Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 29 (2008) 407–415

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A graph-based image annotation framework

Jing Liu [a,*], Bin Wang [b], Hanqing Lu [a], Songde Ma [a]

[a] *Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China*
[b] *University of Science and Technology of China, Hefei 230027, China*

Received 31 May 2007; received in revised form 5 September 2007
Available online 4 November 2007

Communicated by M.-J. Li

Abstract

Automatic image annotation is crucial for keyword-based image retrieval because it can be used to improve the textual description of images. In this paper, we propose a unified framework for image annotation, which contains two kinds of learning processes and incorporates three kinds of relations among images and keywords. In addition, we propose some improvements on its components, i.e. a reinforced image-to-image relation; a combined word-to-word relation; and a progressive learning method. Experiments on the Corel dataset demonstrate their effectiveness. We also show that many existing image annotation algorithms can be formulated into this framework and present an experimental comparison among these algorithms to evaluate their performance comprehensively.

Keywords: Image annotation; Annotation refinement; Graph learning; Image relation; Word relation

1. Introduction

With the advent of digital imagery, the number of images has been growing rapidly and there is an increasing need for effectively indexing and searching these images. Systems using non-textual (image) queries have been proposed but many users found it hard to represent their queries using abstract image features. Most users prefer textual queries, i.e. keyword-based image search, which is typically achieved by manually providing image annotations and searching over these annotations using a textual query. However, manual annotation is an expensive and tedious procedure. Thus, automatic image annotation is necessary for efficient image retrieval.

Many algorithms have been proposed for automatic image annotation. In a straightforward way, each semantic keyword or concept is treated as an independent class and corresponds to one classifier. Methods like linguistic indexing of pictures (Li and Wang, 2003), image annotation using SVM (Cusano and Schettini, 2003) and Bayes point machine (Chang et al., 2003) fall into this category. Other methods try to learn a relevance model associating images and keywords. The early work in (Duygulu et al., 2002) applied a translation model (TM) to translate a set of blob tokens (obtained by clustering image regions) to a set of annotation keywords. Jeon et al. (2003) assumed that image annotation could be viewed as analogous to the cross-lingual retrieval problem and proposed a cross-media relevance model (CMRM). Lavrenko et al. (2003) proposed continuous-space relevance model (CRM) which assumed that every image is divided into regions and each region is described as a continuous-valued feature vector. Given a training set of images with annotations, a joint probabilistic model of image features and words is estimated. Then the probability of generating a word given the image regions can be predicted. Compared with the CMRM, the CRM directly models continuous features, so it does not rely on clustering and consequently avoids the granularity issues. Feng et al. (2004) proposed another relevance model in which a multiple Bernoulli model is used to generate words instead of the multinomial one as

[*] Corresponding author.
E-mail address: liujingmgm@gmail.com (J. Liu).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.10.018

in CRM. Recently, there are some efforts considering the word correlation in the annotation process, such as coherent language model (CLM) (Jin et al., 2004), correlated label propagation (CLP) (Kang et al., 2006), annotation refinement using random walk (Wang et al., 2006) and WordNet-based method (Jin et al., 2005). The graph-based methods also achieved much attention. Pan et al. (2004) firstly proposed a graph-based automatic caption method, in which images, annotations and regions are considered as three types of nodes to construct a mixed media graph so as to perform image annotation. In our previous work (Liu et al., 2006), we proposed an NSC-based method to calculate image similarities on visual features and propagated annotations from training images to their similar test images.

As these algorithms seem to be so different from each other, it is not easy to answer such questions as which models are better, what the connections among them are, and how they should be utilized. In this paper, we conduct a formal study on these issues and find that previous research work can be induced as two kinds of learning processes, which integrate three kinds of relations as shown in Fig. 1: image-to-image relation, word-to-word relation, and image-to-word relation.

We propose a unified framework for image annotation. In the framework, automatic image annotation can be performed across two graph learning processes. The first process (referred as "basic image annotation") aims to obtain the preliminary annotations for each untagged image. It is a learning process on an image-based graph, whose nodes are images and edges are relations between images. The second process (referred as "annotation refinement") aims to refine the candidate annotations obtained from the prior process. It is a word-based graph learning process, where the nodes are words and the edges are relations between words.

The proposed framework allows us to analyze and understand some previous work more clearly, and offers some potential research guidance. In this paper, we propose three improvements on different parts of the framework. First, considering the intra-relations among training images and test images, we propose a reinforced inter-relation between training image and test image. Second, a combined word correlation is designed as a comprehensive estimation, in which not only the statistical distribution in the training dataset, but also the visual-content-based measurement within the context of web are considered. Third, a progressive learning method is proposed to perform image annotation in a greedy manner, while the traditional assumption of word independence for an image is relaxed to the conditional independence. To evaluate the performance of these improvements, we carry out several experiments on benchmark data of Corel images. Besides, we give a systematic comparison among some related work. Exciting performance of our scheme and some consistent conclusions with the theoretical results are achieved under the proposed framework.

The rest of the paper is organized as follows. Section 2 introduces the unified framework of image annotation. Some improvements based on the proposed framework are addressed in Section 3. Section 4 presents the implementation of image annotation with the proposed improvements. Section 5 presents experimental comparisons among several related work and our scheme. Conclusions and future work are given in Section 6.

