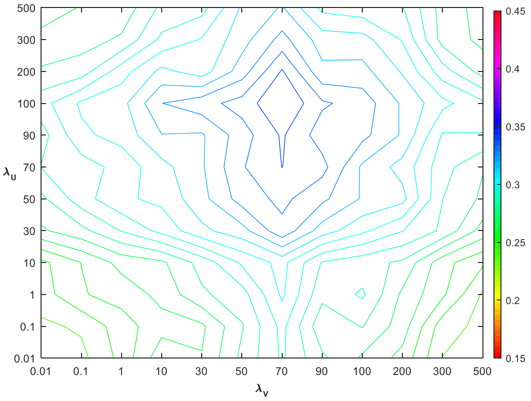

FIGURE 9. Impact of λ_U and λ_V on the IAPR TC12 dataset.

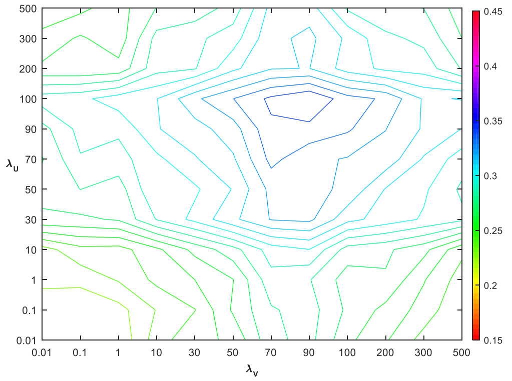

FIGURE 10. Impact of λ_U and λ_V on the ESP dataset.

and $\lambda = 10$. Then, we change the value of K within the following set $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and show the F1 results in Figure 7.

It is worth noting that the higher the dimensionality of the latent features, the better the performance is. This is because the more useful information can be obtained from images or labels. However, when the dimensionality increases to some extent, the performance will be stable or even worse. In our empirical study, the range in Corel5K is $K \geq 50$, in IAPR TC12 is $K \geq 40$ and ESP is also $K \geq 40$. These results indicate that more labels for each image can provide more information for the factorization and there is no need for large dimensionality of latent features. However, the higher the dimensionality, the more computation cost will be needed. Consequently, in our experiments, we set $K = 40$ to obtain a tradeoff.

5) IMPACT OF REGULARIZATION PARAMETERS λ_U AND λ_V

The parameter λ_U denotes the importance of semantic-based image similarity, while λ_V weights the importance of visual-based label similarity; therefore, they should be set with nonnegative values. We evaluate these parameters by empirically fixing others and implement the experiments on the 20% training sets for three datasets. Because there is no prior knowledge about the importance of image or label similarity, we set one of them to a fixed value within the set $\{0.01, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500\}$ and iteratively increase the value to reach a better result. Moreover, we fix the other parameters as $\alpha = 200, \beta = 100, \lambda = 10$ and $K = 40$. The results for the three datasets are shown in Figure 8, 9 and 10, respectively.

From these figures, we can see that optimal value can be achieved when λ_U and λ_V increase to some extent. The region is approximately at $\lambda_U = 90 \sim 100$ and $\lambda_V = 70 \sim 90$ for the three datasets. Moreover, the results also indicate that semantic-based image similarity information is slightly more important than visual-based label similarity in this collaborative-based image annotation method. This proves that the performance can benefit from useful visual-based label cooccurrence information, which compensates the label information from another view.

6) IMPACT OF REGULARIZATION PARAMETERS α AND β

In our method, if $\alpha = 0, \lambda_U = 0$, we only utilize label cooccurrence information. When $\beta = 0, \lambda_V = 0$, we only use the information from the image visual-based similarity matrix to help to factorize the image-label matrix. Furthermore, if $\alpha = \beta = 0$, $\lambda_U, \lambda_V \neq 0$, it is a standard graph nonnegative matrix factorization (GNMF). To obtain the best performance, we will search for the best combination

