# Joint Multi-View Representation Learning and Image Tagging

**Zhe Xue**[1,2], **Guorong Li**[1,2,*], **Qingming Huang**[1,2,3,*]
[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
[2] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[3] Key Lab. of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing 100190, China
xuezhe10@mails.ucas.ac.cn, {liguorong, qmhuang}@ucas.ac.cn

## Abstract

Automatic image annotation is an important problem in several machine learning applications such as image search. Since there exists a semantic gap between low-level image features and high-level semantics, the description ability of image representation can largely affect annotation results. In fact, image representation learning and image tagging are two closely related tasks. A proper image representation can achieve better image annotation results, and image tags can be treated as guidance to learn more effective image representation. In this paper, we present an optimal predictive subspace learning method which jointly conducts multi-view representation learning and image tagging. The two tasks can promote each other and the annotation performance can be further improved. To make the subspace to be more compact and discriminative, both visual structure and semantic information are exploited during learning. Moreover, we introduce powerful predictors (SVM) for image tagging to achieve better annotation performance. Experiments on standard image annotation datasets demonstrate the advantages of our method over the existing image annotation methods.

With the rapid development of social network and digital equipment, large amounts of unlabeled or weakly labeled images are shared on websites such as Flicker and Facebook. To quickly access the interesting content from numerous images, an effective retrieval mechanism is required. Although image search has been studied for years, the search engines still mainly rely on textual queries instead of images. Since automatic image annotation can assign relevant tags to unlabeled images enriching their textual description, it is becoming an essential tool for searchable image databases and attracting more and more research interest.

A variety of methods have been developed for image annotation (Makadia, Pavlovic, and Kumar 2008; Guillaumin et al. 2009; Lin et al. 2013; Gao et al. 2014). Some methods adopt nearest neighbour method to annotate images by propagating tags from their nearest neighbours (Makadia, Pavlovic, and Kumar 2008; Guillaumin et al. 2009). Some methods adopt linear predictors or linear reconstruction for image tagging (Chen, Zheng, and Weinberger 2013; Lin et al. 2013). Moreover, matrix completion is adopted in TMC (Wu, Jin, and Jain 2013) where the tag-image relation is represented by a tag matrix. By searching the optimal tag matrix that is consistent with both the textual and the visual similarity, the images are tagged by relevant tags. A critical factor affecting the annotation performance is the description abilities of the visual features. It is easier to annotate images by using more powerful image descriptors.

Compared with single feature based representation, images can be described by multiple features such as SIFT, HOG and LBP. Different features capture different aspects of visual characteristics, and each type of feature can be treated as a view for describing an instance. Multiple views can complement with each other and generate more powerful descriptions for learning tasks. Thus, some image annotation methods are proposed based on multi-view data (Kalayeh, Idrees, and Shah 2014; Gao et al. 2014). NMF-KNN (Kalayeh, Idrees, and Shah 2014) first learns a query-specific nearest neighbour model and then predict tags by the matrix product of the learned coefficients with the basis of tags. OGL (Gao et al. 2014) is a semi-supervised method that learns an optimal graph from multiple views, which can accurately represent the relationships among image data.

Nevertheless, most of the methods predict tags based on low-level visual descriptors, which may contain noise and redundant features. In addition, since there exists a semantic gap between low-level visual features and high-level semantics, it is difficult to derive the true semantic information directly utilizing the low-level features. A practical solution is to find an appropriate representation to enhance the performance of learning tasks. Subspace learning methods (Belhumeur, Hespanha, and Kriegman 1997; 1997; Roweis and Saul 2000; Belkin and Niyogi 2001) are proposed to obtain new representations from original features and achieves promising performance. However, most of these methods ignore the semantic information associated with images such as tags and the learning process are relatively independent with the follow-up tasks. So the obtained representation are not optimal for the learning tasks and their performance is limited.

