Dual Cross-Media Relevance Model for Image Annotation*

Jing Liu
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
+86-10-62542971
jliu@nlpr.ia.ac.cn

Bin Wang
MOE-MS Key Lab of MCC
University of Science and Technology of China
Hefei 230026, China
+86-551-3600681
binwang@ustc.edu

Mingjing Li, Zhiwei Li, Wei-Ying Ma
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
+86-10-58968888
{mjli, zli, wyma}@microsoft.com

Hanqing Lu, Songde Ma
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
+86-10-62542971
luhq@nlpr.ia.ac.cn, mostma@gmail.com

ABSTRACT
Image annotation has been an active research topic in recent years due to its potential impact on both image understanding and web image retrieval. Existing relevance-model-based methods perform image annotation by maximizing the joint probability of images and words, which is calculated by the expectation over training images. However, the semantic gap and the dependence on training data restrict their performance and scalability. In this paper, a dual cross-media relevance model (DCMRM) is proposed for automatic image annotation, which estimates the joint probability by the expectation over words in a pre-defined lexicon. DCMRM involves two kinds of critical relations in image annotation. One is the word-to-image relation and the other is the word-to-word relation. Both relations can be estimated by using search techniques on the web data as well as available training data. Experiments conducted on the Corel dataset and a web image dataset demonstrate the effectiveness of the proposed model.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms
Algorithms, Measurement, Experimentation

Keywords
Image annotation, Image retrieval, Relevance model, Word correlation

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany

1. INTRODUCTION
With the advent of digital imagery, the number of digital images has been growing rapidly and there is an increasing requirement on indexing and searching these images effectively. 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 [15], image annotation using SVM [3] and Bayes point machine [2] fall into this category. Some other methods try to learn a relevance model associating images and keywords, which is also our focus in this paper. The early work in [4] applied a machine translation model to translate a set of blob tokens (obtained by clustering image regions) to a set of keywords. [9] introduced the Cross-Media Relevance Model (CMRM), which used the keywords shared by the similar images to annotate new images. The CMRM was subsequently improved by the continuous-space relevance model (CRM) [14] and the multiple Bernoulli relevance model (MBRM) [5]. Recently, there are some efforts considering the word correlation in the annotation process, such as the Coherent Language Model (CLM) [11], the Correlated Label Propagation (CLP) [13], and the WordNet-based method (WNM) [12].

All above previous methods suffer from two problems. One is their dependence on the training dataset to learn the models. However, it is really hard to get a well-annotated set, and their scalability is doomed. The other is the well-known semantic gap.

* This work was performed at Microsoft Research Asia.

With traditional simple associations between images (visual content features) and words, the degradation of annotation performance is unavoidable.

The web prosperity brings a huge deposit of almost all kinds of data and provides solutions to many problems that are believed to be unsolvable [6][20]. In recent years, some researchers began to leverage web-scale data for image annotation. A pioneer work was proposed by Wang et al. [19]. In their work, at least one accurate keyword is required by the text-based image searcher to find a set of semantically similar images. Then the content-based image search in the obtained image set is performed to retrieve visually similar images. At last, annotations are mined from the text descriptions (title, URLs and surrounding texts) of these images, which are similar on both semantics and visual content. However, the initial accurate keyword for each image is unreasonable in practice. Moreover, the method needs to perform the content-based search on a well built image set, which is not publicly accessible. Additionally, there is no ideal commercial content-based search engine currently. Thus the method still has some restrictions for the web image annotation.

To address above problems, we propose a relevance model between image and word, which is named as Dual Cross-Media Relevance Model (DCMRM). Similar to traditional relevance models, DCMRM annotates images by maximizing the joint probability of images and words. The "dual" refers to the exchange on the roles of "images" and "words" between DCMRM and those traditional relevance models. In contrast to the traditional models, which estimate the joint probability by the expectation over images in the training set, DCMRM makes the estimation by the expectation over words in a given lexicon. This not only alleviates the dependence on the training set, but also enables the integration of web search techniques into the framework of image annotation. With the development of various search techniques, some commercial image search engines, like Google and Yahoo!, can provide some good search results. Then such publicly available resources can benefit to the image annotation. Namely, the integration of web search techniques in DCMRM is equivalent to standing on the shoulder of a giant.