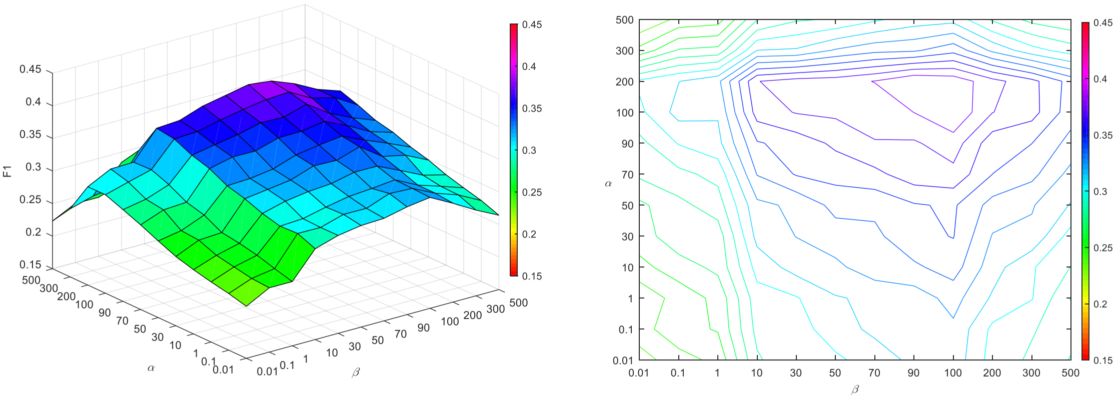

FIGURE 11. The performance of CG-CNMF by varying the bias terms α and β in Corel5k.

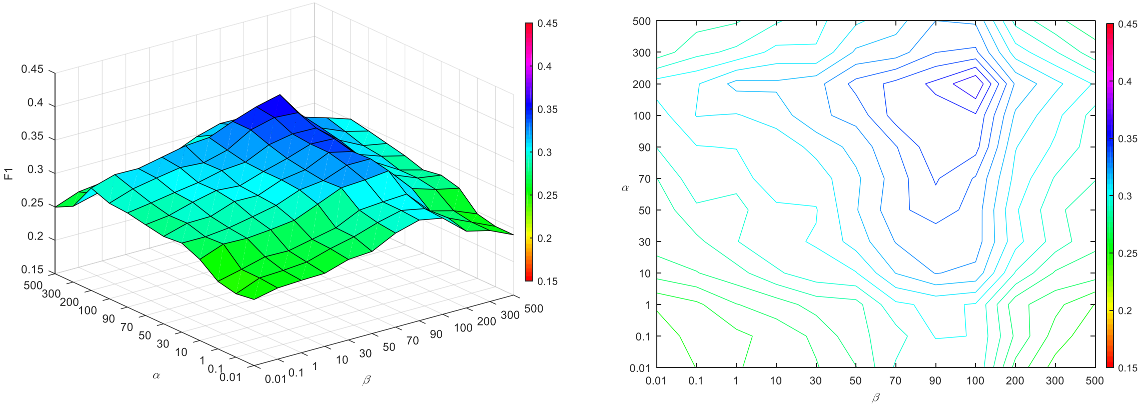

FIGURE 12. The performance of CG-CNMF by varying bias terms α and β in IAPR TC12.

