Information Sciences 451–452 (2018) 180–194

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Joint multi-view representation and image annotation via optimal predictive subspace learning

Zhe Xue[a,d], Guorong Li[a,b,c,*], Qingming Huang[a,b,c,*]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences (CAS), Beijing 100190, China
[b] Key Lab. of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100080, China
[c] Key Laboratory of Big Data Mining and Knowledge Management, CAS, China
[d] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, 100876, Beijing, China

ARTICLE INFO

Article history:
Received 29 September 2017
Revised 19 March 2018
Accepted 22 March 2018
Available online 23 March 2018

MSC:
00-01
99-00

Keywords:
Multi-view data
Image annotation
Representation learning
Subspace learning
Structure preserving

ABSTRACT

Image representation and annotation are two key tasks in practical applications such as image search. Existing methods have tried to learn an effective representation or to predict tags directly using multi-view low-level visual features, which usually contain redundant information. However, these two tasks are closely related and interact on each other. A suitable image representation can yield better image annotation results, which in turn can effectively guide the image representation learning. In this paper, we propose to jointly conduct multi-view representation and image annotation via optimal predictive subspace learning, making the two tasks promote each other. Specifically, for subspace learning, visual structure and semantic information of images are exploited to make the learned subspace more discriminative and compact. For tag prediction, support vector machines (SVM) is adopted to obtain better tag prediction results. Then to simultaneously learn image representation, tag predictors and projection function, the three subproblems are combined into a unified optimization objective function and an alternative optimization algorithm is derived to solve it. Experimental results on four image datasets illustrate that our method is superior to the other image annotation methods.

1. Introduction

As Internet technology and digital equipments develop rapidly in recent years, web users can upload their multi-media data such as photos conveniently. A plenty of unlabeled images are uploaded to the social websites such as Facebook and Flickr every day. Being great helpful to image data management and searching, image annotation has drawn an increasing research interest.

Researchers have proposed a variety of image annotation methods in recent years [17,21,28,33]. Nearest neighbour methods are used to predict image tags by transferring tags from nearest neighbours in the training set [21,33]. A linear reconstruction model is proposed for image tagging [28] by reconstructing an incomplete tag matrix for each image and each tag with sparsity constraints. Moreover, TMC [46] adopts matrix completion method to fill in the missing tags and correct noisy tags for image annotation. Using a tag matrix to represent image-tag relation, the optimal tag matrix is calculated by preserving both visual and textual similarity of images.

[*] Corresponding authors at: School of Computer and Control Engineering, University of Chinese Academy of Sciences (CAS), Beijing, 100190, China.
E-mail addresses: liguorong@ucas.ac.cn, guorong.li@vipl.ict.ac.cn (G. Li), qmhuang@ucas.ac.cn (Q. Huang).

https://doi.org/10.1016/j.ins.2018.03.051
0020-0255/© 2018 Elsevier Inc. All rights reserved.

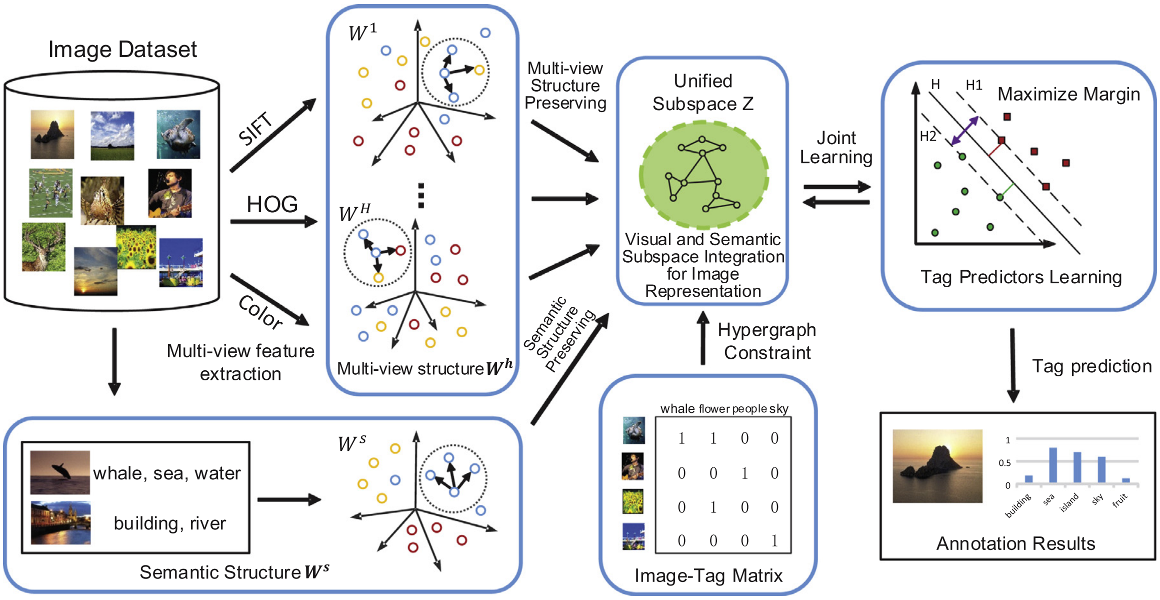


Fig. 1. The framework of OPSL. First, multi-view features are extracted from image data. Then, the multi-view structure and semantic structure are modeled by nearest neighbor graphs. Furthermore, we jointly conduct visual and semantic subspace integration and tag prediction learning, where a unified subspace and tag predictors (SVM) are learned. Finally, OPSL is capable of predicting tags for unlabeled images.

For image tags prediction, description ability of the visual feature is one of the critical factors. Powerful image descriptors can generate more effective annotation results. Compared to single type of feature, multi-view feature can provide more comprehensive description about images and yield more powerful descriptions for learning tasks. Some methods [17,24] are proposed to fuse multi-view features for image annotation. By using k-nearest neighbour method, NMF-KNN [24] constructs multi-view matrix containing different visual features and tags, and then predict tags by jointly factorizing multiple matrices. To make use of unlabeled data, a semi-supervised image annotation method OGL [17] is developed to learn an optimal similarity graph from multi-view graphs, which accurately captures the visual and semantic relationships among images. However, most of the image annotation methods predict tags directly using low-level visual features, which commonly contain redundant and noisy components. Moreover, the gap between high-level semantics and low-level visual features makes it difficult to derive true semantic information directly from the low-level features. To solve this problem, subspace learning methods [3,35] are developed to learn new data representation from original feature space. Nevertheless, most of them ignore the relation between subspace learning and the follow-up task, and just try to learn the subspace without consideration of the task. As a result, the learned representation is not guaranteed to be appropriate for the follow-up task.

In this paper, we propose *Optimal Predictive Subspace Learning* (OPSL) to jointly conduct multi-view representation learning and image annotation, and Fig. 1 shows its framework. Different from existing methods that predict tags from original feature space, OPSL predicts tags from the learned optimal predictive subspace, where the images can be represented more properly for tag prediction. To ensure that the learned subspace can obtain optimal predictive ability, we adopt the following two strategies. First, by performing visual subspace learning and semantic subspace learning, both visual structure and semantic information of images are preserved and exploited for subspace learning. In visual subspace learning, we adopt softmax activation function to minimize the disagreement between different views to make use of multi-view complementary information. In semantic subspace learning, we exploit high-order correlations among images and tags to make the semantic subspace more compact and effective. Second, the correlations between subspace learning and image annotation are leveraged. In the learned subspace, we train support vector machine (SVM) classifier for tag prediction, which in turn is used to guide the subspace learning. In this way, the discriminating power of SVM classifiers and the learned subspace can be mutually improved. Extensive experiments on four datasets demonstrate the effectiveness of the proposed method. The main contributions of this paper are summarized as follows.

- An optimal predictive subspace learning method is developed to simultaneously conduct the image representation learning and image annotation. Through exploiting the correlations of the two tasks and making them promote each other, the proposed method can achieve better image annotation performance.
- To make the learned representation more discriminative and compact, both visual structure and semantic information of images are preserved in the learned subspace. In addition, hypergraph is utilized to capture high-order correlations of images and tags.

- SVM classifiers are adopted to predict image tags for their powerful prediction ability. In turn, the trained SVM classifiers are also used to guide the subspace learning, so that the learned subspace is more suitable for annotation task and yields accurate prediction.

A preliminary version of this paper has been published in [48]. This paper extends [48] from the following aspects. First, we introduce semantic subspace learning to further improve the discriminative ability of image representation in Section 3.3.2. Second, we adopt non-linear method to learn the projection functions in Section 3.5, which accurately maps the unlabeled data into the learned subspace. Finally, we conduct image annotation experiments in another dataset (ICPATC-12) to fully illustrate the effectiveness of OPSL.

This paper is organized as follows. We introduce related work in Section 2. Then the proposed method and optimization algorithm are presented in Section 3 and Section 4, respectively. Section 5 illustrates image annotation performance and parameter analysis. Finally, we conclude this paper in Section 6.

2. Related work

Automatic image annotation is one of the most important tasks in computer vision, which aims to predict proper tags for unlabeled images. A variety of methods are proposed for image annotation in recent years.

