Collaborative Similarity Metric Learning for Semantic Image Annotation and Retrieval

Bin Wang*, Yuncai Liu
Department of Automation, Shanghai Jiao Tong University
Shanghai 200240, China
[e-mail: binwang@sjtu.edu.cn]
*Corresponding author: Bin Wang

Received November 19, 2012; revised February 3, 2013; accepted April 30, 2013; published May 31, 2013

Abstract

Automatic image annotation has become an increasingly important research topic owing to its key role in image retrieval. Simultaneously, it is highly challenging when facing to large-scale dataset with large variance. Practical approaches generally rely on similarity measures defined over images and multi-label prediction methods. More specifically, those approaches usually 1) leverage similarity measures predefined or learned by optimizing for ranking or annotation, which might be not adaptive enough to datasets; and 2) predict labels separately without taking the correlation of labels into account. In this paper, we propose a method for image annotation through collaborative similarity metric learning from dataset and modeling the label correlation of the dataset. The similarity metric is learned by simultaneously optimizing the 1) image ranking using structural SVM (SSVM), and 2) image annotation using correlated label propagation, with respect to the similarity metric. The learned similarity metric, fully exploiting the available information of datasets, would improve the two collaborative components, ranking and annotation, and sequentially the retrieval system itself. We evaluated the proposed method on Corel5k, Corel30k and EspGame databases. The results for annotation and retrieval show the competitive performance of the proposed method.

Keywords: image retrieval, collaborative similarity metric learning, structural SVM , correlated label propagation

This work was supported by National Basic Research Program of China 2011CB302203.

http://dx.doi.org/10.3837/tiis.2013.05.018

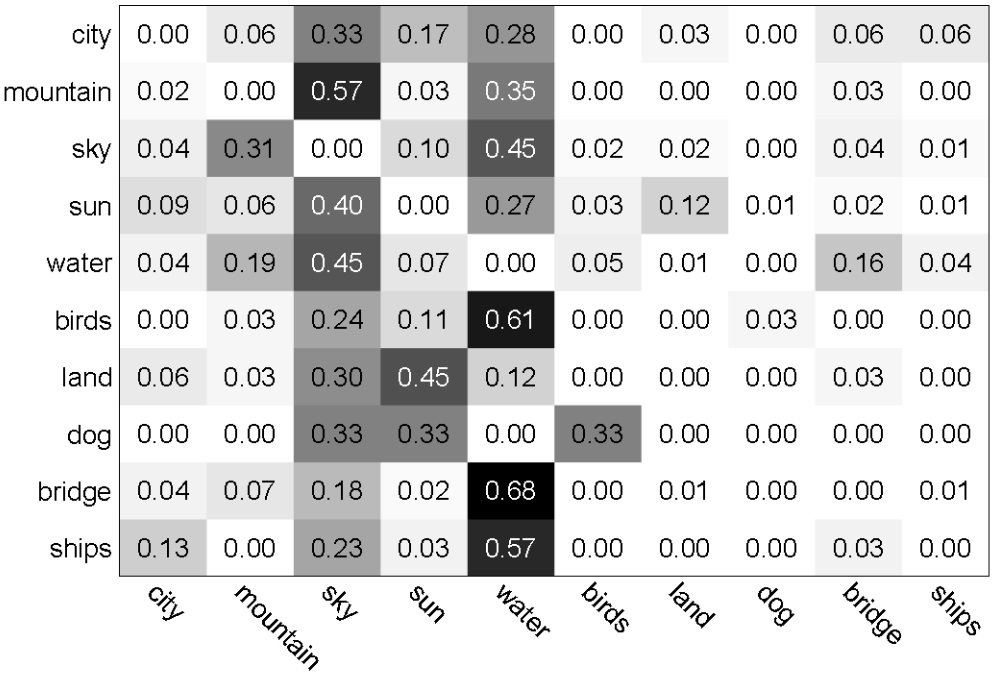
1. Introduction

Retrieving images from a huge database, e.g., Internet image database, has become increasingly important. Most popular search engines use the surrounding texts as keywords to describe images in retrieval. Their performance can be potentially improved by establishing the correspondence between images and keywords well describing the content of these images. The procedure of automatically establishing such correspondence is known as automatic image annotation which has become an active research field in recent years [1][2]. A number of approaches, based on computer vision, machine learning and related fields, have been proposed to attack this problem. However, image annotation is still highly challenging due to the variance of background, the complexity of object overlapping, and so on in images. Even though the general problem is quite difficult, significant progress has been made.

Current image annotation approaches mainly consist of two groups: generative model based approaches [3][4] and discriminative model based approaches [5][6]. We will give a detailed review in Section 2. In this paper, we focus on discriminative model based approaches because of their state-of-the-art performance in image annotation. These approaches are composed of three primary components: feature, similarity/distance metric, and label prediction strategy. Here, we concentrate on similarity metric and label prediction strategy. Our proposed approach is motivated by the limitation issues of the previous approaches with regards to the first two components.

First, some previous approaches utilize predefined similarity/distance measures to judge the similarity between images [7], for instance, Euclidean metrics, Gaussian kernel similarity measure and L1 distance [8]. These predefined distance metrics ignore the distribution of images and their associated keywords, and are less adaptive to the image distribution. To adapt to the data distribution, metric learning approaches [9], through optimizing for either image ranking or annotation, have been proposed to attack the above limitation. These approaches, however, treat ranking and annotation as two separate procedures while they are highly dependent in fact.

Second, previous approaches generally annotates label/keyword independently, rarely considering the inherent correlation between the labels [10]. This is not reasonable because the correlation between class labels does exist. The conditional probability matrix $P(column|row)$ of the keywords of a subset of Corel5K data set is summarized in **Fig. 1**. The matrix shows that the correlation between keywords does exist. For instance, the elements in the 10-th row and 5-th column is $P(water|ships)=0.57$, and the elements in the 10-th row and 8-th column is $P(dog|ships)=0.00$. The conditional probability $P(water|ships)=0.57$ is much greater than $P(dog|ships)=0.00$. That is, the correlation between *ships* and *water* is much closer than the correlation between *ships* and *dog*. This observation suggests that the correlation between class labels can be very informative in the label prediction of image annotation.



	city	mountain	sky	sun	water	birds	land	dog	bridge	ships
city	0.00	0.06	0.33	0.17	0.28	0.00	0.03	0.00	0.06	0.06
mountain	0.02	0.00	0.57	0.03	0.35	0.00	0.00	0.00	0.03	0.00
sky	0.04	0.31	0.00	0.10	0.45	0.02	0.02	0.00	0.04	0.01
sun	0.09	0.06	0.40	0.00	0.27	0.03	0.12	0.01	0.02	0.01
water	0.04	0.19	0.45	0.07	0.00	0.05	0.01	0.00	0.16	0.04
birds	0.00	0.03	0.24	0.11	0.61	0.00	0.00	0.03	0.00	0.00
land	0.06	0.03	0.30	0.45	0.12	0.00	0.00	0.00	0.03	0.00
dog	0.00	0.00	0.33	0.33	0.00	0.33	0.00	0.00	0.00	0.00
bridge	0.04	0.07	0.18	0.02	0.68	0.00	0.01	0.00	0.00	0.01
ships	0.13	0.00	0.23	0.03	0.57	0.00	0.00	0.00	0.03	0.00

Fig. 1. The conditional probability matrix $P(column|row)$ of the keywords of Corel5K data set.