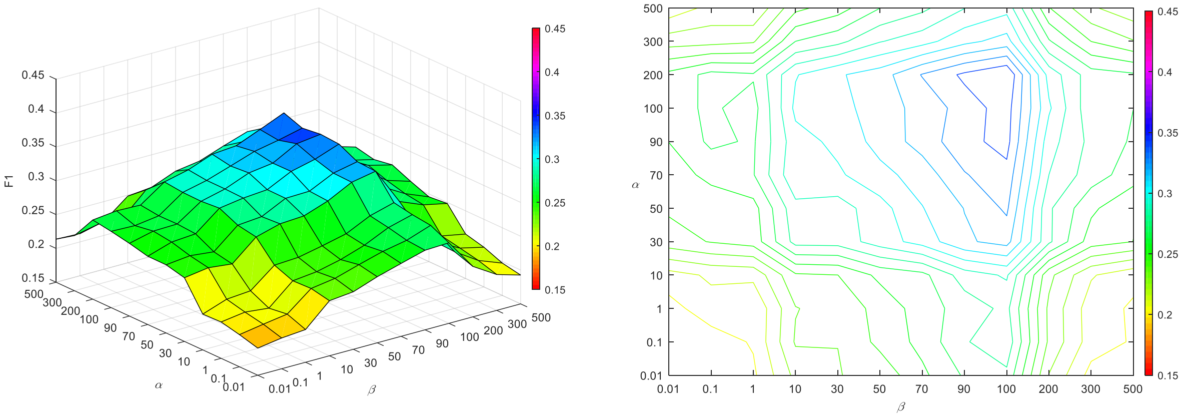

FIGURE 13. The performance of CG-CNMF by varying bias terms α and β in ESP dataset.

of α and β for our method. We fix other parameters as $\lambda_U = 100, \lambda_V = 90, \lambda = 10$ and $K = 40$. The results are shown in Figure 11, 12 and 13 for the three datasets.

From these figures, we can observe that when $\alpha = 0$ and $\beta = 0$, the performance is obviously not satisfactory since less image and label information is involved. However, the performance improves as α and β increase. It is worth noting that there is a region where the optimal values of α and β ensure the best annotation performance. The region is approximately at $\alpha = 100 \sim 200$ and $\beta = 90 \sim 100$. Moreover, the results also indicate that image-image visual-based similarity information plays a more important role than label-label semantic-based cooccurrence information for image annotation in this collaborative-based method; due to the image-label matrix and label-label cooccurrence, both provide some semantic information and visual-based information for images. Thus, the visual-based information is relatively more important in the proposed method.

VI. CONCLUSIONS

In this paper, we present a novel method, named CG-CNMF, for multilabel image annotation. We cast the image annotation problem as a label recommending problem. The first step of our method is factorizing the incomplete image-label matrix into two latent feature matrices: the latent image factor matrix and latent label factor matrix. To fully utilize much information from images and labels and address the sparsity of the image-label matrix, we consider multiple sources from the data to help the matrix factorization procedure find the most interpretable latent features. By sharing some variables, this method investigates two other relationships: the image-to-image relation and the label-to-label relation. These relations can effectively address the issues of sparsity, semantic gap, weak labeling and class imbalance, which can boost the performance of image annotation in turn. To further narrow down the semantic gap, we use a deep neural network architecture to extract high-level visual features and then construct the visual-based image similarity matrix. The results have proven the efficiency of the CNN features.

In the second step, the image annotation task can be achieved by recovering the image-label matrix. We reconstruct the image-label matrix by the product of the learned latent image matrix and label matrix. Finally, we recommend the labels for each test image according to the recovered matrix. Thus, the performance of annotation mainly depends on the latent image feature matrix and label feature matrix. To find the most proper latent image and label features, we not only employ the semantic-based and visual-based similarity for images but also consider the visual-based and semantic-based cooccurrence for labels. It is obvious that such meaningful information can efficiently boost the annotation performance. Experimental results have proven this.

There are remaining issues for us to address. In the future, we will investigate how to accelerate the multiplicative updating process and reduce the time complexity of the method. Furthermore, we will consider the word embedding model in our method to further improve the annotation accuracy. We believe our work will provide a more efficient image annotation framework.

APPENDIX

To prove Theorem 1, we will show that the objective function in Eq. (7) is nonincreasing under the steps in Eqs. (8)-(11) separately and hence converges to a local minimum under each updating rule. To achieve this, we employ an auxiliary function that was first used in [53]. First, let us introduce the following definition and lemmas:

Definition 1: $K(A, A')$ is an auxiliary function of $L(A)$ if the following conditions are satisfied:

$K(A, A') \geq F(A)$, and $K(A, A') = F(A)$.

Lemma 1: If $K(A, A')$ is an auxiliary function of $F(A)$, then $F(A)$ is nonincreasing under the following updating rule:

$$A^{t+1} = \arg\min_{A} K(A, A^t)$$

where A^t is the t^{th} update iteration of A.