Nearest neighbor methods are adopted for tag prediction in [21,33,40]. TagProp [21] fuses several kinds of image features to obtain a discriminative metric, then the nearest neighbor method is used to predict image tags. 2PKNN [40] is a variant of nearest neighbor model, which achieves the benefits of both “image to image” and “image to tag” similarities. In addition, matrix completion is used for image annotation task in [15,46]. The low-rank property of the tag matrix as well as visual similarities of images are leveraged to recover the missing tags. Moreover, dictionary learning is applied for image annotation in [9,23]. SLED [9] learns a semantical dictionary to obtain more discriminative image representation. MLDL [23] introduces dictionary learning and designs label consistency constraint to obtain more effective data representation.

A critical factor in tag prediction is the descriptive and discriminative ability of visual features. Instead of predicting tags using single feature [16,44,46], many methods [2,7,22,31,50] exploit multi-view features for image tagging and achieve competitive annotation performance thanks to abundant and complete descriptions provided by multi-view data. To effectively utilize multi-view data, graph based methods are developed [2,7,17]. Cai et al. [7] learn a shared class indicator matrix from multiple views, and the weights of different views can be learned adaptively. Amiri et al. [2] firstly construct a subgraph for each view and then connect them to form a supergraph. By extracting some prototypes from features, their label inference is scalable. In addition, sparsity constraints are adopted for multi-view image annotation. A Hessian discriminative sparse coding method is developed in [31], which incorporates Hessian regularization into discriminative sparse coding for image annotation.

In order to obtain more reliable and effective representation from original features, researchers have developed many techniques such as dimension reduction, subspace learning and spectral embedding for representation learning. Some linear methods such as principle component analysis (PCA) [3] and non-negative matrix factorization (NMF) [26] are developed to learn new data representation, while these methods may ignore the inherent nonlinearity of the data. Therefore, local manifold structure of data is utilized to learn data representation such as LE [4], LLE [35] and ISOMAP [39].

To learn effective representation for multi-view data, multi-view learning methods are developed [20,25,27,32,47,49]. Some methods adopt graphs to represent multi-view data and then learn a unified representation from multiple views [32,47]. Bo et al. propose a general spectral embedding framework for multi-view dimensionality reduction [32]. Lin et al. perform multi-view dimensionality reduction based on multiple kernel learning [27], which provides convenience of using multiple kinds of visual features. Gui et al. [20] adopt local patches to capture the structure of each view, and obtain the unified representation by global coordinate alignment. Furthermore, matrix factorization methods are adopted in [25,49] to learn low-dimensional representation. By using group sparsity constraint, these methods can well preserve the shared and private information of each view in the learned subspace, so that more accurate and complete description can be obtained for multi-view learning.

Recently, several deep learning based methods are developed for multi-label image classification, which accomplish similar tasks with this work. WARP [18] adopts top-k ranking objective for multi-label prediction, which assigns smaller loss weights to positive labels. CNN-RNN [42] treats multi-label learning as a sequential prediction problem, where the label dependency is solved by Recurrent Neural Networks (RNN). Zhu et al. [54] propose a spatial regularization network to effectively explore the underlying spatial relations between labels. Wang et al. [45] propose a recurrent memorized-attention module to search for meaningful and discriminative regions in terms of multi-label classification. Through leveraging convolutional neural networks, recurrent neural networks and attention mechanism, deep learning based methods achieve promising multi-label learning performance.

3. Methodology

In this section, we will introduce OPSL in detail. For better understanding, the preliminaries that are closely related to OPSL will be provided at first.

Table 1
The notations and their corresponding descriptions.

Notation	Description
$\{\mathbf{X}^h\}_{h=1}^H$	multi-view matrces
$\mathbf{U}$	multi-view data represented by anchor graph
$\mathbf{T}$	the tag matrix
$\mathbf{W}^h$	affinity matrix for structure graph of view h
$\mathbf{W}^s$	affinity matrix for tag graph
$\mathbf{L}_i$	graph Laplacian of image hypergraph
$\mathbf{L}_t$	graph Laplacian of tag hypergraph
$\mathbf{Z}_v$	visual subspace
$\mathbf{Z}_s$	semantic subspace
$\mathbf{Z}$	the learned image representation
$\mathbf{P}$	projection function
α	Lagrange multipliers in SVM
H	number of views
n	number of images
m	number of tags

3.1. Preliminaries

In this subsection, variables that will appear in our method are firstly explained. Then manifold and hypergraph learning and SVM which are used in OPSL are briefly introduced (Table 1).

3.1.1. Notations

Given a matrix **M, M**$_{ij}$ and $\mathbf{M}_{i\cdot}$ denote its (i, j)th entry and ith row, respectively. The trace of $\mathbf{M}$ is $Tr[\mathbf{M}]$. For $\mathbf{M} \in \mathbb{R}^{n\times m}$, let $||\mathbf{M}||_F$ represent its Frobenius norm, and $\ell_{2,1}$-norm is defined as

$$||\mathbf{M}||_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}\mathbf{M}_{ij}^2} \tag{1}$$

Given n multi-view data with H views, we use a set of matrices $\mathcal{X} = \{\mathbf{X}^{(i)} \in \mathbb{R}^{n\times d_i}\}_{i=1}^H$ to denote them, where d_i is the dimensionality of the ith view. For every training data, a m-dimensional binary-valued tag vector $\{t_i\}_{i=1}^n$, $t_i \in \{0,1\}^m$ is assigned, and the tag matrix is formed by $\mathbf{T} = [t_1, \ldots, t_n]^{\mathrm{T}} \in \mathbb{R}^{n\times m}$.

3.1.2. Manifold and hypergraph learning

The study on manifold learning has shown that data are likely to reside on a low-dimensional submanifold of the ambient space [5,35,37]. The discriminating power of the learned subspace can be enhanced by exploiting the intrinsic manifold structure of data. The nearest neighbor graph can effectively model the manifold structure of data. Given image data $\{x_i\}_{i=1}^n$ with H views, we construct a k-nearest neighbor graph G^h for each view $h = 1, \ldots, H$, and the affinity matrix $\mathbf{W}^h \in \mathbb{R}^{n\times n}$ is computed by the following equation,

$$\mathbf{W}_{ij}^h = \begin{cases} \exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) & x_i \in N_k(x_j)\ or\ x_j \in N_k(x_i) \\ 0 & otherwise, \end{cases} \tag{2}$$

where σ is the parameter in gaussian function and can be obtained by self-tuning method [51].

A hypergraph [1] is composed of an edge set $\mathcal{E}$ and a vertex set $\mathcal{V}$. Different from the traditional graph, each edge in the hypergraph connects two or more vertices. A hyperedge $e \in \mathcal{E}$ is a subset of $\mathcal{V}$, and it is assigned with weight $w(e)$. The vertex-edge incidence matrix $\mathbf{Q} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$ is defined as

$$\mathbf{Q}(v, e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & otherwise \end{cases} \tag{3}$$

Given $\mathbf{Q}$, the degree of a hyperedge e is

$$d(e) = \sum_{v\in V}\mathbf{Q}(v, e) \tag{4}$$

and the degree of a vertex $v \in \mathcal{V}$ is defined as

$$d(v) = \sum_{v\in e, e\in E} w(e)\mathbf{Q}(v, e) \tag{5}$$

In hypergraph learning, Zhou et al. [53] develop hypergraph Laplacian $\mathbf{L}$ to encode the higher-order relations of data, which is defined as

$$\mathbf{L} = \mathbf{I} - \mathbf{S}, \quad \mathbf{S} = \mathbf{I} - \mathbf{L} = \mathbf{D}_v^{-\frac{1}{2}}\mathbf{Q}\mathbf{W}_H\mathbf{D}_e^{-1}\mathbf{Q}^T\mathbf{D}_v^{-\frac{1}{2}} \tag{6}$$

where $\mathbf{W}_H$, $\mathbf{D}_e$ and $\mathbf{D}_v$ denote hyperedge weights, diagonal matrices of hyperedge degrees and vertex degrees, respectively.

3.1.3. Support vector machine

To annotate unlabeled images, classifiers are commonly adopted to predict tags. Support vector machines (SVM) is an effective method for data classification and obtains satisfactory learning performance [36]. For a simple binary classification problem with n training data, let $y \in \{-1, +1\}^n$ denote the vector of labels and $\mathbf{Y} = diag(y)$, and $\mathbf{K}$ is the kernel matrix. The formulation of dual SVM is presented as follows,

$$\max_{\{0\leq\alpha\leq C, \alpha^T y=0\}} \alpha^{\mathrm{T}}\mathbf{1} - 0.5Tr\Big(\mathbf{K}(\mathbf{Y}\alpha)(\mathbf{Y}\alpha)^{\mathrm{T}}\Big), \tag{7}$$

where $\mathbf{1} = [1, \dots, 1]^{\mathrm{T}} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}^n$ are Lagrange multipliers, and C is the misclassification penalty.

3.2. Method overview

The proposed method OPSL is composed of three subproblems: visual and semantic subspace integration for image representation, tag predictors learning and projection function learning. In visual and semantic subspace integration (Section 3.3), a new subspace is learned to represent multi-view data where both visual and semantic information can be effectively preserved. In tag predictors learning (Section 3.4), SVM is trained in the learned subspace to predict tags. To annotate unlabeled images in the subspace, projection function learning (Section 3.5) is conducted to map unseen images into the learned subspace. The three subproblems are then integrated into a unified objective function for learning (Section 3.6). Next, we will introduce OPSL in detail.