We observe that ranking and annotation procedures are dependent on each other in practice. At the annotation step, the nearest images of the input image in ranking are selected to train the annotation model; at the retrieval step, images with the same label are reported according to the ranking. Motivated by this observation, we propose an integrated approach to tackle the above two limitation issues by treating ranking and annotation as a whole. We employ the structural SVM [11] with parameterized similarity measure embedding for ranking and correlated label propagation with parameterized similarity measure embedding for annotation. The correlated label propagation, taking the correlation between labels into account, can effectively address the label correlation issue (the second issue). The point as well as the main contribution of our proposed approach lies in the *the similarity metric learning* (the first issue). We propose to learn the similarity metric by simultaneously considering the ranking and annotation. The idea is to optimize the ranking and annotation with respect to similarity metric, encouraging the similarity metric to have less ranking error and annotation error. At the *annotation step*, the learned similarity metric is then used to annotate new input images. The proposed approach has the following advantages:

(1) We present a parameterized similarity measure which allows adapting to the data set, and present similarity measure based SSVM for ranking and similarity measure based label propagation for annotation, by which, both ranking and annotation can benefit from the adaption of similarity measure. Also, label correlation is considered in label propagation process.

(2) We propose a collaborative learning approach for similarity metric, which simultaneously optimizes for ranking and annotation with respect to similarity metric. The learning approach fully exploits the information on hand for ranking and annotation. The resulted problem takes a standard form and can be efficiently solved using cutting-plane algorithm [11].

2. Related Works

In this section, we review previous image annotation approaches. We identify two groups of approaches: discriminative models based approaches and generative models based approaches. We highlight discriminative model based approaches which are closely related to our work.

Generative models model the distribution of images and explain how can they be generated. They can infer and exploit information hidden in images. The hidden information usually closely relates to the high level concepts of images. For instance, in PLSA [12] and LDA [3], the hidden variable 'topic' describes what a set of image patches is. Generative models therefore can utilize these additional high level information for annotation. By means of naive Bayes classifier, they can be used to perform classification. Among generative models, topic models and mixture models are usually used for image annotation. The topic models model the images as samples drawn from a specific mixture of topics, in which each topic is a distribution over images and keywords. Representative models include LDA, MOM-HDP [4] and PLSA. The mixture models define a joint distribution over images and keywords. [13] uses a fixed number of mixture components over the images per keyword. [14][15][16] define a mixture model over images and keywords using the training images as components.

Discriminative models for image annotation have been proposed recent years [5][6]. Discriminative models directly model the map from images to image labels, by capturing the decision bounds among different classes. Specifically, these approaches characterize each keyword as a class and learn a separate classifier for the keyword. The classifier is then used to predict whether the class label can be assigned to the test image. A variety of discriminative models, such as support vector machines (SVM) [6], discriminative kernel type model [5], and multiple-instance learning [17] have been been applied to image annotation. Among these discriminative approaches, we focus on nearest neighbor based discriminative models.

Nearest neighbor based discriminative approaches are becoming attractive due to its computational efficiency and good performance when the amount of training data is relatively large. For example, [18] learns discriminative models in the neighborhoods of test images, while label propagation over a similarity graph of both labeled and unlabeled images are derived to solve the image annotation problem [19][20]. Among these methods, we pay attention to the works most related to our work. A semantic distance function (SDF) is proposed in [9], which is learned based on relative comparison relations. To annotate a new image, training images are ranked according to SDF towards this image, and their labels are then propagated to this image. TagProp [21] is a newly introduced nearest neighbor model which predicts keywords by taking a weighted combination of the label absence or presence among neighbors. There the weights for neighbors are determined either based on the neighbor ranking or its distance. The method assumed that the similarity measure for determining the neighbor ranking are predefined without considering its adaption to data instances. In [22], an adhoc nearest neighbor label propagation mechanism is introduced. In this method, nearest neighbors are determined by a simple combination of several predefined distances. Keywords are then propagated from the neighbors to the test image.

The above approaches vary in formulation and technique. However, they roughly share the same limitations as described in Section 1. To overcome these limitations, we propose a new approach in the next section.

3. Proposed Approach for Content-based Image Annotation

In this section, we will present our method for similarity metric learning by considering both image ranking and annotation, and then use the learned metric for ranking and annotation. The proposed method is graphically illustrated in **Fig. 2**. Specifically, given image features, SSVM with similarity metric embedding is used for image ranking, and label propagation with similarity metric embedding is employed for annotation. In the training procedure, we determine the similarity metric by optimizing for ranking and annotation with respect to similarity metric. In the annotation procedure, we first rank the input image using SSVM with learned similarity metric and then perform annotation using correlated label propagation (over the related training images in ranking) with learned similarity metric.

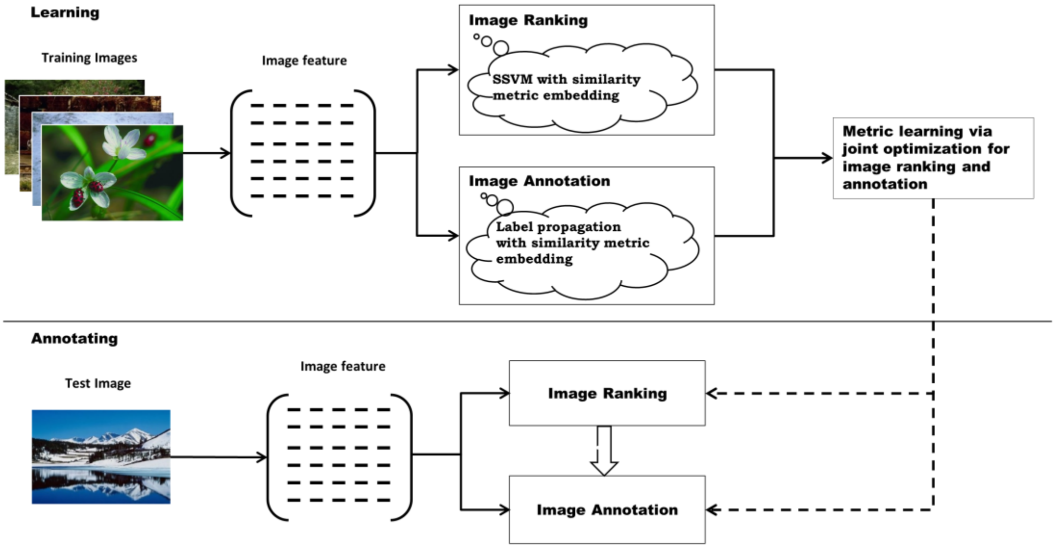

Learning
Training Images
Image feature
Image Ranking
SSVM with similarity metric embedding
Image Annotation
Label propagation with similarity metric embedding
Metric learning via joint optimization for image ranking and annotation
Annotating
Test Image
Image feature
Image Ranking
Image Annotation

Fig. 2. The graphical illustration of the proposed approach.