Because $F(A^{t+1}) \leq K(A^{t+1}, A^t) \leq K(A^t, A^t) = F(A^t)$, thus $F(A)$ is decreasing monotonically. Consequently, to prove $F(A)$ converges to a local minimum, we can find an auxiliary function for it. First, we construct the auxiliary functions for the objection function in Eq. (7) with respect to U, V, P and Z. The following lemmas will be utilized.

Lemma 2 [63]: For any matrices $D \in \mathbb{R}_+^{m\times r}, E \in \mathbb{R}_+^{m\times r}, E' \in \mathbb{R}_+^{m\times r}$, we have the following inequality:

$$Tr(D^T E') \geq \sum\nolimits_{ij} D_{ij}E_{ij}(1 + \log \frac{E'_{ij}}{E_{ij}})$$

Lemma 3 [64]: For any nonnegative matrices $A \in \mathbb{R}_+^{n\times n}, B \in \mathbb{R}_+^{k\times k}, Q \in \mathbb{R}_+^{n\times k}, Q' \in \mathbb{R}_+^{n\times k}$, where A and B are symmetric matrices, we have the following inequality:

$$\sum\nolimits_{ij} \frac{(AQ'B)_{ij}Q_{ij}^2}{Q'_{ij}} \geq Tr(Q^T AQB)$$

Lemma 4 [63]: For any symmetric matrix $O \in \mathbb{R}_+^{r\times r}$, and any matrices $W \in R_+^{m\times r}, W' \in R_+^{m\times r}$, we have the following inequality:

$$\sum\nolimits_{ij} \frac{(WO)_{ij}W'^2_{ij}}{W_{ij}} \geq Tr\left(W'^T W'O\right)$$

Lemma 5 [63]: For $Q \in R_+^{m\times r}, W \in R_+^{m\times r}, W' \in R_+^{m\times r}$, we have the following inequality:

$$Tr(W'^T W'Q) \geq \sum\nolimits_{ijl} B_{jl}W_{ij}W_{il}(1 + \log \frac{W'_{ij}W'_{il}}{W_{ij}W_{il}})$$

In the following, we will prove each updating rule leads the objective function to converge to a local minimum. We first prove Eq. (8) leads Eq. (7) to converge and define the following function.

$$\begin{aligned}
K(U, U') = &-\sum\nolimits_{ij} (Y \odot RV)_{ij}U'_{ij}(1 + \log \frac{U_{ij}}{U'_{ij}}) \\
&+ \frac{1}{2}\sum\nolimits_{ij} \frac{(Y \odot (U'V^T)V)_{ij}U_{ij}^2}{U'_{ij}} \\
&- \alpha \sum\nolimits_{ij} (SP)_{ij}U'_{ij}(1 + \log \frac{U_{ij}}{U'_{ij}}) \\
&+ \frac{\alpha}{2}\sum\nolimits_{ij} \frac{(U'P^T P)_{ij}U_{ij}^2}{U'_{ij}} \\
&+ \frac{\lambda_U}{2}\sum\nolimits_{ij} \frac{(L_U^+ U')_{ij}U_{ij}^2}{U'_{ij}} - \frac{\lambda_U}{2}\sum\nolimits_{ijk} \\
&\times (L_U^-)_{jk}U'_{ji}U'_{ki}(1 + \log \frac{U_{ji}U_{ki}}{U'_{ji}U'_{ki}}) \\
&+ \frac{\lambda}{2}\sum\nolimits_{ij} \frac{(U')_{ij}U_{ij}^2}{U'_{ij}}
\end{aligned}$$

Then, we prove $K(U, U')$ is an auxiliary function of $L(U)$, furthermore, it is a convex function in U and its local minimum is $U_{ij} = U'_{ij}\sqrt{\frac{[Y\odot RV+\alpha SP+\lambda_U L_U^- U]_{ij}}{[Y\odot(UV^T)V+\alpha UP^TP+\lambda_U L_U^+ U+\lambda U]_{ij}}}$.