2. Image annotation framework based on graph learning

The proposed framework consists of two learning processes denoted as "basic image annotation" and "annotation refinement", and three kinds of relations as mentioned above. In the basic image annotation process, image-to-image relation and image-to-word relation are integrated to obtain the candidate annotations. In the annotation refinement process, the word-to-word relation is explored to refine those candidate annotations from the prior process. The both learning processes are performed sequentially. An overview of the framework is shown in Fig. 2.

2.1. Basic image annotation

The basic image annotation can be deemed as a semi-supervised learning process on an image-based graph, i.e. propagating labels (annotations) from annotated images to un-annotated images according to their similarities.

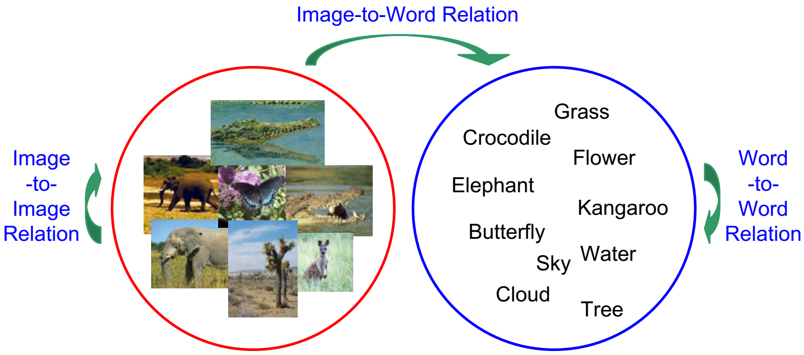


Fig. 1. Illustrative example of image annotation.

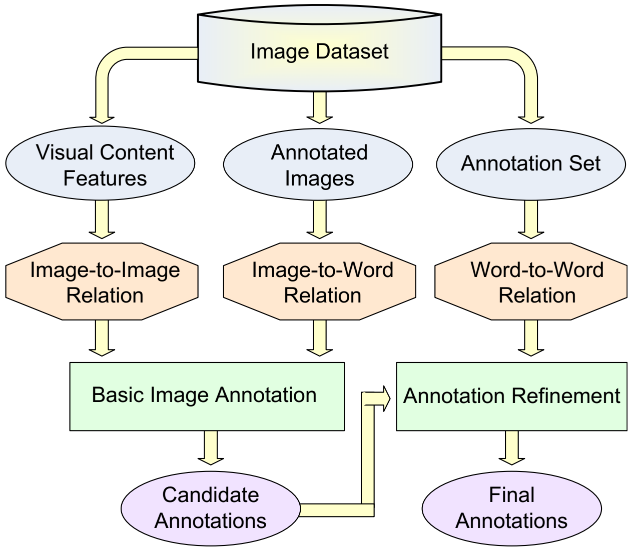


Fig. 2. Overview of image annotation framework.

The learning process includes two key issues: one is how to measure the relations among images (S_{II}), especially the train-to-test image relation, and the other is how to model the distribution of words in annotated images (S_{IW}). Then we formulate it as follows:

$$T^{N_{\mathrm{S}}\times M} = G_1\left(S_{\mathrm{II}}^{N_{\mathrm{S}}\times N_{\mathrm{T}}}, S_{\mathrm{IW}}^{N_{\mathrm{T}}\times M}\right) \tag{1}$$

where N_{T} is the number of training images, N_{S} is the number of test images, M is the size of vocabulary, and $G_1(\cdot)$ is a certain graph learning algorithm, for which the simplest form is given as follows:

$$T = S_{\mathrm{II}} \cdot S_{\mathrm{IW}} \tag{2}$$

Other graph learning methods have been exploited in literatures. The manifold ranking method (Tong et al., 2006) is an example. Generative models (Blei and Jordan, 2003), such as GMM, LSA/pLSA, and LDA, also show good performance in image annotation. Therefore, $G_1(\cdot)$ as Eq. (1) have several variances. It is expected that a better learning method can present superior performance.

In the following, we give a probabilistic analysis for the basic annotation. Usually, the relation between a set of words W and a test image I is defined as the conditional probability $P(W|I)$, which can be calculated as the expectation on the training set T:

$$P(W|I) = \sum_{J\in T} P(W|J)P(J|I) \tag{3}$$

where $P(W|J)$ denotes the probability of W given training image J, i.e. image-to-word relation, and $P(J|I)$ denotes the image-to-image relation.[1] Equivalently, we can calculate the joint probability of words (or a word) and an image:

$$P(W,I) = \sum_{J\in T} P(W,I|J)P(J) \tag{4}$$

Typically, an assumption is made that the test image is independent of the words given a training image, that is

$$P(W,I|J) = P(W|J)P(I|J) \tag{5}$$

So Eq. (4) can be written as

$$P(W,I) = \sum_{J\in T} P(W|J)P(I|J)P(J) = \sum_{J\in T} P(W|J)P(I,J) \tag{6}$$

Then the annotation process becomes the maximization of the joint probability or the conditional probability:

$$W^* = \arg\max_{W\subset V} P(W|I) = \arg\max_{W\subset V} P(W,I) \tag{7}$$

This is a simple learning manner adopted by some relevance model based work, such as CMRM, CRM, MBRM and so on, in which the label information on the training images $P(W|J)$ is propagated to the test image (J) according to the inter relation $P(I,J)$ or $P(I|J)$.