The dual model mainly involves two items, i.e., word-to-word relation and word-to-image relation**.** Certainly, the two types of relations can be well estimated given a training set. When the training set is unavailable, where traditional methods fail, DCMRM is still applicable. In this paper, we design a set of search-based schemes to estimate both relations in DCMRM. First, we present how to calculate the likelihood of an image given certain keyword by exploring the idea of keyword-based image retrieval, in which the top-ranked images obtained by a commercial image search engine are considered. Second, we design two search-based word correlations in the web context and combine them in a linear form. One is the statistical correlation based on the counts of resulting images. The other is the content-based correlation, which is estimated by the visual consistence among resulting images. Obtaining the two relations, we integrate them into the proposed model and perform the image auto-annotation. Exciting performance of our solution based on DCMRM is demonstrated from the experiments on the Corel dataset and the web dataset.

The main contributions in this paper are:

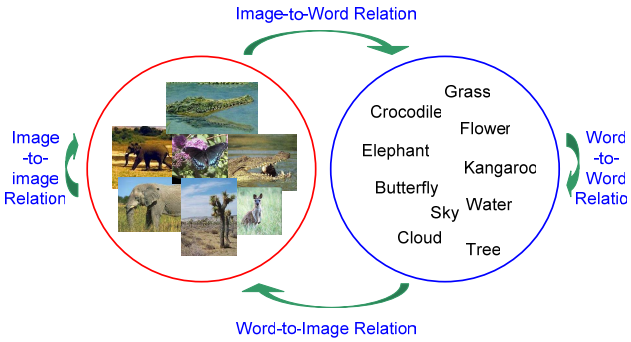

Figure 1. Illustrative example of image annotation

1) The proposed DCMRM provides a new direction to the study of image auto-annotation. To the best of our knowledge, we are the first to formally integrate the word relation, image retrieval, and web search techniques together to solve the image annotation problem.

2) The search-based schemes are designed to obtain two critical relations in DCMRM. This relieves the dependence on training set and makes the large-scale image annotation possible.

3) The proposed solution makes full use of well-indexed web images from commercial search engines to build the correspondence between image and word effectively.

The rest of this paper is organized as follows. In Section 2, a formal study about traditional methods and the proposed DCMRM is presented. The detailed implementations of the search-based image annotation are addressed in Section 3. The experimental results are reported in Section 4. Finally, the conclusion and future work are given in Section 5.

2. RELEVANCE MODELS FOR IMAGE ANNOTATION

In the problem of image annotation, there are two media types: image and word. They form four kinds of relations as illustrated in Fig. 1: image-to-image relation (IIR), word-to-word relation (WWR), image-to-word relation (IWR) and word-to-image relation (WIR). IIR is the relation between images, which is typically built with visual content features. WWR is the relation between words, which is usually built with statistical correlation in the corpus or certain lexicon [10] (such as WordNet [16]). IWR denotes the likelihood of a keyword given an image, which is built on the training set in traditional methods. WIR represents the likelihood of an image given certain keyword, which is similar to the goal of keyword-based image retrieval, i.e., ranking images according to their relevance to the query keyword.

Image annotation can be understood as a learning process, in which the unknown relations between test images and annotated words are estimated by exploring available resources. Thus, how to estimate and integrate these four relations is a key issue. Traditional methods usually focus on IIR and IWR, while WWR and WIR are rarely considered. In the following, we will present a formal study about traditional relevance models and the proposed DCMRM in terms of these relations.