In fact, representation learning and image tagging are two closely correlated works. On one hand, an appropriate image representation can well reflect the semantic relations of images and facilitate tag prediction. On the other hand, image tags are high-level semantic information and can guide the representation learning. Thus, joint conducting the two

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tasks can make them benefit each other and achieve the overall optimal performance. In this paper, we propose an optimal predictive subspace learning (OPSL) method which integrates multi-view representation learning and image tagging into a unified learning framework. Instead of predicting tags from original feature spaces, our method conduct image tagging from the optimal predictive subspace, where images can be appropriately represented for tag prediction. We jointly learn the subspace and tag predictors to make the two tasks promote each other so that the annotation performance can be further improved. To guarantee that the learned subspace achieves good predictive ability, we adopt the following methods. First, the intrinsic geometric structure of multi-view data is preserved in the subspace, and the semantic information is utilized to enhance the structure preserving. Second, we adopt softmax activation function for subspace learning which makes the learned subspace effectively minimize the disagreement between views and better capture the complementary information. Furthermore, we introduce powerful predictors (support vector machines, SVM) for tag prediction. In turn, the trained SVM is used to guide the subspace learning. Thus, the discriminative abilities of the learned subspace and SVM predictors can be mutually improved. Experiments are conducted on standard image annotation datasets and the results show that our method achieves promising image annotation performance compared with the existing methods.

The main contributions are summarised as follows.

- We propose a unified subspace learning framework to simultaneously learn the proper image representation and tag predictors. The correlations of the two tasks are well explored so that they can promote each other and generate promising annotation performance.
- Both visual structure and semantic information are preserved during subspace learning, which makes the learned representation more compact and discriminative. We adopt softmax activation function for multi-view structure preserving which can capture the difference between views and sufficiently encode the multi-view complementary information.
- Instead of adopting linear regression predictors which are commonly used in the existing image tagging works, we introduce SVM for its powerful classification ability to achieve better annotation performance. Moreover, the trained SVM is also used to guide the subspace learning to make the learned subspace more suitable for annotation task and yields accurate prediction.

## Related Work

Image annotation is to assign unlabelled images with semantically related tags. Some annotation methods (Makadia, Pavlovic, and Kumar 2008; Guillaumin et al. 2009; Kalayeh, Idrees, and Shah 2014) adopt nearest neighbour mechanism to propagate tags to unlabeled images. TagProp (Guillaumin et al. 2009) learns a discriminative metric which combines a collection of image similarities and then predict tags using a nearest-neighbor model. In addition, matrix completion are applied to image annotation in (Wu, Jin, and Jain 2013; Feng et al. 2014). Tag matrix is considered to be incomplete and containing noise. The visual similarities of images and the low-rank property of the tag matrix are commonly utilized to recover the missing tags. To infer high-level semantic information from low-level visual features, an important issue is the description ability of visual features. Instead of using single feature (Wu, Jin, and Jain 2013; Wang et al. 2014), some methods (Lin et al. 2013; Chen, Zheng, and Weinberger 2013; Kalayeh, Idrees, and Shah 2014; Gao et al. 2014) exploit multiple features for image tagging, which obtains competitive tagging performance due to multiple features can generate more complete descriptions. However, most of them use task-independent techniques such as PCA (Belhumeur, Hespanha, and Kriegman 1997) to represent multi-view data, which may fail to obtain the optimal representation for the annotation task.

To obtain proper representation from original features, many subspace learning methods have been proposed. Some linear methods are developed such as PCA, LDA (Belhumeur, Hespanha, and Kriegman 1997) and NMF (Lee and Seung 2001). Besides, some methods utilize the local and manifold property to obtain the representation such as ISOMAP (Tenenbaum, De Silva, and Langford 2000), LE (Belkin and Niyogi 2001), LLE (Roweis and Saul 2000). To cope with the multi-view data, several multi-view learning methods are proposed (Long, Philip, and Zhang 2008; Xia et al. 2010; Lin, Liu, and Fuh 2011). A general spectral embedding framework is proposed for multi-view dimensionality reduction in (Long, Philip, and Zhang 2008). In addition, multiple kernel learning is used for multi-view dimensionality reduction in (Lin, Liu, and Fuh 2011), which provides convenience of using multiple types of image features. The obtained multi-view representation can encode the information of each view and provide a more complete and accurate description for learning tasks.

## The Proposed Method

### Preliminary

For an arbitrary matrix $A$, its $(i,j)$-th entry and $i$-th row are denoted by $A_{ij}$ and $A_{i\cdot}$ respectively. $Tr[A]$ is the trace of $A$. For $A \in \mathbb{R}^{n\times m}$, the Frobenius norm is $||A||_F$, and $\ell_{2,1}$-norm is defined as

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

Given multi-view data consisting of $n$ samples with $H$ views, they can be denoted by a set of matrices $\mathcal{X} = \{X^{(i)} \in \mathbb{R}^{n\times d_i}\}_{i=1}^{H}$, where $d_i$ is the dimensionality of the $i$-th view. For image data, each view represents a kind of visual feature. Every training image is assigned with $m$-dimensional binary-valued tag vector $\{t_i\}_{i=1}^{n}, t_i \in \{0,1\}^m$. Let $T = [t_1, ..., t_n]^T$ be the tag matrix of size $n \times m$. We aim to learn the subspace $Z \in \mathbb{R}^{n\times r}$ from multi-view data matrices, where $r$ is the dimensionality of the learned subspace.