2.2. Annotation refinement

Annotation refinement is the second process in the framework. The candidate annotations obtained from the prior process may be unsatisfactory without considering word correlation efficiently. Thus annotation refinement is recognized as a beneficial process to reserve highly correlated annotations and remove irrelevant ones by exploring the word correlations. Actually, the process is performed on a word-based graph. Similarly, the learning in this phase can be denoted as

$$T_{\mathrm{R}}^{N_{\mathrm{S}}\times M} = G_2\left(S_{\mathrm{IW}}^{N_{\mathrm{S}}\times M}, S_{\mathrm{WW}}^{M\times M}\right) \tag{8}$$

where S_{WW} is the word-to-word relation matrix and $G_2(\cdot)$ is a certain graph learning algorithm. Above discussion about $G_1(\cdot)$ is also applicable to $G_2(\cdot)$. The process of annotation refinement can also be expressed in a probabilistic form as follows:

$$P(W|I) = \sum_{v\subset V} P(W|v)P(v|I) \tag{9}$$

where $P(W|v)$ denotes the word correlation and $P(v|I)$ is the prior confidence provided by the basic image annotation.

3. Improvements under the framework

Above analysis about the proposed framework demonstrates that the problem of image annotation can be decomposed into several specific and well defined subproblems. Specially, three kinds of relations and two graph learning processes are referred to. We can improve these items and expect their combination to enhance the overall performance. In the following, we present our improvements focusing on some items in the framework.

[1] Typically, $P(I,J)$ indicates the joint probability between images I and J, which can be estimated as the visual similarity between both images, i.e. image-to-image relation. Assuming that $P(J)$ is a constant, which indicates a uniform distribution, we can roughly regard $P(I|J) = P(I,J)/P(J)$ as the image-to-image relation too.

3.1. Improvement on image-to-image relation

Many previous work only consider the relation between test image and training image, i.e. $P(I|J)$ as in Eq. (6), while the relations within the training images and the relations within the test images are ignored. Based on this view, we propose a method to calculate the image correlation with all available information and make them reinforce each other. Given a training image set and a test image set, there are four kinds of relations: intra-type training image relation (S_{t}), intra-type test image relation (S_{s}), inter-type train-to-test image relation (S_{ts}), and inter-type test-to-train image relation (S_{st}), as illustrated in Fig. 3. The image similarity is usually a symmetric measure. However, in many region-based methods like CRM and MBRM, the image relation is asymmetric as in Eq. (10), which is also our selection in this paper. Thus, we need define S_{ts} and S_{st} separately. Besides, S_{t} and S_{s} may be asymmetric:

$$\begin{aligned} S(I_m, I_n) &= \exp(-\sigma_1 \cdot D(I_m, I_n)) \\ &= \exp\left(-\sigma_1 \cdot \sum_{i,j} D(r_i^m, r_j^n)\right) \end{aligned} \tag{10}$$

where $S(I_m, I_n)$ indicates the similarity between images I_m and I_n, r_i^m is the ith region of I_m, $\sigma_1 > 0$ is a smoothing parameter, and $D(\cdot)$ is certain distance measure. L_1-distance is selected in our implementation due to its better performance (Stricker and Orengo, 1995). Intuitively, the image relation can be reinforced by their "related" images. For example, there are two images t_1 and t_2 in training set and each has one close neighbor in test set denoted as s_1 and s_2, respectively. If s_1 and s_2 are very similar, the similarity between t_1 and t_2 should also be enhanced. Similar discussion is also valid for other types of relations. Thus the four types of relations are dependent and interactive. We can alternatively update them in an iterative way as follows:

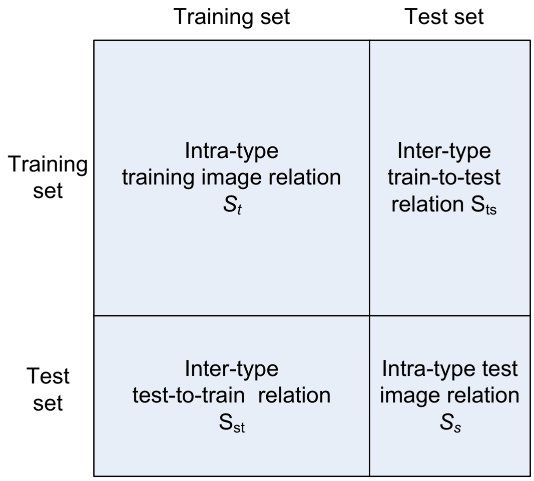
$$S_{\mathrm{t}} = \alpha_1 S_{\mathrm{t}} + (1 - \alpha_1) S_{\mathrm{ts}} \cdot S_{\mathrm{s}} \cdot S_{\mathrm{st}} \tag{11}$$