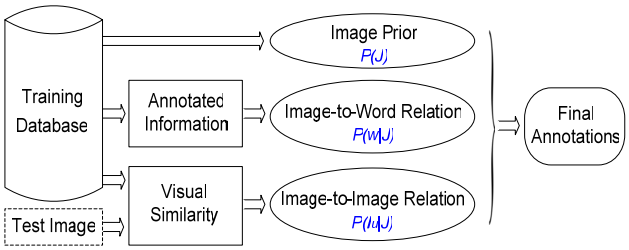

Figure 2. The annotation framework of traditional relevance models

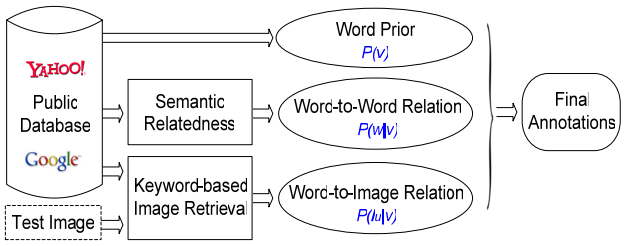

Figure 3. The annotation framework of DCMRM

2.1 Traditional Relevance Models

Image annotation methods based on the relevance model, such as CMRM [9], CRM [14] and MBRM [5], have achieved encouraging performance and become a promising direction in the literature. These relevance models are used to estimate the joint distribution of words and images, which typically require a training database with high quality. The joint distribution can be computed as the expectation over training images, and the annotations for untagged images are obtained by maximizing the expectation as:

$$\begin{aligned} w^* &= \arg\max_{w \subset V} \{P(w \mid I_u)\} \\ &= \arg\max_{w \subset V} \{P(w, I_u)\} \\ &= \arg\max_{w \subset V} \{\sum_{J \in T} P(w, I_u \mid J) P(J)\} \end{aligned} \tag{1}$$

where J is an image in the training set T, w is a word or a set of words in the annotation set V, and I_u is an untagged image.

With the assumption that the probabilities of observing the word w and the image I_u are mutually independent given an image J, the model can be rewritten as:

$$w^* = \arg\max_{w \subset V} \{\sum_{J \in T} P(w \mid J) P(I_u \mid J) P(J)\} \tag{2}$$

where $P(w|J)$ denotes the likelihood of w given the training image J, i.e., IWR. $P(I_u|J)$ denotes the probability of I_u given J, i.e., IIR. $P(J)$ is the probability of selecting the image J.

There are three components in the traditional relevance models.

- $P(J)$ indicates the prior distribution of an image, which is usually given a uniform prior.
- $P(I_u|J)$ (IIR) represents the liklihood of the test image (I_u) given the training image (J), which is estimated by the visual similarity between images.
- $P(w|J)$ (IWR) models the word distribution in the training set, such as the multinomial distribution in CRM [14] and the multiple Bernoulli distribution in MBRM [5].

From the view of the relation exploration, a general explanation can be attached to those models with the form as Eq. 2. The explanation is that the words with the prior confidence (IWR: $P(w|J)$) are propagated from training images to un-annotated images through their visual similarities (IIR: $P(I|J)$), while $P(J)$ can be viewed as the weight of each training image to reflect its importance. The annotation framework for this type of relevance models is illustrated in Fig. 2.

2.2 Dual Cross-Media Relevance Model

The traditional models aim to model the joint probabilities of images and annotations, and they start with the training images to generate the model. It is also reasonable to compute the expectation over words instead of training images, which is given as:

$$\begin{aligned} w^* &= \arg\max_{w \subset V} \{P(w, I_u)\} \\ &= \arg\max_{w \subset V} \sum_{v \in V} P(w, I_u \mid v) P(v) \end{aligned} \tag{3}$$

The reformulated model requires a pre-defined lexicon (V) rather than a clean training image database. Once the lexicon is given, the model alleviates the dependence on training set, and enables much potential to image annotation. The construction of the lexicon can appeal to some developed resources, e.g. the lexicon of WordNet, besides the training dataset. In addition, studies about quantifying the "visualness" of words [21] are also beneficial. Different from the content features of images, words as the most direct representation of semantics are more manageable whether on the scale or on the descriptors.