3.1. Parameterized Similarity Measure

We observe that different data sets usually have different distributions. A similarity measure well adapting to the data distribution would enhance both image ranking and annotation effectively. Suppose that images from a data set lie in the same space $x \in R^m$. For a pair of images $x_i, x_k \in X$, let $S(x_i, x_k)$ denote its similarity measure. We adopt the following generalized inner product as the similarity measure,

$$S_A(x_i, x_k) = \frac{1}{2}\left(x_i^T A x_k + x_k^T A x_i\right) \tag{1}$$

where $A \in R^{m\times m}$ is a semi-definite matrix. It can be further written as,

$$\begin{aligned} S_A(x_i, x_k) &= \frac{1}{2}\left(x_i^T A x_k + x_k^T A x_i\right) = \frac{1}{2}\left(tr(Ax_k x_i^T) + tr(Ax_i x_k^T)\right) \\ &= \frac{1}{2}\left(\left\langle A, x_i x_k^T\right\rangle_F + \left\langle A, x_k x_i^T\right\rangle_F\right) = \left\langle A, \frac{1}{2}(x_i x_k^T + x_k x_i^T)\right\rangle_F \end{aligned} \tag{2}$$

where $tr()$ symbolizes the *trace* operator; $\langle \cdot,\cdot \rangle_F$ is the Frobenius inner product which is the component-wise inner product of two matrices, i.e., $\langle A, B\rangle_F = \sum_{i,j} A_{ij} B_{ij}$ .Although this similarity measure is quite simple, it is very effective in real application, in comparison with some other measures tested in our experiments. Moreover, the simple form allows embedding it to ranking and annotation approaches and developing collaborative learning algorithm, which will be presented in the following parts.

3.2. SSVM with Similarity Metric Embedding for Image Ranking

Image ranking can be casted to a structured output prediction problem, where the prediction result is a complex structure rather than a single quantity. In this paper, we exploit Structural SVM (SSVM) for image ranking [11]. First, we introduce some notations and the problem setting mathematically. Let $X=\{x_i\}_{i=1}^{N}$ be a set of N training images, where $x_i \in R^m$. Let Y be the set of the overall rankings of X. Let X_+ and X_- signify the relevant and irrelevant image sets of image $x \in X$ respectively. For an image x_i, let y_i denote its correct ranking, and $\hat{y}_i$ denote any other ranking in the output space Y.

3.2.1. SSVM for Image Ranking

SSVM has been proved to be promising for building highly accurate models in areas like information retrieval, natural language processing and protein structure prediction [23]. We apply SSVM to ranking where the image $x \in X$ is the input and ranking $\hat{y}$ is the output:

$$\hat{y}(x)=\arg\max_{\hat{y}\in Y} \lambda(x,\hat{y})+\Delta(y,\hat{y}) \tag{3}$$

where y is its correct ranking of x; $\hat{y}$ is other possible ranking in Y; λ is a discriminant function over all input and output pairs, $\lambda: X\times Y \to R$; $\Delta(y,\hat{y})$ is the margin defined between two rankings y and $\hat{y}$, which is a non-negative loss function. For a specific input, a prediction can be given by maximizing the function λ and the loss function over the whole output space. Assuming that the discriminative function λ takes a linear form: $\lambda_\omega(x,\hat{y})=\omega^T\phi(x,\hat{y})$ where ω is a parameter; ϕ is a vector-valued feature map connecting the input x and output $\hat{y}$. $\omega^T\phi(x,\hat{y})$ can be thought to be a compatibility measure that judges how well the output space matches the given input space. Substituting it to Eq. (3), we get:

$$\hat{y}(x)=\arg\max_{\hat{y}\in Y} \omega^T\phi(x,\hat{y})+\Delta(y,\hat{y}) \tag{4}$$

The most commonly used feature map ϕ is the partial order feature mentioned in [24]:

$$\phi_{\mu_d}(x,\hat{y})=\sum_{x_i\in X_+}\sum_{x_j\in X_-} y_{ij}\left(\frac{\mu_d(x,x_i)-\mu_d(x,x_j)}{|X|_+|X|_-}\right) \tag{5}$$

where $\mu_d(x,x_i)$ is vector-valued which characterizes the relationship between x and x_i; $y_{ij}=+1$ if x_i is placed before x_j in $\hat{y}$ and $y_{ij}=-1$ if x_i is placed after x_j in $\hat{y}$;

Substitute Eq. (5) into Eq. (4), we obtain:

$$\hat{y}(x)=\arg\max_{\hat{y}\in Y}\sum_{x_i\in X_+}\sum_{x_j\in X_-} y_{ij}\left(\frac{\omega^T\mu_d(x,x_i)-\omega^T\mu_d(x,x_j)}{|X|_+|X|_-}\right)+\Delta(y,\hat{y}) \tag{6}$$

3.2.2. SSVM with Similarity Measure Embedding

In Eq. (6), $\omega^T\mu_d(x,x_i)$ plays the role of distance and $\mu_d(x,x_i)$ is a distance metric.

Maximizing the objective function equals to put x in a proper place in X. Now, we generalize the distance metric $\mu_d(x,x_i)$ to similarity metric $\mu_s(x,x_i)$ in order to learn the similarity metric directly. $\mu_s(x,x_i)$ is also a vector-valued feature map. Accordingly, we get the following expression with regards to similarity metric $\mu_s(x,x_i)$:

$$\begin{aligned}\hat{y}(x) &= \arg\max_{\hat{y}\in Y} \omega^T \phi_{\mu_s}(x,\hat{y}) + \Delta(y,\hat{y}) \\ &= \arg\max_{\hat{y}\in Y} \sum_{x_i\in X_+}\sum_{x_j\in X_-} y_{ij}\left(\frac{\omega^T\mu_s(x,x_j)-\omega^T\mu_s(x,x_i)}{|X|_+|X|_-}\right)+\Delta(y,\hat{y})\end{aligned} \tag{7}$$

The similarity metric μ_s has the following property.

Remark 1 For a test image x_k and a given ω, the ranking $\hat{y}$ of $x_i \in X$ obtained by ascendingly sorting $\omega^T\mu_s(x_k,x_i)$ $\Longleftrightarrow$ the ranking $\hat{y}$ which maximizes $\omega^T\phi_{\mu_s}(x_k,\hat{y})$. This is a generalization of the property of ϕ_{μ_d} [23].

Now we specify the above similarity measure $\omega^T\mu_s(x_k,x_i)$ to our parameterized similarity measure $\left\langle A, \frac{1}{2}(x_i x_k^T + x_k x_i^T)\right\rangle_F$ in Eq. (2), where $\mu_s(x_k,x_i)=\frac{1}{2}(x_i x_k^T + x_k x_i^T)$. Then we have the following property.

Remark 2 Sort $x_i \in X$ by ascending $S_A(x_k,x_i)$ $\Longleftrightarrow$ the ranking $\hat{y}$ which maximizes the generalized Frobenius inner product $\left\langle A, \phi_{\mu_s}(x_k,\hat{y})\right\rangle_F$. This is because, (1) according to Eq. (2), sort $x_i \in X$ by ascending $S_A(x_k,x_i)$ $\Longleftrightarrow$ sort $x_i \in X$ by ascending $\left\langle A, \mu_s(x_k,x_i)\right\rangle_F$; (2) according to Remark 1, when using $\mu_s(x_k,x_i)$ with ϕ_{μ_s} in Eq. (7), $x_i \in X$ sorted by ascending $\left\langle A, \mu_s(x_k,x_i)\right\rangle_F$ $\Longleftrightarrow$ the ranking $\hat{y}$ which maximizes the generalized Frobenius inner product $\left\langle A, \phi_{\mu_s}(x_k,\hat{y})\right\rangle_F$.