3.3. Visual and semantic subspace integration for image representation

We adopt subspace learning method to learn representation for image data. To enhance the discriminative ability of the learned subspace, the subspace preserves not only multi-view structure of image data, but also semantic information of images. We introduce visual subspace $\mathbf{Z}_v \in \mathbb{R}^{n\times r_v}$ and semantic subspace $\mathbf{Z}_s \in \mathbb{R}^{n\times r_s}$ to encode multi-view structure information and the semantic information, respectively. To make full use of visual and semantic information for tag prediction, we integrate the two above subspaces into a unified one $\mathbf{Z} = [\mathbf{Z}_v, \mathbf{Z}_s]$, where $\mathbf{Z} \in \mathbb{R}^{n\times r}$, $r = r_v + r_s$. The new representation $\mathbf{Z}$ preserves both visual properties and semantic properties of image data, which makes it more compact and discriminative for image tagging. Next, we will introduce how to learn visual subspace $\mathbf{Z}_v$ and semantic subspace $\mathbf{Z}_s$, respectively, and then present the unified image representation learning objective function.

3.3.1. Multi-view structure preserving for visual subspace learning

To effectively preserve the multi-view structure of images in visual subspace $\mathbf{Z}_v$, we propose three conditions that the visual subspace should satisfy. First, the learned subspace should be locally smooth, which is achieved by imposing constraints that the visual subspace should preserve the intrinsic manifold structure of images in the original feature space. Second, the learned subspace should effectively minimize the disagreement between different views, so that multi-view complementary information can be sufficiently preserved in the learned subspace. Last, multi-view structure only provides low-level description of images, so it may contain some inaccurate information. Semantic structure of image data can be leveraged to improve the accuracy of the multi-view structure for subspace learning.

Multi-view structure of multi-view data can be modeled by k-nearest neighbor graphs $\{\mathbf{W}^h\}_{h=1}^H$. We use $\mathbf{W}^s = \mathbf{T}^{\mathrm{T}}\mathbf{T}$ to model the semantic structure of images, where the two images are more similar if they share more common tags. Inspired by the successful application of self-representation [8,14], we encode multi-view structure into visual subspace $\mathbf{Z}_v$ by the following formulation,

$$\begin{aligned}&\min_{\mathbf{Z}_v} \sum_{h=1}^{H} ||\mathbf{Z}_v - (\mathbf{W}^h \odot \mathbf{W}^s)\mathbf{Z}_v||_F^2 + \eta||\mathbf{T} - \mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{T}||_F^2 \\ &s.t.\ \mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v = \mathbf{I}\end{aligned} \tag{8}$$

where $\odot$ denotes Hadamard product. To make multi-view structure $\mathbf{W}^h$ more accurate and reliable, we adopt $\widetilde{\mathbf{W}} = \mathbf{W}^h \odot \mathbf{W}^s$ as the enhanced graph. From the new graph $\widetilde{\mathbf{W}}$, we can observe that two images, which can become the nearest neighbors, should satisfy two conditions: First, the two images should be the nearest neighbors in the visual structure graph $\mathbf{W}^h$. Second, the two images should contain at least one common tag. Furthermore, with the second term, semantic information is encoded into the $\mathbf{Z}_v$. The images sharing common tags are made to be similar in the visual subspace $\mathbf{Z}_v$. η is used to control the strength of semantic information embedding. We impose orthogonal constraint on $\mathbf{Z}_v$ to avoid the trivial solution.

From problem (8), all the views are simply treated as equally important during multi-view structure preserving, which may not be the optimal setting. Since views are different from each other, some views would generate larger disagreements and higher costs during subspace learning. To better exploit the multi-view complementary property, we aim to find a new subspace that accommodates each view well. So we propose to minimize the difference between the learned subspace and the most disagreement view (the one generating the highest embedding cost). In this way, the visual subspace $\mathbf{Z}_v$ can fully preserve the complementary information of multi-view data. In our method, softmax activation function is introduced to approximately find the most disagreement view, and formulation (9) is proposed for visual subspace learning.

$$\begin{aligned}&\min_{\mathbf{Z}_v} \frac{1}{\gamma} \log\left\{\sum_{h=1}^{H} \exp\left[\gamma||\mathbf{Z}_v - (\mathbf{W}^h \odot \mathbf{W}^s)\mathbf{Z}_v||_F^2\right]\right\} + \eta||\mathbf{T} - \mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{T}||_F^2 \\ &s.t.\ \mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v = \mathbf{I}\end{aligned} \tag{9}$$

where the smooth parameter γ is used to control the precision of approximation.

3.3.2. High-order correlation preserving for semantic subspace learning

Multi-view feature provides low-level visual information of images, which cannot generate enough accurate descriptions for images. So we introduce semantic subspace $\mathbf{Z}_s$ to improve the discriminating power of the learned representation. Notably, high-order relations among images and tags are helpful to enhance the effectiveness of the semantic information. On one hand, images having the same tags may share some common visual properties so that their representation in the subspace should be close. On the other hand, there are strong correlations between the tags that always coexist in the same image. For example, it is very likely that “sky”, “cloud” and “bird” coexist in the same image. Exploring high-order correlation of images and tags can effectively handle the noise and missing data problems in $\mathbf{T}$ and further improve the discriminative ability of semantic subspace $\mathbf{Z}_s$.

Considering the above factors, we introduce two hypergraphs to model the high-order relations of tags and images. The first one is image hypergraph. For each tag, we construct a hyperedge which includes all the images containing that tag, so the vertex-edge incidence matrix $\mathbf{Q} = \mathbf{T}$. By using normalized hypergraph Laplacian in Eq. (6), we can obtain the image hypergraph Laplacian $\mathbf{L}_i \in \mathbb{R}^{n\times n}$ which models the high-order relations among images. The second one is tag hypergraph. For each image, we construct a hyperedge that includes all the tags annotated to that image, and the vertex-edge incidence matrix $\mathbf{Q} = \mathbf{T}^{\mathrm{T}}$. Then we use Eq. (6) to obtain tag hypergraph Laplacian $\mathbf{L}_t \in \mathbb{R}^{m\times m}$ which models high-order co-existence relations among tags. Finally, the following objective function is proposed to preserve these high-order relations in the semantic subspace $\mathbf{Z}_s$,

$$\min_{\mathbf{Z}_s}\ \theta Tr(\mathbf{Z}_s\mathbf{L}_t\mathbf{Z}_s^{\mathrm{T}}) + \lambda||\mathbf{Z}_s - \mathbf{T}||_F^2 \tag{10}$$

we expect to learn a more effective semantic representation $\mathbf{Z}_s$ from tag matrix $\mathbf{T}$, and λ controls the similarity of them. The first term preserves high-order relations among images and tags in $\mathbf{Z}_s$, and θ is used to control the strength of hypergraph information preserving. By introducing semantic subspace $\mathbf{Z}_s$ for tag prediction, we can exploit tag dependencies and further improve the tag prediction performance.

3.3.3. Visual and semantic subspace integration

We integrate the visual subspace and semantic subspace into a new subspace $\mathbf{Z} = [\mathbf{Z}_v, \mathbf{Z}_s]$, which is learned by the following objective function,

$$\begin{aligned}f(\mathbf{Z}) = &\frac{1}{\gamma} \log\left\{\sum_{h=1}^{H} \exp\left[\gamma||\mathbf{Z}_v - (\mathbf{W}^h \odot \mathbf{W}^s)\mathbf{Z}_v||_F^2\right]\right\} + \eta||\mathbf{T} - \mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{T}||_F^2 \\ &+ \theta(Tr(\mathbf{Z}^{\mathrm{T}}\mathbf{L}_i\mathbf{Z}) + Tr(\mathbf{Z}_s\mathbf{L}_t\mathbf{Z}_s^{\mathrm{T}})) + \lambda||\mathbf{Z}_s - \mathbf{T}||_F^2 \\ s.t.\ &\mathbf{Z}_v{}^{\mathrm{T}}\mathbf{Z}_v = \mathbf{I}\end{aligned} \tag{11}$$

where the image hypergraph $\mathbf{L}_i$ is adopted to preserve the high-order correlations of images in $\mathbf{Z}$. By minimizing problem (11), we can obtain $\mathbf{Z}$ which preserves both visual structure and semantic information of image data.

3.4. Tag predictors learning

Our method predicts tags based on the learned image representation $\mathbf{Z}$. Considering the powerful prediction ability of SVM, we introduce SVM for tag prediction. To keep the simplicity and effectiveness of the SVM learning problem, linear kernel $\mathbf{K} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ is adopted. Tag predictors learning problem is presented as follows,

$$\begin{aligned}&\min_{\mathbf{Z}} \max_{\alpha_t} g(\mathbf{Z}, \alpha_t) = \sum_{t=1}^{m}\left[\alpha_t^{\mathrm{T}}\mathbf{1} - 0.5Tr\big(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{Y}_t\alpha_t(\mathbf{Y}_t\alpha_t)^{\mathrm{T}}\big)\right]. \\ &\qquad s.t.\ \alpha_t^{\mathrm{T}}y_t = 0, 0 \le \alpha_t \le C,\ t = 1, \ldots, m\end{aligned} \tag{12}$$

There are m tags in the dataset and each tag is trained with a SVM predictor. $y_t \in \{-1, +1\}^n$ and $\mathbf{Y}_t = diag(y_t)$ are the label vector and matrix, respectively. We train SVM based on subspace $\mathbf{Z}$ and obtain the SVM parameters α_t. In turn, the trained SVM is also used to guide the subspace learning, which makes $\mathbf{Z}$ more suitable to predict tags.