Similarly, we assume that the probability of observing the word w and the image I_u are mutually independent given a word v, and the relevance model can be rewritten as:

$$w^* = \arg\max_{w \subset V} \sum_{v \in V} P(I_u \mid v) P(w \mid v) P(v) \tag{4}$$

where $P(I_u|v)$ denotes the probability of an untagged image I_u given a word v, i.e. WIR. $P(w|v)$ denotes the probability of a word w given a word v, i.e. WWR. $P(v)$ is the probability of selecting a word v to generate the observations I_u and w.

Comparing Eq. 3 with Eq. 1 (or Eq. 4 with Eq. 2), it is clear that the proposed model has a dual form of the traditional models in that the roles of the images and words are exchanged. Actually, the essence of the proposed model is far more than a simple exchange. The detailed disucssion will be presented in the following subsection.

Three components are included in the proposed model, which are introduced as follows:

- $P(v)$ indicates the importance (or popularity) of a word, which can be estimated by employing the techniques in textual information retrieval and some domain knowledge.
- $P(w|v)$ (WWR) represents the semantic relation between two words. Corpus-based statistics [18] and WordNet-based word correlation [12] are usual solutions.

- $P(I_u|v)$ (WIR) models how the image (I_u) is relevant to the given word (v). This modeling is consistent with the target of the keyword-based image retrieval. Accordingly, many existing retrieval and search techniques can be adopted.

There have been many studies focusing on $P(w,v)$. To benefit from those work, we unite $P(v)$ and $P(w|v)$ to rewrite Eq. 4 as:

$$P(w,v) = P(w|v) \cdot P(v) \tag{5}$$

$$w^* = \arg\max_{w \subset V} \sum_{v \in V} P(I_u|v) P(w,v) \tag{6}$$

With these components, we can derive a concise interpretation of the proposed model. $P(I_u|v)$ (WIR) is the prior confidence on the correspondence between the untagged image and the word. Based on the prior confidence, the annotation expansion and refinement are realized as Eq. 6 by exploring the word correlation. The framework based on DCMRM is illustrated in Fig. 3.

2.3 Comparison and Discussion

In the following, we will discuss the correspondences between the two types of models and demonstrate the advantages of DCMRM.

First, $P(J)$ corresponds to $P(v)$, because both of them belong to the estimation of a prior distribution. Due to the complex image space with large scale, it is really hard to estimate the prior probability of an image. So, usually a uniform distribution is assumed for $P(J)$. Obviously, this is a compromise as no better solution can be found. However, $P(w)$ can be better estimated. This is because that the scale of a common lexicon is quite small and the distribution of words is relatively simple. Furthermore, the word directly reflects the semantics, and can be easily understood by human. Then many textual resources and a great deal of experiential knowledge can be used to estimate the prior distribution of a word.

Second, $P(w|J)$ corresponds to $P(I_u|v)$, because both of them aim to build the correspondence between an image and a word. Specially, $P(w|J)$ is a mapping from the image J to the word w, while $P(I_u|v)$ is a mapping from the word v to the image I_u. In traditional approaches, $P(w|J)$ is typically estimated by the statistical distribution of words on the training set. Thus its performance deeply depends on the quality of the training set, and the scalability is poor. Even if a large and well-annotated dataset is given**,** this direct mapping suffers from the semantic gap. As for $P(I_u|v)$, it can be estimated using the idea of the keyword-based image retrieval. Due to the rapid development on web image search, well indexing of images usually can be obtained and it provides some beneficial information to build the correspondence. Therefore, employing a commercial web searcher to estimate the relevance of images given a query is preferable to alleviate the semantic gap as well as the dependence on the training set.