SSVM with similarity metric defined in Eq. (2) can be expressed as the following optimization problem,

$$\min_{A\pm 0,\xi\geq 0} tr(A)+c\xi \tag{8}$$

$$s.t. \sum_{i=1}^{N}\left\langle A,[\phi(x_i,y_i)-\phi(x_i,\hat{y}_i)]\right\rangle_F \geq \sum_{i=1}^{N}\Delta(y_i,\hat{y}_i)-N\xi, \quad \forall\, \hat{y}_i\in Y \tag{9}$$

where ξ is a slack variable and c is a weight coefficient.

3.3. Correlated Label Propagation with Similarity Metric Embedding for Image Annotation

In this section, we focus on the label propagation approach over the learned similarity metric, with the correlation between class labels explicitly considered.

Suppose the number of class labels is z. We consider the problem that gives z class labels to an input image x_k, i.e., annotating x_k. Let $V_k=(v_k(1),\ldots,v_k(z))^T$ be its binary label vector, where $v_k(r)=1$ iif the r-th label is given to the image x_k and $v_k(r)=0$ otherwise. Then image annotation is casted to the problem that determines V_k for x_k under a certain criteria. In this paper, we consider label propagation based criteria which essentially assesses how close an image x_k belongs to a class r using a score $p_k(r)$. The score $p_k(r)$ is essentially the summation of the similarity between x_k and all instances with label r.

3.3.1. Independent label Propagation

First we consider the independent label propagation. Kernel-based KNN, as one of the representative approaches for label propagation [25], propagates label r to image x_k according to the confidence score $p_k(r)=\sum_{i=1}^{N} S_A(x_k,x_i)v_i(r)$. The confidence score $p_k(r)$, simply summing all similarities of images with label r, can be overestimated because even if x_i is similar to x_k it is not necessarily true to propagate all the class labels of x_i to x_k. A solution to overcome the overestimation is to release the above equation to the following inequality [25]:

$$p_k(r) \le \sum_{i=1}^{N} S_A(x_k,x_i)v_i(r) \tag{10}$$

It is worth noting that the confidence score $p_k(r)$ of r-th label is independently computed, without considering its correlation with other labels.

3.3.2. Correlated Label Propagation with Similarity Metric Embedding

To overcome the limitation of independent single label propagation, we perform the correlated propagation of multiple labels via considering the correlation of class labels [25]. We define $q(V_k)$ as the confidence score of propagating any subset of V_k to x_k. Similar to Eq. (10), we get the following constraints towards $q(V_k)$:

$$q(V_k) \le \sum_{i=1}^{N} S_A(x_k,x_i))\, I(V_k^T V_i) \tag{11}$$

where $I(a)$ is an indicator function that outputs 1 for $a>0$ and 0 otherwise. Note that $I(V_k^T V_i)=1$ if x_k and x_i have at least one common label. To link $p_k(r)$ with $q(V_k)$, we follow the principle that, the confidence score of propagating the labels individually in a label set V_k to x_k should be no more than that of propagating any subset of V_k to the image x_k. Thus we have,

$$P_k^T V_k \le q(V_k) \le \sum_{i=1}^{N} S_A(x_k,x_i)\, I(V_k^T V_i) \tag{12}$$

where $P_k=(p_k(1),\ldots,p_k(z))^T$.

Assuming that the optimal P_k satisfying Eq. (12) is the one which maximally satisfies the constraints, we obtain the following optimization problem with regards to P_k:

$$\max_{P_k \in R^z} \alpha^T P_k \tag{13}$$

$$s.t.\ \ P_k^T V_k \le \sum_{i=1}^{N} S_A(x_k,x_i)\, I(V_k^T V_i),\ \ \forall V_k \in \{0,1\}^z \tag{14}$$

where $\alpha=(\alpha_1,\cdots,\alpha_z)^T$ are the weights for each class label. To solve the above optimization problem, we generalize the indicator function I to a concave function C, and then use a greedy algorithm [25]. The overall algorithm is summarized in Algorithm 1. We see that the estimating of confidence sore $p_k(r)$ is eventually decided by the concave function C. Here, we choose sigmoid function $C(x)=\frac{1}{1+e^{(-x)}}$ [25]. It is worth noting that the solution only depends on the relative order of α's elements and independent with their exact values. With the confidence score P_k for image x_k, we assign the r-th label to x_k by comparing with a threshold p_{th}, i.e., $v_k(r)=1$ if $p_k(r)>p_{th}$ and $v_k(r)=0$ otherwise.

Algorithm 1 Determine confidence scores based on similarity metric S_A

```
input: an image $x_k$ to be labeled ; $\alpha_1 \ge \alpha_2 \cdots \ge \alpha_z$
for $r=1$ to $z$ do
    construct a label set $L_r=\{i\}_{i=1}^{r}$
    $p_k(r)=f(L_r)-f(L_{r-1})=\sum_{i=1}^{N'} S_A(x_k,x_i)\left[C(V_{L_r}^T V_i)-C(V_{L_{r-1}}^T V_i)\right]$
end for
output: the confidence scores $\{p_k(r)\}_{r=1}^{z}$ for an unlabeled image $x_k$.
```

3.4. Collaborative Metric Learning via Joint Optimization of Ranking and Annotation

So far, we have a flexible similarity measure over the defined similarity metric A, as presented in Eq. (2), which can be embedded into both image ranking and image annotation. In this section, we aim to derive an algorithm to learn the similarity metric which simultaneously optimizes for ranking and annotation.

We note that, correlated label propagation is a nearest neighbor based approach, expecting images with the same label to have high similarity. For situation considering the label correlation as in correlated label propagation, this can be expressed as the following objective function,

$$\max_{A} \sum_{i=1}^{N} \sum_{j \ne i} S_A(x_i,x_j)\, I(V_i^T V_j)$$

Note that, only when the images i and j have common label, i.e., $I(V_i^T V_j)=1$, their similarity will contribute to the objective function. This is pretty intuitive for the correlated label propagation in Eq. (14). That is,when we maximize the above formula, the right side of the inequality of Eq. (14) is also maximized, which enlarges the probable solution space.

In this paper, we parameterized the similarity by a metric A. Then the problem of learning metric A can be expressed as the following optimization problem,

$$\min_{A \pm 0, \xi \ge 0} tr(A)+c\xi-\frac{\rho}{N(N-1)}\sum_{i=1}^{N}\sum_{k \ne i}^{N}\left\langle A, x_i x_k^T \right\rangle_F I(V_k^T V_i) \qquad (15)$$

$$s.t. \sum_{i=1}^{N}\left\langle A,[\phi(x_i,y_i)-\phi(x_i,y_i)]\right\rangle_F \ge \sum_{i=1}^{N}\Delta(y_i,y_i)-N\xi, \quad \forall\, y_i \in Y \qquad (16)$$

where $\rho>0$ is a parameter tuning the balance of the two components.