3.5. Projection function learning

We need to learn a projection function to map unlabeled image data into the subspace for image annotation. As multi-view data maybe highly non-linear distributed, we need to learn an effective non-linear mapping to project multi-view data into the new subspace. Anchor graph [11,30,43] is a fast and effective graph construction method. We adopt this method to capture non-linear structure of multi-view data and learn the projection function.

Given multi-view data matrices $\mathcal{X} = \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^H\}$, the images from the hth view is denoted by $\mathbf{X}^h = [x_1^h, x_2^h, \dots, x_n^h]^{\mathrm{T}} \in \mathbb{R}^{n\times d_h}$. First, we obtain g anchor points from the training set to represent all the training samples, which is achieved by conducting k-means on the training set and the cluster number is set to g. Let $\{c_1^h, c_2^h, \dots, c_g^h\}$ denote the obtained clustering centers, i.e., the anchor points. Then, we calculate the new representation $\mathbf{U}^h \in \mathbb{R}^{n\times g}$ as follows:

$$\mathbf{U}_{ij}^h = e^{-||x_i^h - c_j^h||^2/\sigma^2} \tag{13}$$

The number of anchor points is less than the number of training data, $g \ll n$, so the obtained data representation by anchor graph is more efficient than the traditional kernel matrix. In anchor graph methods, $\mathbf{U}^h$ is usually constructed by nearest anchors to keep its sparsity. In the experiments, we find that completely calculating $\mathbf{U}^h$ can generate better performance, so all the elements in $\mathbf{U}^h$ are calculated in this work.

After calculating the new representation of each view, we concatenates them and obtain $\mathbf{U} = [\mathbf{U}^1, \mathbf{U}^2, \dots, \mathbf{U}^H]$, $\mathbf{U} \in \mathbb{R}^{n\times gH}$. We introduce projection matrix $\mathbf{P} \in \mathbb{R}^{gH\times r}$ to project $\mathbf{U}$ into the subspace $\mathbf{Z}$. In order to reduce the influence brought by the redundant features and noise contained in multi-view data, we adopt $\ell_{2,1}$-norm for projection function learning, which can select effective dimensions by shrinking some rows to zeros. Then the projection function can be learned by solving the following problem,

$$\min_{\mathbf{Z},\mathbf{P}} h(\mathbf{Z}, \mathbf{P}) = ||\mathbf{UP} - \mathbf{Z}||_F^2 + \beta||\mathbf{P}||_{2,1}, \tag{14}$$

where parameter β controls the strength of the regularization term.

3.6. Unified objective function

We integrate image representation learning and image annotation into a unified optimization framework, which aims to make the two highly correlated tasks benefit each other to further improve the learning performance. The ultimate objective function is defined as follows,

$$\begin{aligned} O(\mathbf{Z}, \mathbf{P}, \alpha_t) &= \min_{\mathbf{Z},\mathbf{P}} \max_{\alpha_t} \left(f(\mathbf{Z}) + \mu_1 g(\mathbf{Z}, \alpha_t) + \mu_2 h(\mathbf{Z}, \mathbf{P})\right), \\ s.t.\ \mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v &= \mathbf{I}, \alpha_t^{\mathrm{T}} y_t = 0, 0 \le \alpha_t \le C,\ t = 1, \dots, m \end{aligned} \tag{15}$$

where parameter μ_1 and μ_2 are used to control the weights of the corresponding subproblems.

4. Optimization algorithm

Obviously, problem (15) is not convex over all the variables **P, Z** and α_t simultaneously, so an iterative optimization algorithm is developed to solve the problem. In each iteration, only one variable is updated. Algorithm 1 illustrates the overall optimization procedure. Next, we will present the detailed updating methods for each variable.

Algorithm 1: The algorithm of OPSL.

```
Input: Multi-view matrix $\mathbf{U}$, visual and semantic structure graphs $\{\mathbf{W}^h\}_{h=1}^H$ and $\mathbf{W}^s$, tag matrix $\mathbf{T}$, SVM label matrices
       $\{\mathbf{Y}_t\}_{t=1}^m$, image and tag hypergraphs $\mathbf{L}_i$, $\mathbf{L}_t$ and parameters: $\mu_1$, $\mu_2$, $\gamma$, $\eta$, $\beta$, $\lambda$, $\theta$, $r$.
Initialize $\mathbf{Z}$, $\mathbf{P}$ and $\{\alpha_t\}_{t=1}^m$
for $iter = 1$ to MaxIter do
    Update $\mathbf{Z}_v$ by $\mathbf{Z}_v \leftarrow \mathbf{Z}_v - \delta_1 \nabla_{\mathbf{Z}_v}\mathcal{L}(\mathbf{Z}_v)$;
    Update $\mathbf{Z}_s$ by $\mathbf{Z}_s \leftarrow \mathbf{Z}_s - \delta_2 \nabla_{\mathbf{Z}_s}\mathcal{L}(\mathbf{Z}_s)$;
    Update SVM by SMO method;
    Update $\mathbf{P}$ by equation (17);
    Update $\mathbf{D}$ by $\mathbf{D} \leftarrow \begin{bmatrix} \frac{1}{2||\mathbf{P}_{1\cdot}||_2} & & \\ & \dots & \\ & & \frac{1}{2||\mathbf{P}_{gH\cdot}||_2} \end{bmatrix}$;
end
Output: SVM Lagrange multipliers $\{\alpha_t\}_{t=1}^m$ andprojection function $\mathbf{P}$.
```

*4.1. Update for **P***

To solve projection matrix **P**, we keep the components that are related to **P** in $O(\mathbf{Z}, \mathbf{P}, \alpha_t)$ and obtain

$$\mathcal{L}(\mathbf{P}) = ||\mathbf{UP} - \mathbf{Z}||_F^2 + \beta||\mathbf{P}||_{2,1}. \tag{16}$$

By making the gradient $\nabla_{\mathbf{P}}\mathcal{L}(\mathbf{P}) = 0$, we have

$$\begin{aligned} \nabla_{\mathbf{P}}\mathcal{L}(\mathbf{P}) &= 2\mathbf{U}^{\mathrm{T}}(\mathbf{UP} - \mathbf{Z}) + 2\beta\mathbf{DP} = 0 \\ &\Rightarrow \mathbf{P} = \left(\mathbf{U}^{\mathrm{T}}\mathbf{U} + \beta\mathbf{D}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Z}, \end{aligned} \tag{17}$$

where **D** is a diagonal matrix with elements $\mathbf{D}_{ii} = \frac{1}{2||\mathbf{P}_{i.}||_2}$

*4.2. Update for **Z***

Z is composed of two parts: $\mathbf{Z}_v$ and $\mathbf{Z}_s$, which need to be solved separately. Since $\mathbf{Z}_v$ is imposed with an orthogonal constraint, it is difficult to solve $\mathbf{Z}_v$ exactly. So we use a penalty method to convert the constrained optimization problem into an unconstrained optimization problem. Keep the components that are related to $\mathbf{Z}_v$ from $O(\mathbf{Z}, \mathbf{P}, \alpha_t)$ and we have

$$\begin{aligned} \mathcal{L}(\mathbf{Z}_v) = &\frac{1}{\gamma}\log\left\{\sum_{h=1}^{H}\exp\left[\gamma||\mathbf{Z}_v - (\mathbf{W}^h \odot \mathbf{W}^s)\mathbf{Z}_v||_F^2\right]\right\} \\ &+\eta||\mathbf{T} - \mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{T}||_F^2 + \theta Tr(\mathbf{Z}_v^{\mathrm{T}}\mathbf{L}_i\mathbf{Z}_v) \\ &-0.5\mu_1\sum_{t=1}^{m}Tr\left(\mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{Y}_t\alpha_t(\mathbf{Y}_t\alpha_t)^{\mathrm{T}}\right) \\ &+\mu_2||\mathbf{UP}_v - \mathbf{Z}_v||_F^2 + \xi||\mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v - \mathbf{I}||_F^2, \end{aligned} \tag{18}$$

where we use parameter $\xi > 0$ to control the orthogonality constraint. Generally, ξ should be large to ensure the solution satisfying orthogonality. To obtain $\mathbf{Z}_v$, we divide projection function $\mathbf{P} = [\mathbf{P}_v, \mathbf{P}_s]$, where $\mathbf{P}_v \in \mathbb{R}^{gH\times r_v}$. So the gradient is derived as follows,

$$\begin{aligned} \nabla_{\mathbf{Z}_v}\mathcal{L}(\mathbf{Z}_v) = &\frac{1}{\sum_{i=1}^{H}\Delta_i}\left[\sum_{j=1}^{H}\Delta_j\left(2\mathbf{Z}_v - 4\mathbf{A}^j\mathbf{Z}_v + 2\mathbf{A}^{j^{\mathrm{T}}}\mathbf{A}^j\mathbf{Z}_v\right)\right] \\ &+4\eta(\mathbf{TT}^{\mathrm{T}}\mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v - \mathbf{TT}^{\mathrm{T}}\mathbf{Z}_v) + 2\theta\mathbf{L}_i\mathbf{Z}_v \\ &-\mu_1\sum_{t=1}^{m}\mathbf{Y}_t\alpha_t(\mathbf{Y}_t\alpha_t)^{\mathrm{T}}\mathbf{Z}_v + 2\mu_2(\mathbf{Z}_v - \mathbf{UP}_v) \\ &+4\xi(\mathbf{Z}_v\mathbf{Z}_v^{\mathrm{T}}\mathbf{Z}_v - \mathbf{Z}_v) \end{aligned} \tag{19}$$

where $\Delta_j = \exp\left[\gamma||\mathbf{Z}_v - (\mathbf{W}^j \odot \mathbf{W}^s)\mathbf{Z}_v||_F^2\right]$ and $\mathbf{A}^j = \mathbf{W}^j \odot \mathbf{W}^s$.