Many works have shown that data are more likely to reside on a low-dimensional submanifold of the ambient space (Roweis and Saul 2000; Seung and Lee 2000;

Belkin, Niyogi, and Sindhwani 2006), and exploiting the intrinsic manifold structure of data can enhance the discriminating power of the learned subspace. The geometric structure can be effectively modeled through a nearest neighbor graph. Given image data $\{x_i\}_{i=1}^n$ with $H$ views, we construct $k$-nearest neighbor graph $G^h$ for each view $h = 1, ..., H$ with affinity matrix $W^h \in \mathbb{R}^{n\times n}$. Then each graph encodes the local geometric structure information of the corresponding view and can be used for subspace learning tasks. The graph is constructed as follows,

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

where $\sigma$ is the bandwidth parameter and can be determined by self-tuning method (Zelnik-Manor and Perona 2004).

Image annotation methods commonly adopt classifiers to predict tags for unlabeled images. Support vector machines (SVM) have emerged as a powerful tool for classification tasks and achieved satisfied learning performance (Schölkopf and Smola 2002). In the case of a binary classification problem with $n$ training examples, let $K$ be the kernel matrix and let $y \in \{-1, +1\}^n$ be the vector of labels, with $Y = diag(y)$. The dual SVM solves the following optimization problem:

$$\max_{\{0\le\alpha\le C, \alpha^T y=0\}} \alpha^{\mathrm{T}}\mathbf{1} - 0.5Tr\big(K(Y\alpha)(Y\alpha)^{\mathrm{T}}\big), \tag{3}$$

where $\alpha \in \mathbb{R}^n$ are Lagrange multipliers, $\mathbf{1} = [1, ..., 1]^{\mathrm{T}} \in \mathbb{R}^n$ and $C$ is the misclassification penalty. This is also the classic 1-norm soft margin SVM problem.

## Semantic Information Guided Multi-View Structure Preserving

To guarantee that the learned subspace achieves good predictive ability for image tagging, we expect it to satisfy the following two properties. First, the subspace should be locally smooth, i.e., the local geometric structure of images in the original visual space should be preserved in the learned subspace. Second, different views generate different descriptions about the same sample, and the difference between multiple views make them complement with each other. We expect to learn a universal subspace from multiple views which minimizes the disagreement between them, so that the multi-view complementary information can be sufficiently encoded in the learned subspace. Third, the semantic information of images should be exploited for subspace learning and the learned subspace should be discriminative to well predict image tags.

Geometric structures of multiple views are modeled by several $k$-nearest neighbour graphs $\{W^h\}_{h=1}^H$. We use $W^s = T^{\mathrm{T}}T$ to model the semantic structure of images. The more tags that the two images share, the more similar are the two images. Inspired by the success of self-representation which has been widely used in subspace learning (Elhamifar and Vidal 2013; Cao et al. 2015), we encode the visual and semantic structures into subspace $Z$ by the following formulation,

$$\begin{aligned} &\min_{Z} \textstyle\sum_{h=1}^H ||Z - (W^h \odot W^s)Z||_F^2 + \eta||T - ZZ^{\mathrm{T}}T||_F^2 \\ &s.t.\ Z^{\mathrm{T}}Z = I \end{aligned} \tag{4}$$

where $\odot$ denotes Hadamard product. We utilize semantic structure to enhance the visual structure, and the visual nearest neighbours who share the similar tags are treated as the true neighbours. Semantic correlations are used to filter out inaccurate visual descriptions so that the visual structure are made to be more accurate and reliable. The second term is to encode semantic information into the learned subspace, where we can see that the images possessing similar tags are also made to be similar in the learned subspace $Z$. $\eta$ is the weight parameter to control the strength of semantic information embedding. The orthogonal constraint is to avoid the trivial solution.