Note that the third term of Eq. (15) can be formulated as,

$$\frac{\rho}{N(N-1)}\sum_{i=1}^{N}\sum_{k\neq i}^{N}\left\langle A, x_i x_k^T\right\rangle_F I(V_k^T V_i) = \left\langle A, \frac{\rho}{N(N-1)}\sum_{i=1}^{N}\sum_{k\neq i}^{N} x_i x_k^T I(V_k^T V_i)\right\rangle_F$$

which is the linear function with respect to matrix A. Then the complete learning problem is,

$$\min_{A\pm 0,\xi\geq 0} tr(A)+c\xi-\left\langle A, \frac{\rho}{N(N-1)}\sum_{i=1}^{N}\sum_{k\neq i}^{N} x_i x_k^T I(V_k^T V_i)\right\rangle_F \tag{17}$$

$$s.t.\ \sum_{i=1}^{N}\left\langle A,[\phi(x_i,y_i)-\phi(x_i,\hat{y}_i)]\right\rangle_F \geq \sum_{i=1}^{N}\Delta(y_i,\hat{y}_i)-N\xi, \quad \forall\, \hat{y}_i \in Y \tag{18}$$

This problem can be solved using the cutting plane approach which is used to solve SSVM [11]. In the algorithm, we alternately optimize similarity metric A and update constraint set. The constraints here indicate that, for each image ranking pair (x_i, y_i), the score of $\phi(x_i, y_i)$ for the correct ranking should be much larger than the score of $\phi(x_i, \hat{y}_i)$ for any other ranking in the output space. The algorithm iterates through constructing a working set of the constraints and finding the most violated constraint by the current A and ξ. The algorithm terminates when the constraints are satisfied. The algorithm is summarized in Algorithm 2.

Algorithm 2 Metric learning via joint optimization

```
input: image ranking pairs ${(\mathbf{x}_i, y_i)}_{i=1}^{N}$, $c$, $\varepsilon$
construct a working set $T$ and initialize $T \leftarrow \%$
repeat
    $(A,\xi) \leftarrow$ solve Eq. (17) subject to Eq. (18) on $T$ using cutting plane
    for $i=1$ to $N$ do
        $\hat{y}_i \leftarrow \mathrm{argmax}_{\hat{y}\in Y}\Delta(y_i,\hat{y})+\langle A,\phi(x_i,\hat{y})\rangle_F$
    end for
    $T \leftarrow T \cup \{(\hat{y}_1,\ldots,\hat{y}_N)\}$
until $\sum_{i=1}^{N}\Delta(y_i,\hat{y}_i)-\sum_{i=1}^{N}[(A,\phi(x_i,y_i))_F-(A,\phi(x_i,\hat{y}_i))_F]>N(\phi+\varepsilon)$
output: $A$ and $\xi$.
```

The similarity metric A optimizing both for image ranking and annotation is learned ultimately. For a test image x_k, we only choose the most related images (top ones of induced image ranking) along with the learned similarity metric to participate in label propagation process, which can remove irrelevant information in the dataset and lead to great computation cost reduction. Then confidence score $p_k(r)$ for label r is computed according to Algorithm 1.

4. Experiments

In this section, we extensively evaluate the proposed method on three popular data sets, Corel5K, Corel30K and EspGame, for image annotation and retrieval. The proposed method will compare with several state-of-the-art methods.

4.1. Datasets

Corel5k The dataset has become an important benchmark for semantic-based image annotation and retrieval [14][15][26]. It contains 5,000 images collected from 50 Corel Stock Photo CDs. These 5,000 images are partitioned into a training subset of 4500 images and a testing subset containing 500 images. There are 50 different topics in the dataset. Each topic contains 100 images and every image is annotated with 1 to 5 keywords. The dictionary consists of 260 keywords which appear in both training and testing set.

Corel30k The Corel30k data set is an extension of Corel5k, and is relatively larger than Corel5K. It attempts to overcome the limitations of Corel5k [13][27][28], i.e. the image annotation system trained fromColel5K might has poor generalization ability. Corel30k contains 31,695 images, out of which 90 % images (28,525 images) are used to train system models and 10 % images (3,170 images) are used as a testing set [13]. 950 keywords are selected into the vocabulary [13]. Some sample images from both Corel5k and Corel30k are shown in **Fig. 3**.

EspGame EspGame is obtained from an online game, where two players label the same image without any communication and only the same labels are accepted. The players are in this way to encouraged to provide meaningful labels to images. The subset we use is comprised of 19,659 training images and 2,185 testing images, which is also used in [21]. Each image is associated with 4.6 labels on average. This dataset is very challenging.

Fig. 3. Illustration of Corel data together with two example images and their associated keywords

4.2. Feature Extraction

The performance of color SIFT descriptors [29] have been validated to be effective in image recognition. In this paper, four color SIFT descriptors (OpponentSIFT, C-SIFT, *rg*SIFT and RGB-SIFT) recommended by [29] are used to represent the visual attributes of images. To combine these color SIFT descriptors, we followed a similar setting as [29], 1) simultaneously using dense sampling and Harris-Laplace point sampling; and 2) leveraging spatial pyramid.

4.3. Evaluation Criteria

The performance of image annotation is evaluated by comparing the keywords generated from different methods with the ground truth annotations. Similar to previous approaches [30][16][13], for each image, we use top five keywords with the largest confidence scores obtained in label propagation procedure (i.e., $P_k = (p_k(1),\cdots,p_k(z))^T$ given by Algorithm 1) as its final annotations. Precision and recall of each keyword in the testing set are used as the performance criteria. For a keyword r, the total number of images automatically annotated with r is denoted as N_t, the number of images which are correctly annotated with r is defined as N_c, and the number of images whose ground truth annotations include r is denoted as N_g. Consequently, we present the definitions of precision and recall respectively: $Precision(r) = N_c / N_t$; $Recall(r) = N_c / N_g$. Both of the two criteria are computed over the whole keywords contained in the testing dataset. We also take the number of keywords having non-zero recall into account, which indicates how effectively the system works. And it is defined as N_+.

Moreover, we also evaluate the method for semantic image retrieval. Firstly, the annotation procedure assigns five keywords with the highest confidence scores to every image. Then given a single query, the method can return all the relevant images in the testing dataset('relevant' means the ground truth annotations have the query keyword), which are ranked according to their confidence sores in the label propagation procedure. For the query, we choose the top retrieved images. We use mean average precision (MAP) [31] which is a standard measure to assess the retrieval performance. Precision for image retrieval can be defined as the percentage of images whose ground truth annotations contain the query keyword. Average precision (AP) focuses on ranking relevant images higher [15], and is the average of the precision values at the ranks where relevant items occurs. MAP is given by averaging AP over all the query keywords.

4.4. Experimental Results for Image Annotation

To verify the effectiveness of the integration of collaborative similarity metric learning and correlated label propagation, we firstly present a group of comparative experiments on corel5k and espgame data sets. The experiments are composed of five different implementations: I) predefined similarity measure + label propagation [22]; II) predefined similarity measure + correlated label propagation; III) Similarity metric learning (optimized for ranking) + Correlated label propagation; IV) Similarity metric learning (optimized for annotation) + Correlated label propagation; V) Similarity metric learning (optimized for ranking and annotation)+ Correlated label propagation. Implementation I is the method mentioned in [22], which used predefined similarity measure and propagated keywords without considering the correlation between them. Although implementation II uses the same predefined similarity metric as in [22], it takes the correlation between labels into consideration in label propagation. Implementation III learns the similarity metric via optimizing for image ranking, and performs correlated label propagation. Implementation IV learns the similarity metric via optimizing for annotation, then performs correlated label propagation. Implementation V is the proposed method, in which the similarity metric is learned by simultaneously optimizing for image ranking and annotation. And correlation is taken into account in label propagation. The experimental results are summarized in **Table 1**. As shown in **Table 1**, implementation V shows the best performance. These results suggest that, (1) similarity learning and label correlation is rather important for image annotation performance; (2) our embedding of similarity metric learning into correlated label propagation and the collaborative similarity metric learning accounts for the convincing image annotation results of our approach. **Fig. 4** presents the precision-recall curves of implementation I and V on corel5k, with the number of annotations from 2 to 10. as shown in **Fig. 4**, implementation V (the proposed approach) consistently outperforms other implementations.