$$S_{\mathrm{s}} = \alpha_2 S_{\mathrm{s}} + (1 - \alpha_2) S_{\mathrm{st}} \cdot S_{\mathrm{t}} \cdot S_{\mathrm{ts}} \tag{12}$$

$$S_{\mathrm{st}} = \alpha_3 S_{\mathrm{st}} + (1 - \alpha_3) S_{\mathrm{s}} \cdot S_{\mathrm{st}} \cdot S_{\mathrm{t}} \tag{13}$$

$$S_{\mathrm{ts}} = \alpha_4 S_{\mathrm{ts}} + (1 - \alpha_4) S_{\mathrm{t}} \cdot S_{\mathrm{ts}} \cdot S_{\mathrm{s}} \tag{14}$$

where $\alpha_i \in [0,1]\,(i = 1,2,3,4)$ is the weight to determine the role of the reinforcement. Finally, S_{ts} can be considered as an improved inter-relation for the label propagation from training images to test images.



	Training set	Test set
Training set	Intra-type training image relation S_t	Inter-type train-to-test relation S_{ts}
Test set	Inter-type test-to-train relation S_{st}	Intra-type test image relation S_s

Fig. 3. Four types of image relations.

3.2. Improvement on word-to-word relation

As mentioned above, word-to-word relation is required in the process of annotation refinement. Intuitively, good word correlation ensures the good performance of annotation refinement. In this section, we present a combined correlation with two kinds of correlation measures, i.e. a statistical correlation by co-occurrence in the training set and a content-based word correlation by web search, to estimate the semantic relatedness between words comprehensively.

- *Statistical correlation from training set:* Generally, two words with high co-occurrence in the training set will possibly joint to annotate certain image, such as 'cloud' and 'sky', 'animal' and 'frog'. Therefore, the word co-occurrence becomes an informative representation of word relatedness.
  We get the count of each word-pair as annotations of the same image and obtain the co-occurrence based measure. Usually, the more general a word is, the larger chance it will have to associate with other words to annotate the same image. However, such associations usually have low confidence. Thus, we weight the count according to the frequency of each word, i.e. setting a low weight to a frequent word and a high weight to a rare word. The weighted co-occurrence (K_{SC}) as a statistical word correlation can be calculated as follows:

$$K_{\mathrm{SC}}(v_1, v_2) = K_{\mathrm{C}}(v_1, v_2) \times \log\left(\frac{N_{\mathrm{T}}}{n_i}\right) \tag{15}$$

  where $K_{\mathrm{C}}(v_1, v_2)$ is the number of co-occurrence for word v_1 and v_2, n_1 is the count for v_1 occurring in the training images, and N_{T} is the total number of training images. Note that $K_{\mathrm{SC}}(v_1, v_2)$ may be unequal to $K_{\mathrm{SC}}(v_2, v_1)$. When v_1 (a specific or rare word) occurs, v_2 (a common or frequent word) may have a high probability to occur, but not necessarily vice versa. This is best illustrated by using an example. Considering two words 'animal' and 'frog', we can easily infer 'animal' from 'frog', but not vice versa.
- *Content-based correlation by search:* Word co-occurrence is a locally statistical word correlation dependent on the training set. To get a more robust measure, we should seek other entrance to enrich the estimation of word correlation. Since image is the focus of image annotation,

visual-content as a direct representation of image should also be contributed to the word correlation. Besides, web represents the largest public available corpus with aggregate statistical and indexing information. Such huge and useful resources deserve to our attention. Then, a content-based correlation by web search is designed.
Given a keyword query, image search engines like Google usually return good search results, especially those on the first page. Thus, top-ranked images can be roughly treated as the visual representation of the query word. Then the visual similarity between two resulting sets can be used to estimate the semantic relation between corresponding query words. That is, the estimation of word correlation converts to the measurement of similarity between two sets of visual feature vectors. Here we adopt a simple strategy to calculate the similarity between both vector sets, which is given as follows:

$$K_{\mathrm{CCS}}(v_1,v_2) = S_{\mathrm{set}}(I(v_1),I(v_2)) = \sum_{m,n=1}^{M} S_I(I_m(v_1),I_n(v_2)) \tag{16}$$

where $K_{\mathrm{CCS}}(v_1,v_2)$ indicates the content-based correlation between words v_1 and v_2, $I(v_1)$ and $I(v_2)$ indicate the resulting image sets corresponding to words v_1 and v_2, respectively, $I_m(v_1)$ is the mth image in the image set $I(v_1)$, $S_{\mathrm{set}}(\cdot)$ is the visual similarity between both sets, $S_{\mathrm{I}}(\cdot)$ is the visual similarity between two images as Eq. (10), and M is the number of images from the resulting set (Top 10 images are selected in our implementation).

- *Combined word correlation* Now we get two types of word correlations. Statistical co-occurrence describes the word distribution based on the training set. Although it provides more precise statistical description, it depends on the training data. The content-based correlations by web search represents the word relatedness in the web context. Although the distribution of web data is universal and independent on any corpus, various noises are inevitable. Therefore, they should be jointed to complement each other. We normalize them into $[0,1]$ firstly and combine them in a linear form:

$$S_{\mathrm{WW}} = \varepsilon_1 K_{\mathrm{SC}} + (1-\varepsilon_1)K_{\mathrm{CCS}} \tag{17}$$

where $\varepsilon_1 \in [0,1]$.