Nevertheless, all the views are simply assigned with equal weights during multi-view structure preserving in problem (4), which may not be the optimal setting. The differences of multi-view data make some views generate larger disagreements and higher costs. We expect to find a universal subspace that accommodates each view well and captures the complementary information of multi-view data. To address this problem, we consider to minimize the difference between the learned subspace and the most disagreement view (the one generates the highest embedding cost). Thus, the overall disagreements between the learned subspace and each view can be effectively reduced and the complementary information of multi-view data can be fully preserved in the subspace. We adopt the softmax activation function to approximately find the most disagreement view, and the formulation is revised as follows,

$$\min_{Z}\ f(Z), \quad s.t.\ Z^{\mathrm{T}}Z = I \tag{5}$$

and

$$\begin{aligned} f(Z) =& \frac{1}{\gamma}\log\Big\{\sum_{h=1}^{H}\exp\big[\gamma||Z - (W^h \odot W^s)Z||_F^2\big]\Big\} \\ &+ \eta||T - ZZ^{\mathrm{T}}T||_F^2, \end{aligned} \tag{6}$$

where $\gamma$ is the smooth parameter to control the precision of approximation.

## Tag Predictors Learning

After obtaining the subspace $Z$, we predict the tags in the learned subspace. We introduce SVM to annotate images for its powerful classification ability. To make the subspace more discriminative, we also use the trained SVM to guide the subspace learning. In the dual SVM formulation, we adopt linear kernel $K = ZZ^{\mathrm{T}}$ for its simplicity and effectiveness. Then the objective function of predictors learning can be formulated as follows,

$$\begin{aligned} &\min_{Z}\max_{\alpha_t} g(Z, \alpha_t), \\ &s.t.\ \alpha_t^{\mathrm{T}} y_t = 0, 0 \le \alpha_t \le C,\ t = 1, ..., m \end{aligned} \tag{7}$$

and

$$g(Z,\alpha_t)=\sum_{t=1}^{m}\Big[\alpha_t^{\mathrm{T}}\mathbf{1}-0.5Tr\left(ZZ^{\mathrm{T}}Y_t\alpha_t(Y_t\alpha_t)^{\mathrm{T}}\right)\Big]. \tag{8}$$

It should be noted that each tag is trained with a SVM predictor to separate it from the other tags and totally $m$ SVM predictors are trained. $y_t \in \{-1,+1\}^n$ and $Y_t = diag(y_t)$ are the label vector and matrix for training SVM of tag $t$ respectively, and $\alpha_t$ are the learned Lagrange multipliers. Given the subspace $Z$, SVM is trained based on this representation and the learning results $\alpha_t$ can be obtained. In turn, the trained SVM is also used to guide the subspace learning to make it appropriate for tag prediction and be more discriminative.

### Projection Function Learning

To cope with the out-of-sample problem, we learn the projection function to project the unlabeled multi-view data to the subspace and then conduct image annotation by the trained SVM. We first construct matrix $X = [X^1, X^2, ..., X^H]$, $X \in \mathbb{R}^{n\times d}$ and $d = d_1 + ... + d_H$, which concatenates the feature matrix of each view. Then we use matrix $P \in \mathbb{R}^{d\times r}$ to project multi-view data into the subspace $Z$. Taking account for the noise and redundant features contained in the multi-view data, we adopt $\ell_{2,1}$-norm to regularize the the projection function, which can shrink some rows to zeros to select effective dimensions. The projection function learning is defined as

$$\min_{Z,P} h(Z,P), \tag{9}$$

and

$$h(Z,P)=||XP-Z||_F^2+\beta||P||_{2,1}, \tag{10}$$

where $\beta$ is the parameter to control the strength of the regularization term.

### Unified Objective Function

The ultimate optimization objective is formulated by integrating multi-view subspace learning and image annotation into a unified framework, which aims to make the two tasks promote each other and achieve the overall optimal performance. So the problem is formulated as follows,

$$\begin{aligned} &O(Z,P,\alpha_t)=\min_{Z,P}\max_{\alpha_t}\big(f(Z)+\mu_1 g(Z,\alpha_t)+\mu_2 h(Z,P)\big),\\ &s.t.\ Z^TZ=I, \alpha_t^{\mathrm{T}}y_t=0, 0\le\alpha_t\le C,\ t=1,...,m \end{aligned} \tag{11}$$

where $\mu_1$ and $\mu_2$ are two parameters to control the weights of the corresponding subproblems.

## Optimization Algorithm

Apparently, problem (11) is not convex over all variables $Z$, $P$ and $\alpha_t$ simultaneously, so we derive an iteration optimization algorithm to solve it. In each iteration, only one variable is updated while the others remain unchanged. The optimization process is summarized in Algorithm 1.