Table 1. Performance comparison of different implementations

	Corel5k			EspGame		
Implementation	Precision	Recall	N+	Precision	Recall	N+
I	0.24	0.29	127	0.21	0.24	224
II	0.26	0.30	130	0.22	0.25	224
III	0.28	0.31	135	0.24	0.26	227
VI	0.30	0.32	137	0.23	0.26	226
V	**0.31**	**0.34**	**142**	**0.26**	**0.27**	**230**

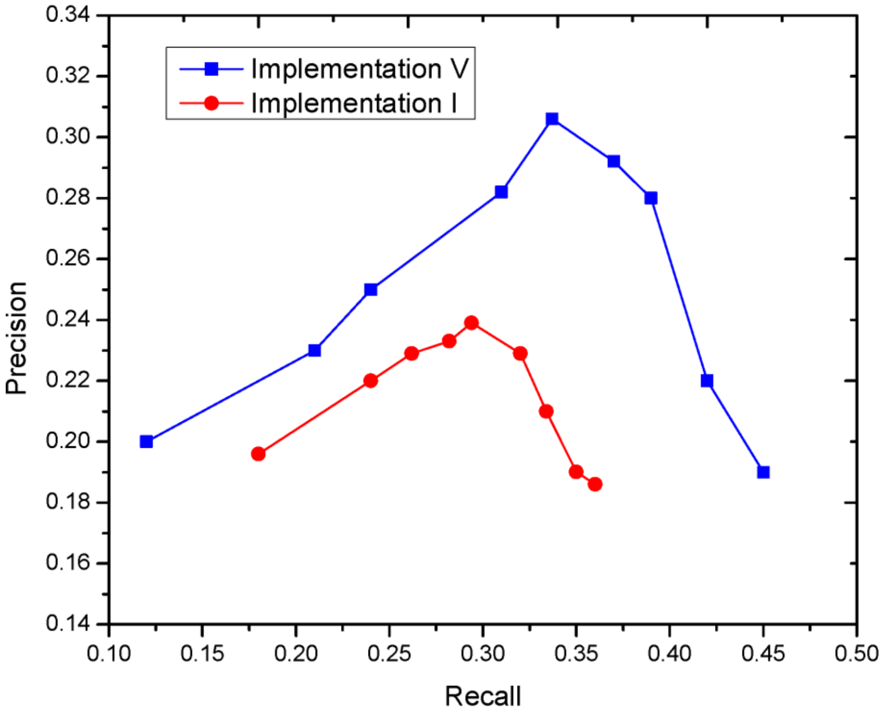


Fig. 4. Comparison precision-recall curves of implementation I and V for automatic image annotation on Corel5k

Table 2. Performance comparison of different automatic image annotation models on the Corel5K

Method	Precision	Recall	N+
CRM [16]	0.16	0.19	107
InfNet[32]	0.17	0.24	112
PLSA-F[33]	0.19	0.22	112
MBRM [15]	0.24	0.25	122
TGLM [19]	0.25	0.29	131
LASSO [22]	0.24	0.29	127
MSC [30]	0.25	0.32	136
SML [13]	0.23	0.29	137
SML-cDCT [34]	0.28	0.31	132
GS [35]	0.30	0.33	**146**
Ours	**0.31**	**0.34**	142

Table 3. Performance comparison of different automatic image annotation models on the EspGame

Method	Precision	Recall	N+
MBRM [15]	0.18	0.19	209
JEC [22]	0.22	0.25	224
JEC-15 [21]	0.24	0.19	222
Ours	**0.26**	**0.27**	**230**

To further validate the effectiveness of the proposed method, we compare with several state-of-the-art methods for image annotation from different perspectives: CRM [16], InfNet [32], PLSA-F [33], MBRM [15], TGLM [19], LASSO [22], MSC [30], SML [13], SML-cDCT [34], and GS [35]. We compute the precision and recall of each keyword and use the mean of these values to evaluate the proposed method. The experimental results on the corel dataset are summarized in **Table 2**. The results present that our method achieves the best performance in comparison with the other approaches. This is due to the successful embedding of parameterized similarity metric learning into both image ranking and correlated label propagation. Also, the collaborative similarity metric learning enables our approach to be adapted to database. To validate this, we also conduct comparative experiments on Espgame set using MBRM [15], JEC [22], JEC-15 [21] and report the experimental results in **Table 3**. The results demonstrate that our approach has superior performance as compared to other three models in the sense of precision and recall. The above results validate that our method is adaptive to data, which benefits from our parameterized similarity measure learning. **Fig. 5** shows the experimental results of some specific keywords in precision and recall. **Fig. 6** presents some examples of the annotation results produced by the proposed approach on

Corel5k and Corel30k datasets. The annotated keywords in italic font are those not contained in the ground truth annotations. Even if sometimes our method assigns keywords excluded in the ground truth annotations to an image, these keywords are still meaningful for the image in fact. This is because the correlated label propagation, which is modeled based on the parameterized similarity measure, takes the correlation between keywords into consideration. And the similarity metric is collaboratively optimized both for image ranking and image annotation.

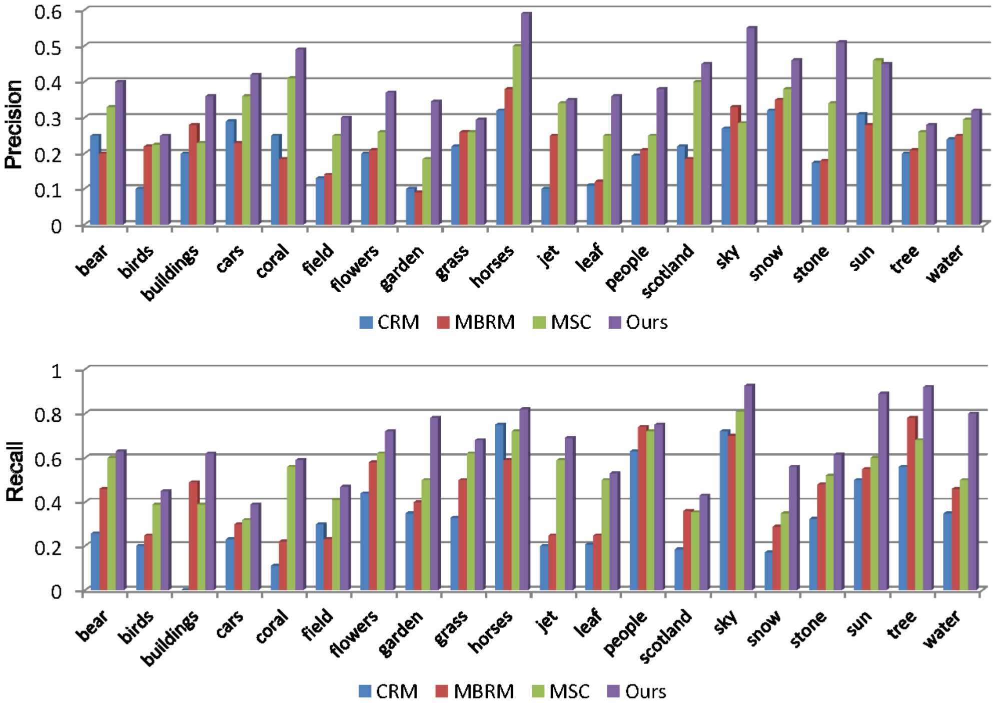
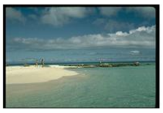
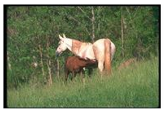
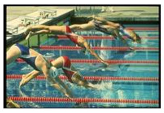
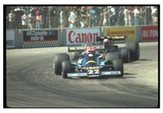
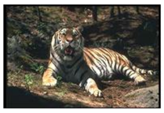