3.3. Improvement on learning method

Usually, one image is annotated with several annotation words. Then the images annotation problem is formulated as maximizing the probability of a set of words W given an image I as follows:

$$W: \quad \{w_1,w_2,\ldots,w_n\} = \arg\max_{W\subset V} P(W|I) = \arg\max_{w_i\in V} P(w_1,w_2,\ldots,w_i,\ldots,w_n|I) \tag{18}$$

In previous work, they assume the predication of annotations for an image is independent from one word to another and make the joint probabilities of multiple words be factorized as

$$P(W|I) = P(w_1|I)P(w_2|I)\cdots P(w_n|I) \tag{19}$$

In the process of annotation, the probability of each word given image I is estimated individually and those words with larger probabilities are selected to annotate image I.

However, such an process makes the model only captures the correlation between word and image, while the word correlation is not considered. However, direct estimation of $P(W|I)$ is computationally prohibitive due to the problem of combination explosion brought with the large size of the vocabulary. To avoid the expense in computation and yet utilize the word correlation in predicating annotations, we adopt a greedy solution to the problem. Specially, the model is factorized as follows:

$$P(W|I) = P(w_1|I)P(w_2|w_1,I)\cdots P(w_n|w_1,w_2,\ldots,w_{n-1},I) \tag{20}$$

According to Eq. (20), we can annotate images in a progressive manner. That is, the annotation prediction is performed multiple times. In the first run, only the most probable word w_1 is chosen. In the second run, the conditional probability $P(w_2|w_1,I)$ instead of $P(w_2|I)$, is used to decide the optimal word w_2. The similar process is repeated until the desired annotation length is reached or the probabilities of left words are below a certain threshold. Actually, most of relevance models can be easily incorporated into the progressive learning process and perform image annotation effectively.

4. Implementation details

4.1. Basic image annotation

Obtaining the reinforced relation S_{ts} in Section 3.1, we prepare the image correlation (S_{II}) for the basic image annotation. The image-to-word relation (S_{IW}) based on the training set is another key part. Here, we select the Multi-Bernoulli model to model the word distribution as (Feng et al., 2004)

$$S_{\mathrm{IW}}(i,j) = P(w_j|I_i) = \frac{\mu\delta_{w_j,I_i} + N_{w_j}}{\mu + N_{\mathrm{T}}} \tag{21}$$

where $S_{\mathrm{IW}}(i,j)$ indicates the probability of the word w_j given the image I_i, μ is a smoothing parameter estimated by the cross-validation, $\delta_{w_j,I_i} = 1$ if the word w_j occurs in the annotations of image I_i and $\delta_{w_j,I_i} = 0$ otherwise, N_{w_j} is the number of training images annotated with w_j.

We select Eq. (2) as the graph learning model and perform basic image annotation through the progressive learning process (as discussed in Section 3.3) to obtain the preliminary annotation result. The result provides a guideline to select the candidate annotations and further gives a prior confidence for every candidate one, which is required by the annotation refinement.

4.2. Annotation refinement

As we know, annotation refinement can be regarded as a re-ranking process for candidate annotations. To re-rank these annotations, the graph ranking algorithm (Zhou et al., 2003) is used to leverage both the word correlations and the prior confidence of each word. We carry out the iteration of the following Eq. (22) until it coverage to obtain the final re-ranking matrix T_{R}.

$$T_{\mathrm{R}}^{t+1} = \beta \cdot T_{\mathrm{R}}^{t} \cdot S_{\mathrm{WW}} + (1-\beta)T \tag{22}$$

where T is the resulting matrix from the prior process, S_{WW} is the combined word correlation obtained in Section 3.2, and β is the weight to regulate the role of annotation refinement using the word correlation. Considering that words do not have the transitive characteristic strictly, we usually carry out the iteration for a few times (1–3 times) to obtain the refined result.

5. Experiments

5.1. Experimental design

To present a fair comparison with some previous work, we use the Corel dataset provided by Duygulu et al. (2002) without any modification. The dataset contains 5000 images. Each image is segmented into 1–10 regions. A 36-dimensional feature for each region is extracted, which includes color, texture and area features as in (Duygulu et al., 2002). All the regions are clustered into 500 clusters (called as blobs). Each image is annotated with 1–5 words. The total number of words is 371. The dataset is divided into two parts: 4500 images for training and rest 500 for test.

Similar to previous work, the quality of automatic image annotation is evaluated through the process of retrieving test images with single keyword. For each keyword, the number of correctly annotated images is denoted as N_{c}, the number of retrieved images is denoted as N_{r}, and the number of truly related images in test set is denoted as N_{t}. Then the precision, the recall and the $F1$ measure are computed as follows:

$$precision(w) = \frac{N_{\mathrm{c}}}{N_{\mathrm{r}}}, \quad recall(w) = \frac{N_{\mathrm{c}}}{N_{\mathrm{t}}} \tag{23}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{24}$$

We further average the values of the precision and the recall, respectively over all the words in test set (260 words) to evaluate the performance. Besides, we give a measure to evaluate the coverage of correctly annotated words, i.e. the number of words with non-zero recall, which is denoted as "*NumWord*" for short. The measure is important because a biased model can achieve high precision and recall values by only performing quite well on a small number of common words.