To solve $\mathbf{Z}_s$, we keep the parts that are related to $\mathbf{Z}_s$ in $O(\mathbf{Z}, \mathbf{P}, \alpha_t)$ and obtain

$$\begin{aligned} \mathcal{L}(\mathbf{Z}_s) = &\theta(Tr(\mathbf{Z}_s^{\mathrm{T}}\mathbf{L}_i\mathbf{Z}_s) + Tr(\mathbf{Z}_s\mathbf{L}_t\mathbf{Z}_s^{\mathrm{T}})) + \lambda||\mathbf{Z}_s - \mathbf{T}||_F^2 \\ &+\mu_2||\mathbf{UP}_s - \mathbf{Z}_s||_F^2 - 0.5\mu_1\sum_{t=1}^{m}Tr\left(\mathbf{Z}_s\mathbf{Z}_s^{\mathrm{T}}\mathbf{Y}_t\alpha_t(\mathbf{Y}_t\alpha_t)^{\mathrm{T}}\right), \end{aligned} \tag{20}$$

The derivative of $\mathcal{L}(\mathbf{Z}_s)$ with respect to $\mathbf{Z}_s$ is

$$\begin{aligned} \nabla_{\mathbf{Z}_s}\mathcal{L}(\mathbf{Z}_s) = &2\theta(\mathbf{L}_i\mathbf{Z}_s + \mathbf{Z}_s\mathbf{L}_t) + 2\lambda(\mathbf{Z}_s - \mathbf{T}) \\ &+2\mu_2(\mathbf{Z}_s - \mathbf{UP}_s) - \mu_1\sum_{t=1}^{m}\mathbf{Y}_t\alpha_t(\mathbf{Y}_t\alpha_t)^{\mathrm{T}}\mathbf{Z}_s \end{aligned} \tag{21}$$

The above presents the gradients for solving $\mathbf{Z}_v$ and $\mathbf{Z}_s$. We adopt gradient descent method to solve them, and Armijo line search [6] is used to determine the step size δ.

4.3. Update for α_{t}

Given **P** and **Z**, solving α_t is the standard SVM optimization problem which can be effectively solved by SMO method [34]. By decomposing the quadratic programming problem into some solvable quadratic programming problems, SMO can obtain the solution quickly and effectively.

5. Experiments

To verify the effectiveness of our method OPSL in the task of image annotation, we perform several experiments on four image datasets. We firstly introduce four public image datasets and the compared methods used in our experiments. Then we analyze the influence of the parameters in our model. Last, comparison results with representative image annotation methods and baselines are provided to demonstrate the effectiveness of our method.

5.1. Datasets

The following four publicly available image annotation datasets are used to test our method.

1) Corel5k [13]. It contains 5,000 manually annotated images selected from a larger Corel CD set. Every image contains 3.5 tags on average. Following the standard setting, 4,500 samples are used as the training set and the rest of samples are used for testing.
2) ESP Game [41]. It consists of more than 20,000 images from many classes such as personal photos and logos. Every image is annotated with 4.6 tags on average. We adopt the standard experimental setting, so 18,689 samples are used as the training set and the remaining ones constitute the test set.
3) IAPRTC-12 [19]. It consists of 19,267 images of landscapes, people, sports, animals and other aspects of daily life. Following [10], 17,665 images are used for training and 1,962 images are used for testing.
4) NUS-WIDE [12]. It collects 55,615 images from Flickr. To reduce the noisy images and tags, we remove the images that annotated less than 3 tags and tags whose occurrence numbers are below 100. Then about 13,000 images are used as our dataset, where 10,000 samples are randomly selected for training and the remaining samples are used for testing.

5.2. Compared methods and evaluation metrics

To illustrate the effectiveness of our method, we compare it with several representative image annotation methods (1–5). Another two multi-view representation learning methods (6, 7) are introduced to verify the predictive ability of the learned image representation. In addition, as OPSL is extended from our previous method [48], we introduce several baselines (8–11) to demonstrate the effectiveness of each extended component.

1) FastTag [10]: A fast image tagging method that co-regularizes two simple linear mappings in a joint convex loss function.
2) NMF-KNN [24]: It combines nearest neighbour model with multi-view non-negative matrix factorization for image annotation.
3) LSR [28]: It is a tag completion method based on tag-specific and image-specific linear sparse reconstructions.
4) TMC [46]: It finds the optimal tag matrix which accords with both the visual and semantical similarity of images.
5) OGL [17]: It conducts image annotation by learning an optimal graph from multiple cues including different visual features and tags.
6) lrMVL [29]: It is a matrix completion based annotation method which seeks a common representation of multi-view data by utilizing low-rank multi-view learning.
7) MVLR [52]: It is a multi-view regression model with low-rank constraints, and it can provide a closed-form solution to the regression model.
8) OPSLP [48]: It is our previous method. We compare with it to illustrate the performance improvement achieved by this extended work.
9) OPSLP-V: It replaces softmax activation function in the objective function of OPSLP with equal weight learning strategy. This baseline method is to verify whether the softmax activation function can help better capturing the multi-view information.
10) OPSLP+AG: It is achieved by adding anchor graph method (Section 3.5) to OPSLP for projection function learning.
11) OPSLP+AG+SS: It is achieved by adding semantic subspace learning for OPSLP+AG (adopt problem (11) in the objective function). In this method, we do not add hypergraph regularization term and set $\theta = 0$.
12) OPSL: The method proposed in this paper, which is achieved by adding hypergraph regularization term in problem (11) to OPSLP+AG+SS, i.e., let $\theta > 0$.

To conduct image annotation, we annotate the five most relevant tags to each image. Then four standard performance measures are adopted for performance evaluation. Average precision (P), average recall (R) and F1-score (F1) [38] are computed for each test image, and the reported results are averaged across all test images. Moreover, similar to [17] and [46], Mean Average Precision (MAP), which is defined as the mean of average precision over each tag after retrieving relevant images, is adopted.

Table 2
P, R and F1 results on different datasets.

Methods	Corel5k			ESP Game			NUS			IAPRTC-12		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
FastTag	32.2	45.7	37.8	29.0	32.1	30.5	58.0	26.6	36.5	35.9	33.1	34.4
NMF-KNN	35.0	49.6	41.0	28.4	31.6	29.4	51.6	23.8	32.5	34.6	33.4	34.1
LSR	33.1	46.8	38.8	28.5	32.4	30.3	52.8	24.2	33.2	34.6	32.2	33.4
TMC	31.7	37.1	33.9	21.1	23.2	22.1	39.2	17.9	24.6	30.0	28.4	29.2
OGL	34.7	49.0	40.7	31.0	34.1	32.5	57.2	26.2	35.9	39.3	36.6	37.9
lrMVL	29.9	42.0	34.9	25.9	28.5	27.1	48.6	22.3	30.6	29.1	27.0	28.0
MVLR	25.9	37.2	30.5	24.5	27.2	25.8	37.7	17.3	23.7	28.1	26.7	27.4
OPSLP	37.0	52.1	43.3	32.3	35.6	33.9	**60.9**	**28.0**	**38.4**	39.8	37.4	38.6
OPSLP-V	36.0	50.7	42.1	31.3	34.5	32.8	59.1	27.1	37.2	39.3	36.9	38.1
OPSLP+AG	37.5	52.3	43.6	32.4	35.9	34.1	57.8	26.6	36.4	40.9	38.5	39.7
OPSLP+AG+SS	38.0	54.3	44.7	33.3	36.7	34.9	59.0	27.1	37.2	42.4	39.9	41.1
OPSL	**38.3**	**55.0**	**45.2**	**33.5**	**36.9**	**35.1**	59.2	27.2	37.3	**42.5**	**40.0**	**41.2**

5.3. Experimental settings

We extract several kinds of visual features to construct multiple views of images. For Corel5k, ESP Game and IAPRTC-12 datasets, we adopt seven kinds of visual features: HarrisHue, HarrisHueV3H1, DenseHue, DenseHueV3H1, HarrisSift, DenseSift, and Gist. These features can be obtained from [21]. For NUS dataset, we adopt six kinds of features: BoW based on SIFT, color moments, color correlation, color histogram, wavelet texture and edge direction histogram. Since some of the compared methods cannot directly leverage multi-view data, a preprocessing step is needed. We conduct PCA for each view and then concatenate the results as the new feature matrix.