Fig. 5. Comparison results of four methods: CRM [16], MBRM [15], MSC [30] and ours

					
Human Annotation	beach sand sky water	foals grass horses mare	people pool swimmers water	cars formula tracks wall	bengal cat forest tiger
Predicted Annotation	sky *ocean* beach sand water	foals grass *tree* horses mare	water pool people swimmers *athlete*	cars formula tracks *turn* wall	tiger bengal cat *rocks* forest

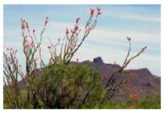
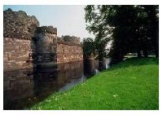
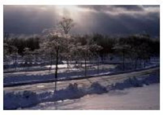
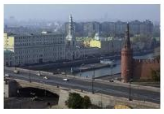
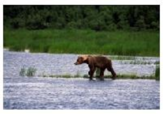
					
Human Annotation	desert blossoms hill sky	water castle tree grass	mountain ice trees winter	bridge sky cars water	bear grizzly trees water
Predicted Annotation	sky *clouds* desert blossoms hill	water castle grass tree *sky*	*snow* ice trees winter mountain	bridge sky *building* cars water	bear grizzly water *grass* trees

Fig. 6. Comparison of predicted annotations with manual annotations on the Corel dataset

4.5. Experimental Results for Semantic Image Retrieval

In this section, we evaluate our approach for the semantic-based image retrieval task. It is well known that users always favors ranked retrieval results, where the top ranked images are the most revelent ones. Actually, most users just want to see no more than 10 images for a query. Thus, ranking order is of great importance for image retrieval. We use mean average precision(MAP) to evaluate the performance of single-keyword retrieval. In the current implementation, we firstly compare our method with four related algorithms on Corel5k : the continuous-space relevance model (CRM) [16], CRM with rectangular regions as input, called CRM-Rectangles [15]; the Multiple-Bernoulli Relevance Model (MBRM) [15]; the cross-media relevance model (CMRM) [14].

Table 4. Performance comparison of semantic image retrieval results on Corel5k

Algorithms	All words	Words (recall＞0)
Mean average precision on corel5k		
CMRM [14]	0.17	0.20
CRM [16]	0.24	0.27
CRM-Rectangles[15]	0.26	0.30
MBRM [15]	0.30	0.35
Ours	**0.31**	**0.37**

For a query keyword, the proposed approach returns those images annotated with the keyword. Meanwhile, our approach rank these returned images according to the confidence score of the keyword. The experimental results are presented in **Table 4** our approach significantly outperforms the other four methods. More specifically, our approach achieves an improvement up to 19% in MAP on all 260 keywords over CRM-Rectangles. When compared with MBRM, the proposed approach has an improvement of 3% on all 260 words. For the keywords which have positive recall, our approach also exhibits superior performance, achieving a gain of 23% and 6% over CRM-Rectangles and MBRM respectively. The effective image annotation performance of our approach directly leads to these significant improvements over other approaches we compared with in the image retrieval task. The similarity measure based SSVM in the image annotation procedure makes sure that the keywords we assigned to an image is optimized, which improves the effectiveness of our keyword-based image retrieval.

Table 5. Performance comparison of semantic image retrieval results on Corel30k

Algorithms	All words	Words (recall＞0)
Mean average precision on corel30k		
PLSA-WORDS [33]	0.14	0.17
GM-PLSA [31]	0.23	**0.28**
Ours	**0.25**	**0.28**

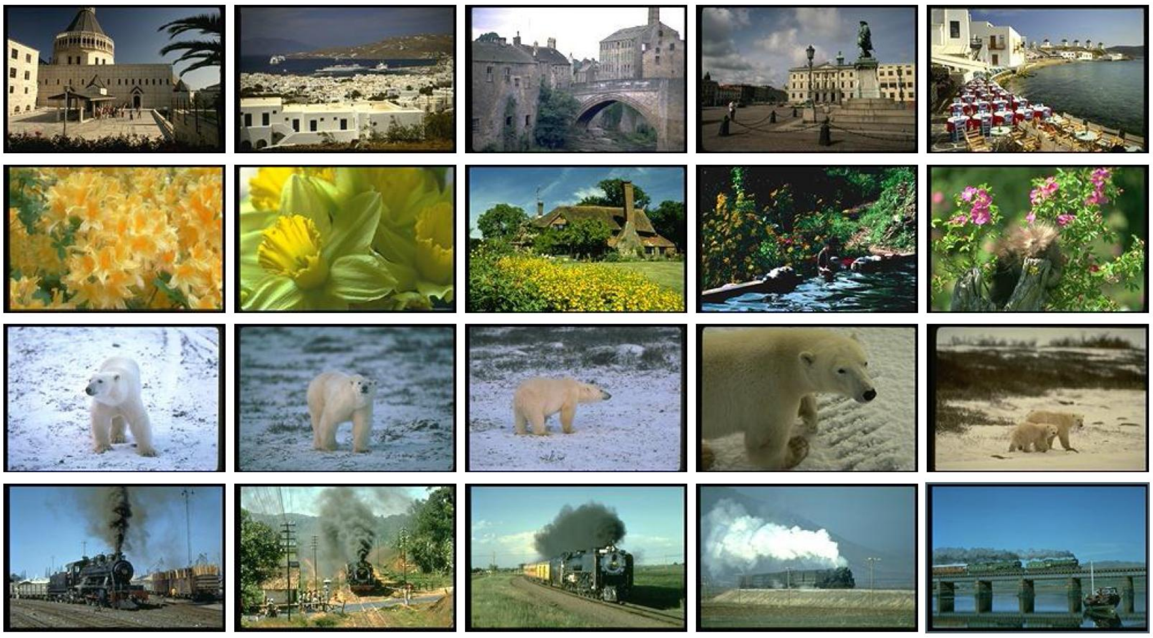


Fig. 7. Retrieval results using the proposed approach on Corel dataset.

We further experiment on Corel30k for image retrieval. Compared with Corel5k, Corel30k is relative larger. So far, only a few approaches have experimented on this data set [31]. In this experiment, we compare our approach with PLSA-WORDS [33] and GM-PLSA [31] For semantic image retrieval. The results are reported in **Table 5**, where the proposed approach significantly outperforms PLSA-WORDS, no matter on all 950 keyword sets or on the keyword sets with positive recall, and is competitive with GM-PLSA. These convincing results are consistent with that on Corel5k, which validates that our approach can adapt to data and exploit more information hidden in the images. Fig. 7 presents some retrieval results using our approach with several keywords as queries. Each row presents the first five retrieved images towards a query. From top to bottom, the semantic queries are *buildings*, *flower*, *bear and railway*. The diverse visual appearance of the returned images demonstrates that the proposed approach has a good generalization ability.