In the following, we will present a series of comparisons among some related work focusing on different components in the proposed framework: image-to-image relation, image-to-word relation, word-to-word relation, and the learning method. By the cross-validation, the parameters in our scheme are set as follows: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.05$, $\beta = 0.30$, $\mu = 7.33$, $\varepsilon_1 = 0.50$.

5.2. Comparison on image-to-image relation

Firstly, we present the comparison about the image-to-image relation. When the reinforced image relation as in Section 3.1 is incorporated to CRM and MBRM, the improved models are denoted as IIR_CRM and IIR_MBRM, respectively. All the methods listed in Table 1 are divided into two groups for analyzing convenience:

CMRM and CRM: Both methods adopt the probabilistic relevance model between images and words to perform basic image annotation. However, they differ in the representation of visual feature. CRM uses the continuous region features to calculate the image similarity, while CMRM uses blob histograms. Because the blobs are obtained by clustering the region features, much information has been lost. Therefore, CRM can better reflect the image relation than CMRM. This difference results in the large gap in their performances. CRM achieves noticeable improvement. The *NumWord* of CRM is 107, much larger than 66 in CMRM. The precision, recall and $F1$-measures of CMRM are almost doubled in CRM.

CRM and IIR_CRM (*MBRM and IIR_MBRM*)*:* The methods in the group adopt similar learning model and they all use the region-based visual features.[2] However, they differ in the estimation of the image-to-image relation. CRM adopts the typical method to estimate the relation as Eq. (10), while IIR_CRM exploits the improved image relation, which effectively considers the intra-relation among test images and training images to reinforce the original inter-relation. As a result, the $F1$-measure is boosted to 0.23, and the *NumWord* is further increased to 118 words. Similar improvement can also be seen from the comparison between MBRM and IIR_MBRM.

Therefore, the better image-to-image relation is beneficial to image annotation. Obviously, the reinforced image-to-image relation achieves the best performance in the comparison.

[2] In order to present a fair comparison, we use the features of segmented regions rather than those of rectangles. Thus, the results of MBRM (Feng et al., 2004) and CLP (Kang et al., 2006) in this paper may be different from their original papers using the rectangular features.

Table 1
Performance comparison on image-to-image relation

Model	Precision	Recall	$F1$	NumWords
CMRM	0.09	0.10	0.09	66
CRM	0.16	0.19	0.17	107
IIR_CRM	0.20	0.23	0.21	118
MBRM	0.22	0.24	0.23	121
IIR_MBRM	0.23	0.25	0.24	122

5.3. Comparison on image-to-word relation

In Table 2, we present the comparison among some related methods, which differ in the image-to-word relation. The discussion about the comparison is also carried out on two groups:

CRM and MBRM: In the comparison, MBRM has the *NumWord* of 121 compared to 107 in CRM. The increased *NumWord* in turn enhances precision, recall and $F1$-measure. The improvement is largely attributed to that MBRM adopts multiple Bernoulli distribution to replace the multinomial distribution in CRM. Actually, it is because that MBRM provides a solution based on the multi-label learning instead of a multi-class one as CRM. Similar results can also be observed from the comparison between IIR_MBRM and IIR_CRM.

MBRM and CLP: They both provide the solutions within the context of the multi-label learning, while CLP employs a more sophisticated form. In the label propagation, CLP give more chance to rare word and relatively weaken the bias to common words, while MBRM suffer from the bias. Accordingly, CLP achieves wider coverage of correctly annotated words, i.e. more gains on *NumWord* and recall than MBRM.

Better performance of MBRM and CLP than CRM implies that the formulation of image annotation as a multi-label learning problem is really preferable to as a multi-class learning problem.

5.4. Comparison on word-to-word relation

In the following, we make a comparison among various word correlations, in which WordNet-based correlation (WNC) (Budanitsky and Hirst, 2001), statistical correlation by co-occurrence (SC), and content-based correlation by search (CCS) are considered. For the WNC, we adopt the measure proposed by Jiang and Conrath (1997), which is demonstrated to be a more effective measure based on WordNet (Budanitsky and Hirst, 2001).

When candidate annotations are provided by MBRM, we present a comparison among MBRM, MBRM + WNC, MBRM + SC, MBRM + SCS, MBRM + CCS, and MBRM + (SC − CCS), as listed in Table 3. Compared with MBRM, the gains on $F1$-measure are 3.0% (MBRM + SC), 4.8% (MBRM + CCS), and 8.3% (MBRM + (SC − CCS)), respectively, but MBRM + WNC gets a poorer performance. Considering each word correlation individually, their different roles on the performance improvement can be observed. First, the statistical correlation by co-occurrence, i.e. SC, gains obvious improvement on the measure of *NumWord*, but it losses on the average precision. This indicates that the method is capable of connecting more words through the statistical information, but the connections cannot ensure the relatedness on the semantic level. Second, CCS achieves overall improvements on three measures. This is because that the content-based correlation by search is estimated in the web context and provide the word relatedness from a more general and reasonable level indeed. Third, WNC shows the worst performance. Specially, the WordNet-based correlation takes a negative role through the annotation refinement. There are 49 words out of 371 words in the Corel dataset that either do not exist in the lexicon of the WordNet or have no available relations with other words. The sparse relation largely weakens the performance of WNC. Thus we do not consider the addition of WNC in our combined correlation. Finally, the combination of SC and CCS achieves the best performance. It shares the advantages from both correlations and gives a relatively precise and comprehensive representation of word semantic relatedness.