### Update for P

In order to solve $P$, we keep the parts which are related to $P$ from $O(Z,P,\alpha_t)$ and define

$$\mathcal{L}(P)=||XP-Z||_F^2+\beta||P||_{2,1}. \tag{12}$$

Setting the gradient $\nabla_P\mathcal{L}(P)=0$, we have

$$\begin{aligned} \nabla_P\mathcal{L}(P)&=2X^{\mathrm{T}}(XP-Z)+2\beta DP=0\\ &\Rightarrow P=\left(X^{\mathrm{T}}X+\beta D\right)^{-1}X^{\mathrm{T}}Z, \end{aligned} \tag{13}$$

where $D$ is a diagonal matrix with elements $D_{ii}=\frac{1}{2||P_{i\cdot}||_2}$

### Update for Z

There is an orthogonal constraint for learning $Z$ and it is difficult to solve this problem exactly. So we adopt an approximation procedure and convert the original constrained optimization problem into an unconstrained optimization problem by using a penalty method. Keep the parts which are related to $Z$ from $O(Z,P,\alpha_t)$ and we obtain

$$\begin{aligned} \mathcal{L}(Z)&=\tfrac{1}{\gamma}\log\Big\{\sum_{h=1}^{H}\exp\big[\gamma||Z-(W^h\odot W^s)Z||_F^2\big]\Big\}\\ &+\eta||T-ZZ^{\mathrm{T}}T||_F^2-0.5\mu_1\sum_{t=1}^{m}Tr\big(ZZ^{\mathrm{T}}Y_t\alpha_t(Y_t\alpha_t)^{\mathrm{T}}\big)\\ &+\mu_2||XP-Z||_F^2+\lambda||Z^{\mathrm{T}}Z-I||_F^2, \end{aligned} \tag{14}$$

where $\lambda > 0$ is a parameter to control the orthogonality condition. It should be large to ensure the orthogonality is satisfied. Then the gradient is computed as

$$\begin{aligned} \nabla_Z\mathcal{L}(Z)&=\tfrac{1}{\sum_{i=1}^{H}\Delta_i}\big[\sum_{j=1}^{H}\Delta_j\big(2Z-4A^jZ+2A^{j^{\mathrm{T}}}A^jZ\big)\big]\\ &+4\eta(TT^{\mathrm{T}}ZZ^{\mathrm{T}}Z-TT^{\mathrm{T}}Z)+4\lambda(ZZ^{\mathrm{T}}Z-Z)\\ &-\mu_1\sum_{t=1}^{m}Y_t\alpha_t(Y_t\alpha_t)^{\mathrm{T}}Z+2\mu_2\left(Z-XP\right) \end{aligned} \tag{15}$$

where $\Delta_j = \exp\big[\gamma||Z-(W^j\odot W^s)Z||_F^2\big]$ and $A^j = W^j\odot W^s$. After obtaining the gradient, we adopt gradient descent method to solve $Z$. The step size $\delta$ is determined by Armijo linesearch (Bertsekas 1999).

### Update for $\alpha_t$

By fixing $Z$ and $P$, learning $\alpha_t$ is the classic SVM optimization problem which can be effectively solved by several methods such as SMO (Platt and others 1999). SMO can break this quadratic programming (QP) problem into a series of smallest possible QP problems and obtain the solution effectively and quickly.

## Experiments

### Datasets

We adopt three publicly available image tagging datasets that have been widely used in previous works.

**Corel5k**. It (Duygulu et al. 2002) contains $5,000$ manually annotated images collected from the larger Corel CD set and each image is annotated with $3.5$ tags on average.

**Algorithm 1:** The algorithm of OPSL

**Input**: Multi-view matrices $\mathcal{X} = \{\mathbf{X}^{(i)}\}_{i=1}^{H}$, visual and semantic structure graphs $\{W^h\}_{h=1}^{H}$ and $W^s$, tag matrix $T$, SVM label matrices $\{Y_t\}_{t=1}^{m}$ and parameters: $\mu_1$, $\mu_2$, $\gamma$, $\eta$, $\beta$, $\lambda$, $r$, $C$

Initialize $Z$, $P$ and $\{\alpha_t\}_{t=1}^{m}$

**for** $iter = 1$ *to* $MaxIter$ **do**

  Update $Z$ by $Z \leftarrow Z - \delta \nabla_Z \mathcal{L}(Z)$;

  Update SVM by SMO method;

  Update $P$ by equation (13);

  Update $D$ by $D \leftarrow \begin{bmatrix} \frac{1}{2||P_{1\cdot}||_2} & & \\ & \ldots & \\ & & \frac{1}{2||P_{d\cdot}||_2} \end{bmatrix}$;

**end**

**Output**: Projection function $P$ and SVM Lagrange multipliers $\{\alpha_t\}_{t=1}^{m}$.

**ESP Game**. It (Von Ahn and Dabbish 2004) contains about $20,000$ images from several varieties such as logos and personal photos and each image contains 4.6 tags on average.