To determine the parameters of each method, 1/10 of training data are used as the validation set. For the proposed method OPSL, the local structure information is modeled by constructing a 10-NN graph for each view. We set $\xi = 10^5$ to ensure that the orthogonality is satisfied. The number of anchor points g is set to 1000 for each dataset. The rest of the parameters in OPSL are determined on the validation set. The parameter sensitivity analysis are presented in Section 5.4. To effectively initialize the variables in the objective function, $\mathbf{Z}$ is initialized by conducting PCA on $\mathbf{U}$, then $\mathbf{P}$ and α_t are initialized by solving problem (14) and (12), respectively. In our experiments, we observe that the variables commonly remain unchanged after 20 iterations, so we set $MaxIter = 20$.

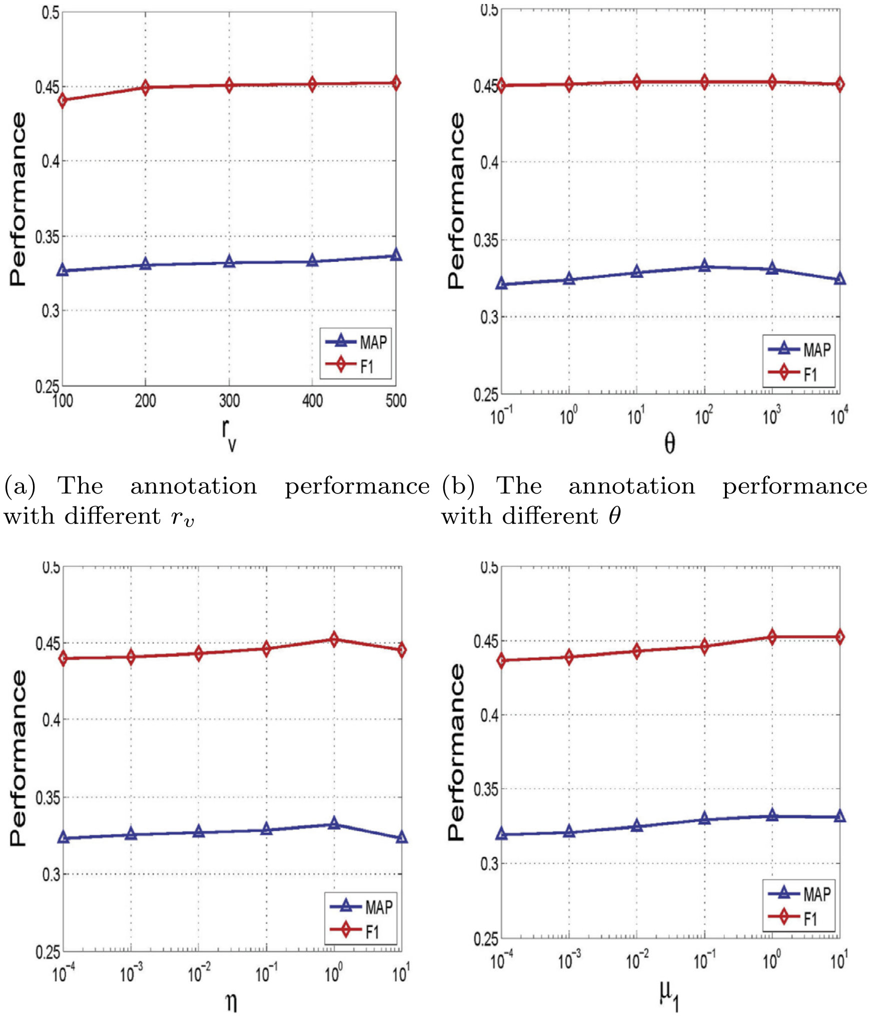
5.4. Parameter analysis

There are four important parameters r_v, θ, η and μ_1 in our method, and we study how the image annotation results change with different values of the parameters. We only present the results on the Corel5k dataset due to the space limit. To clearly illustrate the performance variations when changing parameters, we vary one parameter at a time while fixing the others. The default settings are: $r_v = 300,\ \theta = 100,\ \eta = 1$ and $\mu_1 = 1$. Fig. 2 shows the image annotation results when changing the values of different parameters.

r_v represents the dimensionality of visual subspace. From Fig. 2(a), we can see that when the dimensionality of visual subspace r_v is small ($r < 200$), the subspace cannot well preserve the multi-view information and the annotation performance is limited. When $r_v \geq 200$, the performance can be improved and it is stable with different values of r_v. θ controls the strength of hypergraph structure embedding. Although F1 score has little changes with different θ in Fig. 2(b), MAP can be effectively improved until reaching the peak at $\theta = 10^2$. This indicates that hypergraph is capable of capturing the high-order semantic correlations among images and tags, which is helpful to enhance the discriminative ability of learned subspace. η is the weight of semantic information preserving. From Fig. 2(c) we can observe that as η becomes larger, which means more semantic information is preserved, the performance is increased. However, when $\eta \geq 10$, the strength may become too strong, which leads to a slight drop of annotation performance. $\eta = 1$ generates the best annotation performance. μ_1 controls the weight for learning SVM classifiers. The larger of μ_1, SVM classifiers can transfer more guidance to the learned subspace. From Fig. 2(d), we can see that annotation performance increases when μ_1 becomes larger, which indicates that the supervision information of SVM can effectively guide the subspace learning. While the performance basically unchanged when $\mu_1 > 1$. So the proper range for this parameter is around $\mu_1 = 1$. Generally, the proposed method OPSL is not sensitive to the above parameters, and we can obtain competitive annotation performance in a wide range of parameters.

5.5. Results and analysis

The experimental results conducted on four datasets are summarized in Tables 2 and 3. Table 2 presents P, R and F1 results on each dataset, and MAP results are shown in Table 3. First, we compare the proposed method OPSL with the other image annotation methods except the baselines. We can observe that OPSL obtains competitive image annotation



(a) The annotation performance with different r_v

(b) The annotation performance with different θ

(c) The annotation performance with different η

(d) The annotation performance with different μ_1

Fig. 2. The annotation performance variance with different parameter settings on Corel5k dataset.

results compared to the other methods, which verifies the effectiveness of OPSL. Compared to the best results obtained by the other methods except the baselines, OPSL obtains performance improvement of 4.2% in F1 and 5.7% in MAP on Corel5k dataset, 2.6% in F1 and 3.1% in MAP on ESP Game dataset, 0.8% in F1 and 5.3% in MAP on NUS dataset, 3.3% in F1 and 1.8% in MAP on IAPRTC-12 dataset. From the experimental results presented in Tables 2 and 3, we have the following observations.

- NMF-KNN, OGL and OPSL are multi-view image annotation methods, and they generally achieve better annotation performance than single view image annotation methods (TMC and LSR). This is because multi-view learning methods can well exploit the complementary nature of different views, so they can obtain more comprehensive and accurate multi-view information for image annotation. Simply concatenating multi-view features cannot fully preserve multi-view information, so the image annotation performance of single view based methods are limited.
- The local manifold structure of image data can enhance the effectiveness of visual descriptions, which is beneficial to image annotation. LSR and NMF-KNN adopt knn strategies to learn the latent space for tag completion. OGL and OPSL model the underlying manifold structure of multi-view data by constructing nearest neighbor graphs. These methods generally achieve better performance than lrMVL and MVLR, which do not exploit the local structure information.
- With the help of semantic information, the discriminative ability of the learned multi-view representation can be improved. Semantic correlations of images are utilized in OGL and OPSL to guide the representation learning, which makes the two methods learn proper image representation for annotation task and obtain better performance.
- Compared to the other methods, OPSL achieves competitive image annotation performance because of several reasons. First, we use the semantic information to enhance the local manifold structure of multi-view data, which improves the