Table 2
Performance comparison on image-to-word relation

Model	Precision	Recall	$F1$	NumWords
CRM	0.16	0.19	0.17	107
MBRM	0.22	0.24	0.23	121
CLP	0.21	0.26	0.23	125
IIR_CRM	0.20	0.23	0.21	118
IIR_MBRM	0.23	0.26	0.24	122

Table 3
Performance comparison on word-to-word relation based on MBRM

Model	Precision	Recall	$F1$	NumWords
MBRM	0.218	0.243	0.230	121
+WNC	0.208	0.221	0.219	117
+SC	0.215	0.265	0.237	126
+CCS	0.230	0.253	0.241	124
+SC − CCS	0.236	0.264	0.249	128

Table 4
Performance comparison on word-to-word relation based on IIR_MB + Prog

Model	Precision	Recall	$F1$	NumWords
IIR_MB + Prog	0.223	0.269	0.244	130
+WNC	0.212	0.251	0.230	125
+SC	0.213	0.275	0.240	133
+CCS	0.229	0.273	0.249	132
+SC − CCS	0.232	0.281	0.254	136

Table 4 presents the similar comparison, in which the improved image-to-image relation and the improved learning method are utilized together (denoted as IIR_MB + Prog) to prepare the candidate annotations for the annotation refinement. Similarly, the encouraging performance of the combined correlation (SC + CCS) is achieved.

5.5. Comparison on learning method

In Table 5, we present the performance comparison among some related work with different learning methods. The discussion about the comparison is carried out on two groups of methods. CRM_Prog (or MBRM_Prog) denotes the annotation process is performed with CRM (or MBRM) in the progressive learning manner:

CMRM and CLM: Both models share blob-based image features, while CLM designs a new language model to represent the image-to-word relation and adopts the EM algorithm to update its image-to-image relation and the language model. Simply, CLM utilize a more sophisticated learning model to improve the image-to-image relation and the image-to-word relation. Accordingly, it boosts its performance. The *NumWord* is increased from 66 words in CMRM to 79 in CLM, and the *F*1-measure is increased from about 0.09 to around 0.11.

CRM and CRM_Prog (MBRM and MBRM_Prog): It can be observed that all the measures are significantly improved by applying the progressive learning method, especially on *NumWord*. Because the progressive learning algorithm considers the word correlation, more correct words are annotated. Similar improvement can also be demonstrated from the comparison between MBRM and MBRM_Prog.

5.6. Overall comparison

In the experiment, we integrate all the proposed improvements, i.e. the reinforced image-to-image relation, the combined word-to-word relation, and the progressive learning method, into the framework to evaluate the overall performance compared with other related work. For clarity, we denote the combined scheme as “OurComb”. The overall comparison is illustrated in Fig. 4. Obviously, OurComb outperforms all the other methods in the comparison.

To specify the effect of each improved part, we give a comparison among their partial combinations in Table 6. Specially, they are MBRM, IIR_MBRM (applying the improved image-to-image relation to MBRM), IIR_MB + Prog (applying the progressive learning method to IIR_MBRM), IIR_MB + WWR (using the combined word correlation to refine the annotations from IIR_MBRM), and OurComb (three improvements are combined together as Section 4).

According to Table 6, some useful conclusions can be derived. First, with the reinforced image correlation, IIR_MBRM achieves better performance than MBRM. Second, IIR_MB+Prog shows the most noticeable improvement on the metric of *NumWord*. This is because that the progressive method considers the word correlation in the learning process instead of the independence among words as MBRM or IIR_MBRM. Third, IIR_MB+WWR shows better than IIR_MBRM, especially on the average precision. The improvement is attributed to the annotation refinement by exploring the proposed word correlation. Several noise annotations are removed and relevant annotations are complemented. Finally, the combination of the three improved parts further enhances the performance compared with any part alone. Thus, these three parts benefit to each other, and they can be integrated to perform image annotation effectively.

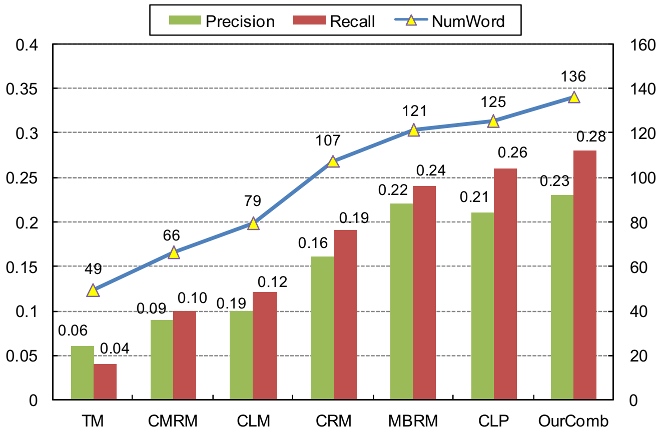


Fig. 4. Overview of image annotation framework.