**NUS-WIDE**. It (Chua et al. 2009) is a web image dataset which includes $55,615$ images and $5,018$ tags from Flickr. To reduce noisy tags and data, we removed tags whose occurrence numbers are below $100$ and images that contain less than 3 tags. Then we obtained around $13,000$ images, where $10,000$ of them are randomly sampled as training set and the rest are used as testing set.

## Compared Methods and Evaluation Metrics

To demonstrate the effectiveness of OPSL, we compare it with several representative image tagging methods, such as FastTag (Chen, Zheng, and Weinberger 2013), LSR (Lin et al. 2013), TMC (Wu, Jin, and Jain 2013), NMF-KNN (Kalayeh, Idrees, and Shah 2014) and OGL (Gao et al. 2014). To indicate the predictive ability of the learned multi-view subspace, we also compare our method with two multi-view learning methods lrMVL (Liu et al. 2015) and MVLR (Zheng et al. 2015). The two methods also learn new subspaces from original multi-view data and then predict labels based on the new learned representation. Furthermore, we conduct OPSL with equal weight learning manner which adopts function (4) instead of $f(Z)$ in objective function (11) and denote it by OPSL-V. This setting is to validate wether the adopted softmax activation function can help better capturing the multi-view information.

During image tagging, each image is annotated with the five most relevant keywords as in (Chen, Zheng, and Weinberger 2013). Then we evaluate our method with standard performance measures. Average precision (P), average recall (R) and F1-score (F1) (Singhal 2001) are computed for each test image by comparing annotation results with ground truth. The reported results are averaged across all test images. In addition, we adopt Mean Average Precision (MAP) as in (Wu, Jin, and Jain 2013; Gao et al. 2014), which can be calculated by checking the correctness of retrieved images.

## Experimental Setting

In our experiments, different kinds of visual features are used to construct different views of image data. For Corel5k and ESP Game datasets, we adopt seven kinds of visual features: Gist, HarrisSift, HarrisHue, HarrisHueV3H1, DenseSift, DenseHue and DenseHueV3H1 which are provided by (Guillaumin et al. 2009). For NUS dataset, six kinds of features are adopted: color histogram, color correlation, color moments, edge direction histogram, wavelet texture and BoW based on SIFT. As some of the compared methods cannot directly utilize multi-view data, we adopt PCA to separately perform dimensionality reduction for each view and then concatenate them into a new merged feature matrix.

Around $1/10$ of the training set are taken as the validate set for parameter tuning. Parameters in the compared methods are set as suggested in their works. For our method, we construct a 10-NN graph for each view to model the visual local structure, and Euclidean distance is used for finding the neighbors. To guarantee the orthogonality is satisfied, we fix $\lambda = 10^5$ in our experiments. The other parameters in our method are tuned on the validation set to determine their values. We will present the detailed parameter sensitivity analysis on Corel5k dataset in the following part. To obtain effective initial value, we first initialize $Z$ by conducting PCA on $X$, then initialize $P$ and $\alpha_t$ by solving problem (9) and (7) respectively. In experiments, we find the optimization variables generally remain unchanged after 20 iterations, so $MaxIter$ is set to 20.

## Results and Analysis

Image annotation performance on three datasets are shown in Table 1. It is obvious that the proposed method OPSL achieves promising image tagging performance compared with the other methods, which demonstrates the effectiveness of our method. OPSL achieves improvement of $2.3\%$ in F1 and $2.1\%$ in MAP for Corel5k dataset and $1.9\%$ in F1 and $0.9\%$ in MAP for NUS dataset. Although our method does not obtain the best MAP score in ESP Game dataset, it achieves the highest F1 score by improving $1.4\%$ compared with the other methods.