Third, $P(I_u|J)$ corresponds to $P(w|v)$ in that they are both pairwise relations between two instances from the same type of media. Specially, $P(I_u|J)$ is the image-based similarity, while $P(w|v)$ is the word-based similarity. $P(I_u|J)$ is typically built with the content features of images. Usually, the similarities on different image-pairs attribute to different objects in images and are also reflected on different types of features, e.g. color, texture or shape. Thus the quality of the image-based similarity suffers from the difficulties in image understanding and analysis. The number of common words is basically finite, while the number of images is always infinite. The meaning of a word is relatively limited and fixed, while the expression of image is rich and diverse. Furthermore, many well-developed techniques in the textual information retrieval and the natural language processing can be applied to solve the case of words. Therefore, $P(w|v)$ (WWR) can be estimated more effectively.

Finally, although both frameworks are based on the relevance model, their estimating processes are different in essence. The traditional one is regarded as a process of the label propagation. It aims at propagating semantics from training images to test images according to their visual similarities. In the process, semantic relatedness between images is desired originally, but $P(I_u|J)$ is actually the similarity computed with content features. Obviously, the substitution makes the propagation suffer from the semantic gap. As for DCMRM, it is a process of the annotation expansion and refinement by exploring the word correlation, while the correlation is built on the semantic relevance. Thus the process is more reasonable and effective.

To sum up, the proposed DCMRM has great potential to relieve the dependence on the training database and alleviate the semantic gap to some extent. Similar to other models, DCMRM can equally benefit from a clean training set. In case that the training set is unavailable, DCMRM shows its particular advantage by enabling web search techniques to estimate the required information. As we know, the web represents the largest publicly available corpus with aggregate statistical and indexing information. Thus such huge and valuable web resources deserve our attention. Then, we design a search-based solution for DCMRM, which will be presented in the following section.

3. SEARCH-BASED IMAGE ANNOTATION

As above mentioned, WWR and WIR are two critical items in the proposed DCMRM. In this section, we present the details on the estimation of both relations. At first, we will discuss how to estimate WIR by using a web image searcher. Next, assuming a lexicon has been given, two types of search-based WWRs are designed in the web context.

3.1 Search-based Word-to-Image Relation

The WIR is important in DCMRM because it enables the keyword-based image search to be applicable for image annotation. This makes image annotation benefit from the encouraging performance of the web search engines. Accordingly, the scalability of the model also becomes possible. In the following, we will detail the method to calculate the word-to-image relation, i.e., $P(I_u|v)$, by using a commercial image search engine.

Given a keyword query, a web image search engine, e.g. Google image searcher, usually returns good search results, especially those on the first page. Accordingly, top-ranked images can be roughly treated as the visual representation of the query. Then we use the similarity between the untagged image and the resulting images to represent the relation between the untagged image and the query word.

Generally, a search engine gives its attention to the relevance and the diversity of resulting images simultaneously. For example, if we submit ‘jaguar’, images about an animal, a car or a plane may

appear in the resulting set. That is, returned images usually are diverse on semantic meaning and visual appearance. So, if partial resulting images are visually similar to the untagged image, we can deduce that the untagged image is likely to be annotated by the query word. Based on this consideration, larger weights are set to the images which are more similar to the untagged image. The calculation of image-based similarity is given as follows:

$$S(I_u, R_v) = \sum_{r_i \in R_v} \alpha_i S_{IIR}(I_u, r_i) = \sum_{r_i \in R_v} \alpha_i \exp(-\frac{d(f_u, f_{r_i})}{\sigma_1}) \quad (7)$$

where v is a query word, and $S_{IIR}(I_u,r_i)$ is the similarity between the untagged image (I_u) and the image r_i in the resulting image set R_v (experientially including top-ranked 20 images). α_i is a adjustable parameter, which aims to support similar image-pairs and penalize dissimilar image-pairs. And $d(.)$ is certain distance metric between feature vectors f_u and f_{ri}, which is L_1-distance in our implementation.