It is obvious that $K(U, U') = L(U)$ when $U' = U$. Thus, we only need to prove $K(U, U') \geq L(U)$. From $K(U, U')$, we can find that: (a) due to Lemma 2, the first term in $K(U, U')$ is always smaller than the first term in $L(U)$; (b) due to Lemma 3, the second term in $K(U, U')$ is always larger than the second term in $L(U)$; (c) due to Lemma 2, the third term in $K(U, U')$ is always smaller than the third term in $L(U)$; (d) due to Lemma 3, the fourth term in $K(U, U')$ is always smaller than the fourth term in $L(U)$; (e) due to Lemma 4, the fifth term in $K(U, U')$ is always larger than the fifth term in $L(U)$; (f) due to Lemma 5, the sixth term in $K(U, U')$ is always smaller than the sixth term in $L(U)$; and (g) the seventh term in $K(U, U')$ is always larger than the seventh term in $L(U)$ due to Lemma 4. By summing over all the bounds, we obtain $K(U, U') \geq L(U)$. Thus, $K(U, U')$ is an auxiliary function of $L(U)$ according to Definition 1.

Third, we can find a local minimum of $\min_U K(U, U')$ by calculating the partial derivative of $K(U, U')$ and setting it to zero.

$$
\begin{aligned}
0 = \frac{\partial K(U, U')}{\partial U_{ij}} &= -(Y \odot RV)_{ij}\frac{U'_{ij}}{U_{ij}} \\
&+ \frac{(Y \odot (U'V^T)V)_{ij}U_{ij}}{U'_{ij}} - \alpha(SP)_{ij}\frac{U'_{ij}}{U_{ij}} \\
&+ \alpha\frac{(U'P^TP)_{ij}U_{ij}}{U'_{ij}} + \lambda_U\frac{(L_U^+U')_{ij}U_{ij}}{U'_{ij}} \\
&- \lambda_U(L_U^-U')_{ij}\frac{U'_{ij}}{U_{ij}} + \lambda\frac{(U')_{ij}U_{ij}}{U'_{ij}}
\end{aligned}
$$

By solving the above equation for U_{ij}, we obtain the following minimum

$$
U_{ij} = U'_{ij}\sqrt{\frac{[Y \odot RV + \alpha SP + \lambda_U L_U^- U]_{ij}}{[Y \odot (UV^T)V + \alpha UP^TP + \lambda_U L_U^+ U + \lambda U]_{ij}}}
$$

Set $U^{t+1} = U$ and $U' = U^t$ according to Lemma 1; then, we recover Eq. (8). Thus, $L(U)$ deceases monotonically and converges to a local minimum.

To prove $K(U, U')$ is convex with respect to U, we derive the following Hessian matrix,

$$
\begin{aligned}
&\frac{\partial K^2(U, U')}{\partial U_{ij}\partial U_{kl}} \\
&= \sigma_{ik}\sigma_{jl}\left[\begin{array}{l}
(Y \odot RV)_{ij}\frac{U'_{ij}}{U_{ij}^2} + \frac{[Y \odot (U'V^T)V]_{ij}}{U'_{ij}} \\
+\alpha(SP)_{ij}\frac{U'_{ij}}{U_{ij}^2} + \frac{(U'P^TP)_{ij}}{U'_{ij}} \\
+\lambda_U\frac{(L_U^+U')_{ij}}{U'_{ij}} + \lambda_U(L_U^-U')_{ij}\frac{U'_{ij}}{U_{ij}^2} + \lambda\frac{(U')_{ij}}{U'_{ij}}
\end{array}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
\Lambda_{ij} = &\frac{[Y \odot RV + SP + \lambda_U(L_U^-U')]_{ij}U'_{ij}}{U_{ij}^2} \\
&+ \frac{[Y \odot (U'V^T)V + U'P^TP + \lambda_U L_U^+ U' + \lambda U']_{ij}}{U'_{ij}}
\end{aligned}
$$

is a diagonal matrix with positive diagonal elements and σ_{ik} is a function such that $\sigma_{ik} = 1$ if $i = k$; otherwise, $\sigma_{ik} = 0$. Therefore, $K\left(U, U'\right)$ is a convex function with respect to U.