Throughout the experiments, we reveal several interesting points. First, multi-view based annotation methods (NMF-KNN, OGL and OPSL) generally obtain better annotation performance than the single view based methods. Multi-view methods considers the difference and complementarity between views, so they can obtain more accurate representation for image tagging. Second, local structure of images can enhance the accuracy of visual descriptions, which is very helpful for improving the annotation performance. NMF-KNN, LSR, OGL and OPSL explore the local structure of multi-view data generally achieve better performance than lrMVL and MVLR which do not consider the local information. Third, exploiting semantic information can make the learned representation more discriminative. Both OGL and OPSL utilize semantic correlations of images to guide the representation learning, which makes them obtain more appropriate representation for image annotation and achieve

Table 1: Image Annotation Comparison on Different Datasets.

| Method | Corel5k | | | | ESP Game | | | | NUS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | MAP | P | R | F1 | MAP | P | R | F1 | MAP |
| FastTag | 32.2 | 45.7 | 37.8 | 25.3 | 29.0 | 32.1 | 30.5 | 12.2 | 58.0 | 26.6 | 36.5 | 11.2 |
| NMF-KNN | 35.0 | 49.6 | 41.0 | 26.2 | 28.4 | 31.6 | 29.4 | 13.7 | 51.6 | 23.8 | 32.5 | 10.5 |
| LSR | 33.1 | 46.8 | 38.8 | 24.8 | 28.5 | 32.4 | 30.3 | 14.9 | 52.8 | 24.2 | 33.2 | 13.6 |
| TMC | 31.7 | 37.1 | 33.9 | 17.3 | 21.1 | 23.2 | 22.1 | 9.8 | 39.2 | 17.9 | 24.6 | 9.4 |
| OGL | 34.7 | 49.0 | 40.7 | 27.5 | 31.0 | 34.1 | 32.5 | **17.0** | 57.2 | 26.2 | 35.9 | 13.3 |
| lrMVL | 29.9 | 42.0 | 34.9 | 20.4 | 25.9 | 28.5 | 27.1 | 10.3 | 48.6 | 22.3 | 30.6 | 9.4 |
| MVLR | 25.9 | 37.2 | 30.5 | 16.9 | 24.5 | 27.2 | 25.8 | 9.5 | 37.7 | 17.3 | 23.7 | 7.9 |
| OPSL-V | 36.0 | 50.7 | 42.1 | 28.8 | 31.3 | 34.5 | 32.8 | 15.3 | 59.1 | 27.1 | 37.2 | 13.9 |
| **OPSL** | **37.0** | **52.1** | **43.3** | **29.6** | **32.3** | **35.6** | **33.9** | 16.1 | **60.9** | **28.0** | **38.4** | **14.5** |

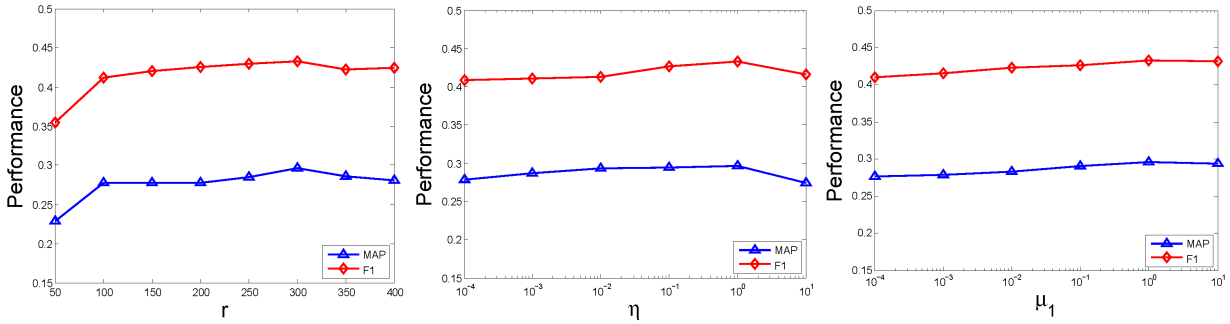


Figure 1: The parameter sensitivity analysis on Corel5k dataset.

better performance. Furthermore, OPSL-V does not perform as good as OPSL, indicating that our softmax activation function can effectively minimize the disagreement between the learned subspace and each view and better capture the multi-view complementary information. Finally, OPSL achieves the best image tagging performance in most cases mainly due to the following reasons. The local geometric structure of multi-view data is effectively enhanced by using the semantic structure constraints, and both visual structure and semantic information are preserved in the learned subspace. This guarantees the obtained subspace to be compact and discriminative. In addition, we use the learned SVM predictors to guide the subspace learning which makes the subspace to be more proper for tag prediction task. Moreover, OPSL jointly conducts multi-view representation learning and image tagging. The two tasks promote each other so that the overall performance can be improved.