5. Conclusions

In this paper, we propose a new approach for automatic image annotation and retrieval. The approach learns similarity metric from data sets and embeds it to both ranking and annotation procedures. Inspired by the observation that ranking and annotation are highly dependent each other, we take the two procedures into account when learning similarity metric, i.e., simultaneously optimizing the objective function for ranking (using structural SVM) and annotation (using correlated label propagation) with respect to similarity metric. The collaborative metric learning method fully exploits ranking and annotation information, and can be effectively solved using cutting plane. Specifically, in the annotation procedure, we employ correlated label propagation to utilize the correlation information among the labels. To evaluate the proposed approach, we extensively experiment on three large data sets. The results show its competitive performance in both annotation and retrieval.

References

[1] M. Jaber and E. Saber, "Probabilistic approach for extracting regions of interest in digital images," *Journal of Electronic Imaging*, vol. 19, 2010.
Article (CrossRef Link)

[2] S. Zhang, J. Huang, H. Li and D. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 99, pp.1–12, 2012.
Article (CrossRef Link)

[3] D. Putthividhy, H. Attias and S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3408–3415, *2010.*
Article (CrossRef Link)

[4] O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proc. of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD*, pp. 1–7, 2008.
Article (CrossRef Link)

[5] D. Grangier and S. Bengio, "A discriminative kernel-based model to rank images from text queries," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 30, no. 8, pp. 1371–1384, 2008.
Article (CrossRef Link)

[6] C. Cusano, G. Ciocca and R. Schettini, "Image annotation using svm," in *Proc. of Internet imaging IV*, vol. SPIE 5304, 2004.
Article (CrossRef Link)

[7] S. Hoi,W. Liu and S. Chang, "Semi-supervised distance metric learning for collaborative image retrieval," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2008.
Article (CrossRef Link)

[8] H. Cheng, Z. Liu and J. Yang, "Sparsity induced similarity measure for label propagation," in *Proc. of IEEE International Conference on Computer Vision,* pp. 317–324, 2009.
Article (CrossRef Link)

[9] T. Mei, Y. Wang, X. Hua, S. Gong and S. Li, "Coherent image annotation by learning semantic distance," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
Article (CrossRef Link)

[10] C. Yang, M. Dong and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2057–2063, 2006.
Article (CrossRef Link)

[11] T. Joachims, T. Finley and C. Yu, "Cutting plane training of structural svms," *Machine Learning,* vol. 77, no. 1, pp. 27–59, 2009.
Article (CrossRef Link)

[12] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: constraining the latent space," in *Proc. of ACM International Conference on Multimedia, ACM*, pp. 348–351, 2004.
Article (CrossRef Link)

[13] G. Carneiro, A. Chan, P. Moreno and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transaction on Pattern Analysis and Machine Intelligenc, vol.* 29, no. 3, pp. 394–410, 2007.
Article (CrossRef Link)

[14] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 119–126, 2003.
Article (CrossRef Link)

[15] S. Feng, R.Manmatha and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Article (CrossRef Link)

[16] V. Lavrenko, R. Manmatha and J. Jeon, “A model for learning the semantics of pictures,” *Advances in Neural Information Processing Systems*, 2003.

[17] N.Loe and A. Farhadi, “Scene discovery by matrix factorization,” in *Proc. of European Conference on Computer Vision,* pp. 451–464, 2008.
Article (CrossRef Link)

[18] H. Zhang, A. Berg, M. Maire and J. Malik, “Svm-knn: Discriminative nearest neighbor classification for visual category recognition,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2126–2136, 2006.
Article (CrossRef Link)

[19] J. Liu, M. Li, Q. Liu, H. Lu and S. Ma, “Image annotation via graph learning,” *Pattern recognition*, vol. 42, no. 2, pp. 218–228, 2009.
Article (CrossRef Link)

[20] J. Pan, H. Yang, C. Faloutsos and P. Duygulu, “Automatic multimedia cross-modal correlation discovery,” in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 653–658, 2004.
Article (CrossRef Link)

[21] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid, “Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation,” in *Proc. of IEEE International Conference on Computer Vision*, pp. 309–316,2009.
Article (CrossRef Link)

[22] A. Makadia, V. Pavlovic and S. Kumar, “A new baseline for image annotation,” in *Proc. of European Conference on Computer Vision*, 2008.
Article (CrossRef Link)

[23] B. McFee and G. Lanckriet, “Metric learning to rank,” in *Proc. of International Conference on Machine Learning*, 2010.

[24] T. Joachims, “A support vector method for multivariate performance measures,” in *Proc. of International Conference on Machine learning*, pp. 377–384, 2005.
Article (CrossRef Link)

[25] F. Kang, R. Jin and R. Sukthankar, “Correlated label propagation with application to multi-label learning,” in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 1719–1726, 2006.
Article (CrossRef Link)

[26] P. Duygulu and K. Barnard, “Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in *Proc. of European Conference on Computer Vision*,” pp. 97–112, 2002.
Article (CrossRef Link)

[27] B.Wang, Y. Shen and Y. Liu, “Integrating distance metric learning into label propagation model for multi-label image annotation,” in *Proc. of IEEE Conference Image Processing*, 2011.
Article (CrossRef Link)

[28] J. Tang, H. Li, G. Qi and T. Chua, “Image annotation by graph-based inference with integrated multiple/single instance representation,” *IEEE Transactions on Multimedia*, vol. 12, no. 2, pp. 131–141, 2010.
Article (CrossRef Link)

[29] K. Van De Sande, T. Gevers and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32 (9) (2010) 1582–1596.
Article (CrossRef Link)

[30] C. Wang, S. Yan, L. Zhang and H. Zhang, “Multi-label Sparse Coding for Automatic Image Annotation,” in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 1643–1650, 2009.
Article (CrossRef Link)

[31] Z. Li, Z. Shi, X. Liu and Z. Shi, “Modeling continuous visual features for semantic image annotation and retrieval, ” *Pattern Recognition Letters*, vol. 32, no. 3, pp. 516–523, 2010.
Article (CrossRef Link)

[32] D. Metzler and R. Manmatha, "An inference network approach to image retrieval," *Image and Video Retrieval,* vol. 3115*,* pp 42-50, 2004.
Article (CrossRef Link)

[33] Z. Li, Z. Shi, X. Liu, Z. Li and Z. Shi, "Fusing semantic aspects for image annotation and retrieval, " J*ournal of Visual Communication and Image Representation,* vol. 21, no.8, pp. 798–805, 2010.
Article (CrossRef Link)

[34] M. Fukui, N. Kato and W. Qi, "Multi-class labeling improved by random forest for automatic image annotation," in *Proc. of IAPR Conference on Machine Vision Applications*, pp. 202–205, 2011.

[35] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. L and D. Metaxas, "Automatic image annotation using group sparsity," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3312–3319, 2011.
Article (CrossRef Link)

Bin Wang received her BE degree in information and computional science from Shandong Normal University, Ji'nan, China, in 2006; and the MS degree in applied mathmatics from University of Science and Technology Beijing, Beijing, China. She is currently a PhD candidate at Department of Automation, Shanghai Jiao Tong University, Shanghai, China. Her research interests include computer vision, machine learning, image processing, multimedia analysis.

Yuncai Liu received the Ph.D. degree in the Department of Electrical and Computer Science Engineering in 1990 from the University of Illinois at Urbana-Champaign (UIUC), and worked as an associate researcher at the Beckman Institute of Science and Technology from 1990 to 1991. Since 1991, he had been a system consultant and then a chief consultant of research in Sumitomo Electric Industries, Ltd., Japan. In October 2000, he jointed the Shanghai Jiao Tong University as a distinguished professor. His research interests are in image processing and computer vision.