Analogous to Eq. (8), we can prove the updating rules in Eq. (9), (10) and (11) lead Eq. (7) to converge to a local minimum.

In summary, we have proven Theorem 1.

REFERENCES

[1] 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.

[2] Q. Liu and Z. Li, "Projective nonnegative matrix factorization for social image retrieval," *Neurocomputing*, vol. 172, pp. 19–26, Jan. 2016.

[3] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[4] A.-M. Tousch, S. Herbin, and J.-Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognit.*, vol. 45, no. 1, pp. 333–345, 2012.

[5] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution Satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[6] P. Wennerberg, K. Schulz, and P. Buitelaar, "Ontology modularization to improve semantic medical image annotation," *J. Biomed. Inform.*, vol. 44, no. 1, pp. 155–162, 2011.

[7] M. A. Reicher, "Customizing annotations on medical images," Google Patents 20 180 060 488 A1, Mar. 1, 2018.

[8] Y. Deng, Y. Sun, Y. Zhu, Y. Xu, Q. Yang, S. Zhang, M. Zhu, J. Sun, W. Zhao, X. Zhou, and K. Yuan, "Efforts estimation of doctors annotating medical image," 2019, *arXiv:1901.02355*. [Online]. Available: https://arxiv.org/abs/1901.02355

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

[10] Z. He, C. Chen, J. Bu, P. Li, and D. Cai, "Multi-view based multi-label propagation for image annotation," *Neurocomputing*, vol. 168, pp. 853–860, Nov. 2015.

[11] R. Rad and M. Jamzad, "Image Annotation using Multi-view Non-negative Matrix Factorization with Different Number of Basis Vectors," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 1–12, Jul. 2017.

[12] H. Ma, J. Zhu, M. R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, Aug. 2010.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[15] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, 2019.

[16] Y. Verma and C. V. Jawahar, "Image annotation by propagating labels from semantic neighbourhoods," *Int. J. Comput. Vis.*, vol. 121, no. 1, pp. 126–148, 2016.

[17] H. Wang, H. Huang, and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 793–800.

[18] H. Ge, Z. Yan, J. Dou, Z. Wang, and Z. Wang, "A semisupervised framework for automatic image annotation based on graph embedding and multiview nonnegative matrix factorization," *Math. Problems Eng.*, vol. 2018, Jun. 2018, Art. no. 5987906.

[19] Y. Gao, R. Ji, W. Liu, Q. Dai, and G. Hua, "Weakly supervised visual dictionary learning by harnessing image attributes," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5400–5411, Dec. 2014.

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

[21] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1643–1650.

[22] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 605–614.

[23] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Image annotation via graph learning," *Pattern Recognit*, vol. 42, no. 2, pp. 218–228, 2009.

[24] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2014, pp. 512–519.

[26] S. Rajaraman, S. K. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. R. Thoma, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, Apr. 2018, Art. no. e4568.

[27] R. Merris, "Laplacian matrices of graphs: A survey," *Linear Algebra Appl.*, vols. 197–198, no. 2, pp. 143–176, 1994.

[28] Z. Wang, H. Yi, J. Wang, and D. Feng, "Hierarchical Gaussian mixture model for image annotation via PLSA," in *Proc. 5th Int. Conf. Image Graph.*, Sep. 2009, pp. 384–389.

[29] L. Du, L. Ren, L. Carin, and D. B. Dunson, "A Bayesian model for simultaneous image clustering, annotation and object segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 486–494.

[30] D. Tian and Z. Shi, "Automatic image annotation based on Gaussian mixture model considering cross-modal correlations," *J. Vis. Commun. Image Represent.*, vol. 44, pp. 50–60, Apr. 2017.