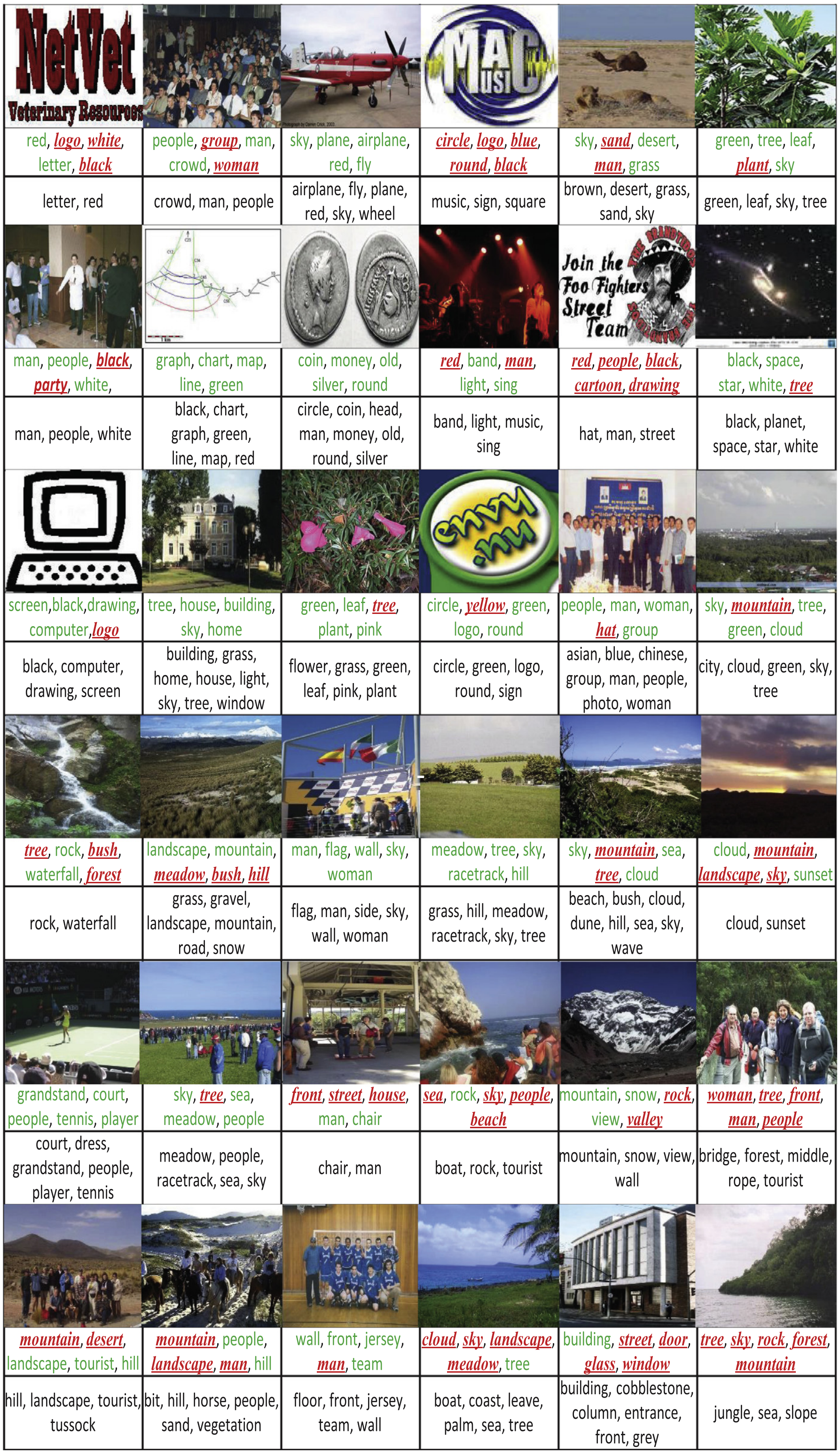

red, *logo*, *white*, letter, *black*
letter, red
people, *group*, man, crowd, *woman*
crowd, man, people
sky, plane, airplane, red, fly
airplane, fly, plane, red, sky, wheel
circle, *logo*, *blue*, *round*, *black*
music, sign, square
sky, *sand*, desert, *man*, grass
brown, desert, grass, sand, sky
green, tree, leaf, *plant*, sky
green, leaf, sky, tree
man, people, *black*, *party*, white,
man, people, white
graph, chart, map, line, green
black, chart, graph, green, line, map, red
coin, money, old, silver, round
circle, coin, head, man, money, old, round, silver
red, band, *man*, light, sing
band, light, music, sing
red, *people*, *black*, *cartoon*, *drawing*
hat, man, street
black, space, star, white, *tree*
black, planet, space, star, white
screen,black,drawing, computer,*logo*
black, computer, drawing, screen
tree, house, building, sky, home
building, grass, home, house, light, sky, tree, window
green, leaf, *tree*, plant, pink
flower, grass, green, leaf, pink, plant
circle, *yellow*, green, logo, round
circle, green, logo, round, sign
people, man, woman, *hat*, group
asian, blue, chinese, group, man, people, photo, woman
sky, *mountain*, tree, green, cloud
city, cloud, green, sky, tree
tree, rock, *bush*, waterfall, *forest*
rock, waterfall
landscape, mountain, *meadow*, *bush*, *hill*
grass, gravel, landscape, mountain, road, snow
man, flag, wall, sky, woman
flag, man, side, sky, wall, woman
meadow, tree, sky, racetrack, hill
grass, hill, meadow, racetrack, sky, tree
sky, *mountain*, sea, *tree*, cloud
beach, bush, cloud, dune, hill, sea, sky, wave
cloud, *mountain*, *landscape*, *sky*, sunset
cloud, sunset
grandstand, court, people, tennis, player
court, dress, grandstand, people, player, tennis
sky, *tree*, sea, meadow, people
meadow, people, racetrack, sea, sky
front, *street*, *house*, man, chair
chair, man
sea, rock, *sky*, *people*, *beach*
boat, rock, tourist
mountain, snow, *rock*, view, *valley*
mountain, snow, view, wall
woman, *tree*, *front*, *man*, *people*
bridge, forest, middle, rope, tourist
mountain, *desert*, landscape, tourist, hill
hill, landscape, tourist, tussock
mountain, people, *landscape*, *man*, hill
bit, hill, horse, people, sand, vegetation
wall, front, jersey, *man*, team
floor, front, jersey, team, wall
cloud, *sky*, *landscape*, *meadow*, tree
boat, coast, leave, palm, sea, tree
building, *street*, *door*, *glass*, *window*
building, cobblestone, column, entrance, front, grey
tree, *sky*, *rock*, *forest*, *mountain*
jungle, sea, slope

Fig. 3. Image annotation examples on ESP Game and ICPRTC-12 datasets. The first three rows are the examples from ESP Game dataset, and the last three rows are the examples from ICPRTC-12 datasets. The tags in black are the groundtruth given by the dataset. The tags in color are the predicted ones obtained by OPSL, where the green ones are those match with the groundtruth, and the red and italic ones are missing in the groundtruth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
MAP results on different datasets.

Methods	Corel5k	ESP Game	NUS	IAPR
FastTag	25.3	12.2	11.2	25.8
NMF-KNN	26.2	13.7	10.5	24.4
LSR	24.8	14.9	13.6	22.7
TMC	17.3	9.8	9.4	19.8
OGL	27.5	17.0	13.3	28.3
lrMVL	20.4	10.3	9.4	20.4
MVLR	16.9	9.5	7.9	19.2
OPSLP	29.6	16.1	14.5	29.2
OPSLP-V	28.8	15.3	13.9	28.6
OPSLP+AG	31.8	17.4	16.5	26.9
OPSLP+AG+SS	32.0	19.2	18.1	29.8
OPSL	**33.2**	**20.1**	**18.9**	**30.1**

accuracy of visual structure. By preserving both visual structure and semantic information in the learned subspace, the learned subspace becomes more discriminative and compact. Moreover, the trained SVM predictors are used to guide the subspace learning which makes the subspace appropriate for image annotation. Furthermore, our method performs multi-view representation and tag prediction simultaneously. We make the two tasks promote each other so that the image annotation results can be further enhanced.

Then the proposed method OPSL is compared to the baselines (OPSLP, OPSLP-V, OPSLP+AG, OPSLP+AG+SS) to illustrate the effectiveness of OPSL. Compared to our previous method OPSLP, OPSL obtains performance improvement of 1.9% in F1 and 3.6% in MAP on Corel5k dataset, 1.2% in F1 and 4.0% in MAP on ESP Game dataset, 2.6% in F1 and 0.9% in MAP on IAPRTC-12 dataset. Although the F1 score of OPSL is slightly less than OPSLP on NUS dataset, OPSL obtains better comprehensive results with a MAP score 4.4% higher than OPSLP. These performance improvements indicate that the extended components in OPSL is beneficial to improving annotation performance. The detailed reasons and analysis are listed as follows.

- OPSLP-V does not perform as well as OPSLP, indicating that the softmax activation function can effectively reduce the difference between each view and the learned subspace and make better use of multi-view complementary nature. OPSLP-V equally combines each view which cannot sufficiently capture the multi-view complementary information.
- By introducing anchor graph method for projection function learning, OPSLP+AG generally achieves better performance than OPSLP. The performance improvement is mainly because that image data are always nonlinear distributed, the linear projection function cannot well cope with the nonlinearity inherent in data. Anchor graph method can represent data by a set of anchor points and generate nonlinear data representation. So the projection function learned by OPSLP+AG achieves better generalization ability and provides more accurate mapping results for images. It should be noted that OPSLP+AG leads to F1 score on NUS and MAP on IAPRTC-12 slightly decrease, which further limits the performance of OPSLP+AG+SS and OPSL. This indicates that anchor graph method cannot guarantee improving the learning performance for all the situations, but it is still an effective method since it can improve image annotation performance in most cases.
- OPSLP+AG+SS obtains better performance than OPSLP+AG on all the datasets, which demonstrates that the proposed semantic subspace learning in Section 3.3.2 can effectively enhance the discriminative ability of image representation. As the learned semantic subspace contains tag information of images, tag dependency can be exploited for tag prediction. Moreover, by integrating visual and semantic subspace into a unified subspace, the semantic information can further complement the visual structure of images so that better annotation results can be achieved.
- With the help of the hypergraph regularization for image representation learning, OPSL performs better than OPSLP+AG+SS. Although the improvements of F1 score in Table 2 are limited, the performance improvements are obvious in terms of MAP (Table 3) on each dataset. Hypergraph captures high-order correlations among images and tags, therefore, conducting hypergraph regularization can further enrich the semantic information of images and make the learned subspace more compact and discriminative for annotation task.