Table 5
Performance comparison on learning methods

Model	Precision	Recall	*F*1	NumWords
CMRM	0.09	0.10	0.09	66
CLM	0.10	0.12	0.11	79
CRM	0.16	0.19	0.17	107
CRM_Prog	0.20	0.24	0.22	127
MBRM	0.22	0.24	0.23	121
MBRM_Prog	0.22	0.26	0.24	130

Table 6
Performance comparison on partial combinations of our proposed improvements

Model	Precision	Recall	*F*1	NumWords
MBRM	0.218	0.243	0.230	121
IIR_MBRM	0.226	0.259	0.241	122
IIR_MB + Prog	0.223	0.269	0.244	130
IIR_MB + WWR	0.240	0.266	0.252	128
OurComb	0.232	0.281	0.254	136

6. Conclusions and future work

In this paper, we propose a unified framework for the automatic image annotation. It includes two graph learning processes by exploring three kinds of relations. The reinforced image-to-image relation, the combined word-to-word relation, and the progressive learning method are proposed to effectively improve the performance of image annotation. The comprehensive experiments and discussion demonstrate that any improvement in the framework is beneficial to image annotation.

In future work, we will perform more improvements with the assistance of the proposed framework and strive to explore web resources to make the large-scale images annotation possible and effective.

Acknowledgement

The research was supported by National Natural Science Foundation of China (60605004, 60675003) and National 863 Project (2006AA01Z315, 2006AA01Z117).

References

Blei, D., Jordan, M., 2003. Modeling annotated data. In: Proc. 26th Internat. Conf. on Research and Development in Information Retrieval.

Budanitsky, A., Hirst, G., 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. North American Chapter of the Association for Computational Linguistics (NAACL-2001).

Chang, E., Kingshy, G., Sychay, G., Wu, G., 2003. Cbsa: Content-based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans. CSVT 13 (1), 26–38.

C. Cusano, G.C., Schettini, R., 2003. Image annotation using svm. In: Proc. Internet Imaging IV, SPIE 5304, vol. 5304, December, pp. 330–338.

Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.A., 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. 7th Eur. Conf. on Computer Vision, UK, pp. 97–112.

Feng, S.L., Manmatha, R., Lavrenko, V., 2004. Multiple bernoulli relevance models for image and video annotation. In: IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 1002–1009.

Jeon, J., Lavrenko, V., Manmatha, R., 2003. Automatic image annotation and retrieval using cross-media relevance models. In: Proc. 26th Ann. Internat. ACM SIGIR, pp. 119–126.

Jiang, J., Conrath, D., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. Internat. Conf. on Research in Computational Linguistics.

Jin, R., Chai, J., Si, L., 2004. Effective automatic image annotation via a coherent language model and active learning. In: Proc. 12th Ann. ACM Internat. Conf. on Multimedia, pp. 892–899.

Jin, Y., Khan, L., Wang, L., 2005. Image annotations by combining multiple evidence wordnet. In: Proc. 13th Ann. ACM Internat. Conf. on Multimedia, pp. 706–715.

Kang, F., Jin, R., Sukthankar, R., 2006. Correlated label propagation with application to multi-label learning. In: IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 1719–1726.

Lavrenko, V., Manmatha, R., Jeon, J., 2003. A model for learning the semantics of pictures. In: Proc. Advance Neutral Information Processing.

Li, J., Wang, J., 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Machine Intell. 25 (19), 1075–1088.

Liu, J., Li, M.J., Ma, W., Liu, Q., Lu, H.Q., 2006. An adaptive graph model for automatic image annotation. In: 8th ACM International Workshop on Multimedia Information Retrieval, pp. 61–70.

Pan, J., Yang, H., Pinar, D., 2004. Automatic multimedia cross-modal correlation discovery. In: 10th ACM Internat. Conf. on Knowledge Discovery and Data Mining, pp. 653–658.

Stricker, M., Orengo, M., 1995. Similarity of color images. In: Storage and Retrieval of Image and Video Databases III, vol. 2, pp. 381–392.

Tong, H., He, J., Li, M., Ma, W., Zhang, H.J., Zhang, C., 2006. Manifold-ranking based keyword propagation for image retrieval. EURASIP J. Appl. Signal Process. 21, 1–10, Special Issue on Information Mining from Multimedia Database.

Wang, C.H., Jing, F., Zhang, L., Zhang, H.J., 2006. Image annotation refinement using random walk with restarts. In: Proc. 14th Ann. ACM Internat. Conf. on Multimedia, pp. 647–650.

Zhou, D., Bousquet, O., Lal, T., Weston, J., Scholkopf, B., 2003. Ranking on data manifolds. In: Proc. 18th Ann. Conf. on Neural Information Processing System, pp. 169–176.