Next, we study the sensitiveness of three important parameters $r$, $\eta$ and $\mu_1$ on Corel5k dataset. The performance of different parameter settings are shown in Figure 1. We can observe that when the dimensionality of the learned subspace $r$ is small ($r < 200$), the subspace cannot fully encode the information of multi-view data so that the annotation performance is seriously influenced. The proper dimensionality of the subspace is from 200 to 350. $\eta$ controls the strength of semantic information preserving. As can be seen from Figure 1, the larger $\eta$ the more semantic information can be preserved, and the performance is increasing. While the strength becomes too strong for $\eta \geq 10$ and the performance is influenced. The best performance is achieved for $\eta = 1$. $\mu_1$ is the weight of SVM predictors learning term. The larger of $\mu_1$, the more strength that SVM predictors impose on the learned subspace. As shown in Figure 1, the annotation performance can be improved as the value of $\mu_1$ increases, which utilizes more supervised information from SVM to guide subspace learning. For $\mu_1 > 1$, the performance varies very little. The best performance can be achieved around $\mu_1 = 1$.

## Conclusion

In this paper, we propose a novel image tagging method that jointly conducts multi-view representation learning and image tagging. The two tasks can promote each other to achieve the better annotation performance. We learn an optimal predictive subspace from multi-view data to obtain a proper representation for image tagging, which builds a bridge between low-level features and high-level semantic information. Both local geometric properties of multi-view data and semantic information are exploited for subspace learning which makes the learned representation more compact and discriminative. Experimental results on three standard image annotation datasets demonstrate the effectiveness of the proposed method.

## Acknowledgments

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351800, in part by National Natural Science Foundation of China: 61332016 and 61303153.

## References


Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19(7):711–720.

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, 585–591.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research* 7:2399–2434.

Bertsekas, D. P. 1999. Nonlinear programming.

Cao, X.; Zhang, C.; Fu, H.; Liu, S.; and Zhang, H. 2015. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–594.

Chen, M.; Zheng, A.; and Weinberger, K. 2013. Fast image tagging. In *Proceedings of the 30th international conference on Machine Learning*, 1274–1282.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 48. ACM.

Duygulu, P.; Barnard, K.; de Freitas, J. F.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision–ECCV 2002*. Springer. 97–112.

Elhamifar, E., and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(11):2765–2781.

Feng, Z.; Feng, S.; Jin, R.; and Jain, A. K. 2014. Image tag completion by noisy matrix recovery. In *Computer Vision–ECCV 2014*. Springer. 424–438.

Gao, L.; Song, J.; Nie, F.; Yan, Y.; Sebe, N.; and Shen, H. T. 2014. Optimal graph learning with partial tags and multiple features for image and video annotation. *Trans. Cybernetics* 44(7):1225–1236.

Guillaumin, M.; Mensink, T.; Verbeek, J.; and Schmid, C. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, 309–316. IEEE.

Kalayeh, M. M.; Idrees, H.; and Shah, M. 2014. Nmf-knn: image annotation using weighted multi-view non-negative matrix factorization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 184–191. IEEE.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 556–562.

Lin, Z.; Ding, G.; Hu, M.; Wang, J.; and Ye, X. 2013. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 1618–1625. IEEE.

Lin, Y. Y.; Liu, T. L.; and Fuh, C. S. 2011. Multiple kernel learning for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33(6):1147–1160.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Long, B.; Philip, S. Y.; and Zhang, Z. M. 2008. A general model for multiple view unsupervised learning. In *SDM*, 822–833. SIAM.

Makadia, A.; Pavlovic, V.; and Kumar, S. 2008. A new baseline for image annotation. In *Computer Vision–ECCV 2008*. Springer. 316–329.

Platt, J., et al. 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodsłsupport vector learning* 3.

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

Schölkopf, B., and Smola, A. J. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Seung, H. S., and Lee, D. D. 2000. The manifold ways of perception. *Science* 290(5500):2268–2269.

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

Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. ACM.

Wang, Q.; Shen, B.; Wang, S.; Li, L.; and Si, L. 2014. Binary codes embedding for fast image tagging with incomplete labels. In *Computer Vision–ECCV 2014*. Springer. 425–439.

Wu, L.; Jin, R.; and Jain, A. K. 2013. Tag completion for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(3):716–727.

Xia, T.; Tao, D.; Mei, T.; and Zhang, Y. 2010. Multiview spectral embedding. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 40(6):1438–1446.

Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *Advances in neural information processing systems*, 1601–1608.

Zheng, S.; Cai, X.; Ding, C.; Nie, F.; and Huang, H. 2015. A closed form solution to multi-view low-rank regression. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.