Finally, we present some image annotation examples obtained by OPSL on ESP Game and ICPRTC-12 datasets. From Fig. 3, we can see that the proposed method accurately predicts image tags in most cases, which indicates the effectiveness of OPSL. Some of the prediction results can well match with the groundtruth image tags such as (row 1, col 3), (row 5, col 1) and (row 2, col 2), etc. While some annotation results fail to predict groundtruth tags such as (row 2, col 5), (row 5, col 6) and (row 6, col 6). By observing the wrong predicted cases, we find that some of them actually provides accurate prediction results. Specifically, for image (row 2, col 4), the predicted tags “red” and “man” accurately depict the content of that image. For image (row 2, col 5), the predicted tags “cartoon” and “black” accord with that image. For image (row 4, col 5), the predicted tags include “mountain” and “tree”, which accurately describes the content. So some of the wrong predictions are caused by the incomplete groundtruth tags. Since the visual content of an image is rich and diversified, it is difficult for people to provide perfect groudtruth tags for each image. Generally, our method OPSL can accurately annotate images according to their visual contents and achieve satisfying annotation results.

6. Conclusion

In this paper, we proposed OPSL, an optimal predictive subspace learning method designed to exploit correlations of multi-view representation and image annotation to jointly perform the two tasks, and derived an alternative optimization algorithm to obtain projection function and the final SVM classifier quickly and effectively. Data representations in the learned subspace become more discriminative and compact by preserving local manifold structure of multi-view data with hypergraph. Furthermore, anchor graph method is adopted to incorporate the nonlinearity into the projection function, which provides more accurate mapping results. Experiments are conducted on four image datasets and the results demonstrate that the proposed method achieves promising image annotation performance compared to the other methods. The experimental results also verify that each component of our method is reasonable and effective.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China: 61772494, 61332016, 61620106009, U1636214, 61650202, 61532006 and 61772083, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, in part by Youth Innovation Promotion Association CAS, in part by Director Foundation of Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia: 1TSM20180102.

References

[1] S. Agarwal, K. Branson, S. Belongie, Higher order learning with graphs, in: International Conference on Machine Learning, 2006, pp. 17–24.
[2] S.H. Amiri, M. Jamzad, Efficient multi-modal fusion on supergraph for scalable image annotation, Pattern Recognit. 48 (7) (2015) 2241–2253.
[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[4] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering., in: Advances in Neural Information Processing Systems (NIPS), vol. 14, 2001, pp. 585–591.
[5] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[6] D.P. Bertsekas, Nonlinear programming (1999).
[7] X. Cai, F. Nie, W. Cai, H. Huang, Heterogeneous image features integration via multi-modal semi-supervised learning model, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1737–1744.
[8] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 586–594.
[9] X. Cao, H. Zhang, X. Guo, S. Liu, D. Meng, SLED: Semantic label embedding dictionary representation for multilabel image annotation, IEEE Trans. Image Process. 24 (9) (2015) 2746–2759.
[10] M. Chen, A. Zheng, K. Weinberger, Fast image tagging, in: International Conference on Machine Learning, 2013, pp. 1274–1282.
[11] J. Cheng, C. Leng, P. Li, M. Wang, H. Lu, Semi-supervised multi-graph hashing for scalable similarity search, Comput. Vision Image Understand. 124 (1) (2014) 12–21.
[12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: ACM International Conference on Image and Video Retrieval, 2009, p. 48.
[13] P. Duygulu, K. Barnard, J.F. de Freitas, D.A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: European Conference on Computer Vision, 2002, pp. 97–112.
[14] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
[15] Z. Feng, S. Feng, R. Jin, A.K. Jain, Image tag completion by noisy matrix recovery, in: European Conference on Computer Vision, 2014, pp. 424–438.
[16] Z. Feng, R. Jin, A.K. Jain, Large-scale image annotation by efficient and robust kernel metric learning, in: IEEE International Conference on Computer Vision, 2013, pp. 1609–1616.
[17] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, H.T. Shen, Optimal graph learning with partial tags and multiple features for image and video annotation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4371–4379.
[18] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, arXiv:1312.4894 (2013).
[19] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The iapr tc12 benchmark: a new evaluation resource for visual information systems, in: In International Workshop OntoImage, 2006, pp. 13–23.
[20] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, Y.Y. Tang, Group sparse multiview patch alignment framework with view consistency for image classification, IEEE Trans. Image Process. 23 (7) (2014) 3126–3137.
[21] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation, in: International Conference on Computer Vision, 2009, pp. 309–316.
[22] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, X. Wu, Image annotation by multiple-instance learning with discriminative feature mapping and selection, IEEE Trans. Cybernet. 44 (5) (2014) 669–680.
[23] X. Jing, F. Wu, Z. Li, R. Hu, D. Zhang, Multi-label dictionary learning for image annotation, IEEE Trans. Image Process. 25 (6) (2016) 2712–2725.
[24] M.M. Kalayeh, H. Idrees, M. Shah, Nmf-knn: image annotation using weighted multi-view non-negative matrix factorization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 184–191.
[25] J. Kim, R. Monteiro, H. Park, Group sparsity in nonnegative matrix factorization., in: SIAM International Conference on Data Mining (SDM), 2012, pp. 851–862.
[26] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2001, pp. 556–562.
[27] Y.Y. Lin, T.L. Liu, C.S. Fuh, Multiple kernel learning for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1147–1160.
[28] Z. Lin, G. Ding, M. Hu, J. Wang, X. Ye, Image tag completion via image-specific and tag-specific linear sparse reconstructions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1618–1625.
[29] M. Liu, Y. Luo, D. Tao, C. Xu, Y. Wen, Low-rank multi-view learning in matrix completion for multi-label image classification, in: AAAI Conference on Artificial Intelligence, 2015.
[30] W. Liu, J. He, S.F. Chang, Large graph construction for scalable semi-supervised learning, in: International Conference on Machine Learning, 2010, pp. 679–686.
[31] W. Liu, D. Tao, J. Cheng, Y. Tang, Multiview hessian discriminative sparse coding for image annotation, Comput. Vision Image Understand. 118 (2014) 50–60.

[32] B. Long, S.Y. Philip, Z.M. Zhang, A general model for multiple view unsupervised learning, in: SIAM International Conference on Data Mining (SDM), 2008, pp. 822–833.

[33] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: European Conference on Computer Vision, 2008, pp. 316–329.

[34] J. Platt, et al., Fast training of support vector machines using sequential minimal optimization, Adv. Kernel Methods Support Vector Learn. 3 (1999).

[35] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[36] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2002.

[37] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (5500) (2000) 2268–2269.

[38] A. Singhal, Modern information retrieval: a brief overview, IEEE Data Eng. Bull. 24 (4) (2001) 35–43.

[39] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[40] Y. Verma, C.V. Jawahar, Image annotation by propagating labels from semantic neighbourhoods, Int. J. Comput. Vis. 121 (1) (2017) 126–148.

[41] L. Von Ahn, L. Dabbish, Labeling images with a computer game, in: SIGCHI Conference on Human Factors in Computing Systems, 2004, pp. 319–326.

[42] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: a unified framework for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2285–2294.

[43] M. Wang, W. Fu, S. Hao, D. Tao, Scalable semi-supervised learning by efficient anchor graph regularization, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1864–1877.

[44] Q. Wang, B. Shen, S. Wang, L. Li, L. Si, Binary codes embedding for fast image tagging with incomplete labels, in: European Conference on Computer Vision, 2014, pp. 425–439.

[45] Z. Wang, T. Chen, G. Li, R. Xu, L. Lin, Multi-label image recognition by recurrently discovering attentional regions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 464–472.

[46] L. Wu, R. Jin, A.K. Jain, Tag completion for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 716–727.

[47] T. Xia, D. Tao, T. Mei, Y. Zhang, Multiview spectral embedding, IEEE Trans. Systems, Man Cybernet Part B 40 (6) (2010) 1438–1446.

[48] Z. Xue, G. Li, Q. Huang, Joint multi-view representation learning and image tagging, in: AAAI Conference on Artificial Intelligence, 2016, pp. 1366–1372.

[49] Z. Xue, G. Li, S. Wang, W. Zhang, Q. Huang, Bi-level multi-view latent space learning, IEEE Trans. Circuits Syst. Video Technol. PP (99) (2016) 1.

[50] J. Yu, Y. Rui, Y.Y. Tang, D. Tao, High-order distance-based multiview stochastic learning in image classification, IEEE Trans. Cybernet. 44 (12) (2014) 2431–2442.

[51] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: Advances in Neural Information Processing Systems, 2004, pp. 1601–1608.

[52] S. Zheng, X. Cai, C. Ding, F. Nie, H. Huang, A closed form solution to multi-view low-rank regression, in: AAAI Conference on Artificial Intelligence, 2015.

[53] D. Zhou, J. Huang, B. Scholkopf, Learning with hypergraphs: clustering, classification, and embedding, in: Advances in Neural Information Processing Systems, 2006, pp. 1601–1608.

[54] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.