[31] K. Pliakos and C. Kotropoulos, "PLSA driven image annotation, classification, and tourism recommendation," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 3003–3007.

[32] A. A. Olaode, G. Naghdy, and C. A. Todd, "Unsupervised image classification by probabilistic latent semantic analysis for the annotation of images," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Nov. 2014, pp. 1–8.

[33] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[34] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent Dirichlet allocation for image annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3408–3415.

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

[36] X. Jia, F. Sun, H. Li, Y. Cao, and X. Zhang, "Image multi-label annotation based on supervised nonnegative matrix factorization with new matching measurement," *Neurocomputing*, vol. 219, pp. 518–525, Jan. 2017.

[37] F. Briggs, X. Z. Fern, R. Raich, and Q. Lou, "Instance annotation for multi-instance multi-label learning," *ACM Trans. Knowl. Discovery Data*, vol. 7, no. 3, 2013, Art. no. 14.

[38] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2126–2136.

[39] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[40] F. Su and L. Xue, "Graph learning on K nearest neighbours for automatic image annotation," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 403–410.

[41] 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.

[42] Y. Verma and C. V. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.

[43] Y. Ma, Q. Xie, Y. Liu, and S. Xiong, "A weighted KNN-based automatic image annotation method," *Neural Comput. Appl.*, vol. 2019, pp. 1–12, Mar. 2019.

[44] X. Hu and X. Qian, "A novel graph-based image annotation with two level bag generators," in *Proc. Int. Conf. Comput. Intell. Secur.*, Dec. 2009, pp. 71–75.

[45] L. Xi, R. Liu, L. Fei, and Q. Cao, "Graph-based dimensionality reduction for KNN-based image annotation," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 1253–1256.

[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 359–368.

[52] S. Boyd, L. Vandenberghe, and L. Faybusovich, "Convex optimization," *IEEE Trans. Autom. Control*, vol. 51, no. 11, p. 1859, Nov. 2006.

[53] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 556–562.

[54] 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 *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.

[55] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. Int. Workshop Image*, vol. 2, 2006, pp. 1–55.

[56] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2004, pp. 319–326.

[57] M. Chen, A. Zheng, and K. Q. Weinberger, "Fast image tagging," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2013, pp. 1274–1282.

[58] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 603–606.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[60] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang, "Multi-label dictionary learning for image annotation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2712–2725, Jun. 2016.

[61] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 316–329.

[62] Y. Yao, X. Xin, and P. Guo, "A rank minimization-based late fusion method for multi-label image annotation," in *Proc. 23rd Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 847–852.

[63] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 734–749, May 2010.

[64] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 126–135.

JULI ZHANG received the B.S. degree in computer science from Xi'an University of Posts and Telecommunications, Xi'an, China, in 2007, and the M.S. degree in computer application technology from Xi'an Microelectronics Technology Institute, Xi'an, in 2011, where she is currently pursuing the Ph.D. degree in computer architecture. Her main research interests include image processing, machine learning, and pattern recognition.

ZHANZHUANG HE received the Ph.D. degree in computer science from Xi'an Microelectronics Technology Institute, Xi'an, China, in 2006, where he is currently a Professor and a Doctoral Supervisor. He is also an Adjunct Professor with the Xi'an University of Technology and Xi'an Technological University. His research interests include embedded system architecture, computer operating systems, computer control, and pattern recognition.

JUNYI ZHANG received the B.S. degree in software engineering from Northeast Normal University, Changchun, China, in 2006, and the M.S. degree in computer application technology from the Xi'an Microelectronics Technology Institute, Xi'an, in 2009. He is currently pursuing the Ph.D. degree in software engineering with Xi'an Jiaotong University. He is also a Senior Engineer with China Minsheng Bank. His research interests include information retrieval, machine learning, and data mining.

TAO DAI received the B.E. and M.S. degrees in software engineering from Xi'an Jiaotong University, China, in 2008 and 2011, respectively, where he is currently pursuing the Ph.D. degree with the School of Software Engineering. His main research interests include machine learning and information retrieval.