Usually, all the top-ranked images can be crawled when the lexicon is given. Then both the image crawling and the feature extraction can be done a prior once and for ever. To annotate an untagged image, only the calculation of image similarity as Eq. 7, needs to be conducted. So the cost of online computation is not expensive.

If an untagged image is extracted from a web page, some textual information, such as title, URL and surrounding text, can also be utilized to measure the relation between the image and certain word in a lexicon. Here, we adopt a simple manner to calculate the semantic relatedness as follows:

$$S(W_u, v) = \sum_{w_i \in W_u} \beta_i S_{WWR}(w_i, v) \quad (8)$$

where W_u is a set of words representing the textual information of I_u, and β_i is the weight positively relevant to the *tf-idf* value [17] of w_i, $S_{WWR}(w_i,v)$ is the correlation between words w_i and v, which will be discussed in Section 3.2.

Then the final word-to-image relation can be approximated by integrating above two measures, which is given as:

$$S_{WIR}(I_u, v) = [S(I_u, R_v)]^{\eta_1} \cdot [S(W_u, v)]^{\eta_2} \quad (9)$$

where $\eta_1, \eta_2 \geq 0$ are parameters to control the reliabilities of both measure. Specially, $\eta_2 = 0$, if there is no textual information for I_u.

3.2 Word Correlation on the Web

Within DCMRM, word correlation is used to refine and expand candidate annotations obtained from the above WIR indeed. A well-defined word correlation enhances the performance of image annotation. In the following, we will detail the calculation of word correlation using a web search engine in two different manners. One is a statistical correlation within the web context, and the other is a content-based correlation obtained from resulting images of Google image searcher.

- **Statistical Correlation by Search**

We consider a hypothesis as in [1]: the relative frequency whereupon two words appear on the web within the same documents gives an idea of their semantic distance. In [1], they define a Normalized Google Distance (NGD) for the text retrieval as follows:

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log G - \min\{\log f(x), \log(f(y)\}} \quad (10)$$

where x and y are two query words, G is the total number of web pages indexed by Google, $f(x)$ is the count of pages where word x appears, and $f(x,y)$ is the count of pages where both x and y appear.

The smaller the value of NGD is, the more relevant the two words are on semantics. Though the NGD fails on the triangular inequality property of distance, it provides a measure of how far two terms are related semantically. Furthermore, experimental results in [1] demonstrate that the NGD stabilizes with the growing Google dataset, i.e., scale invariant.

Inspired by [1], we define a NGD-based word correlation using Google image searcher according to Eq. (10), so as to obtain a general correlation measure for any word-pair. As the word is used to describe image semantics, the image search engine is preferable to the textual search engines. In our implementation, x and y are two words in the lexicon, $f(x)$ is the count of images which are indexed by the word x in Google image searcher, and $f(x,y)$ is the count of images which are indexed by both x and y. Nevertheless, we need a bounded measure in [0, 1], which increases according to the degree of semantic correlation. So a general transform $F(.)$ to obtain the NGD-based correlation is defined as:

$$K_{SCS}(x, y) = F(NGD(x, y)) = \exp[-\gamma \cdot NGD(x, y)] \quad (11)$$

where γ is an adjustable factor.

Since the range of NDG values is from 0 to ∞, $S_{WWR}(x,y)$ ranges from 0 to 1. In addition, the value of G, i.e., the total number of indexed images has not been published officially. So, we should infer it from some available reported results or regulate it in experiments. Here, we refer to a report in Aug. 2005 [8], in which the total number of indexed images in Google image searcher is reported to be 2,187,212,422. With the fast development of various web techniques, we can conservatively double the count. Then we set G to be about 4.5 billions in our experiments.

- **Content-Based Correlation by Search**

To get a more robust measure, we try to seek other entrances to enrich the representation of word correlation. Since image is the focus of image annotation, visual content as a direct representation of image should also contribute to the word correlation. Then, a content-based correlation by web search is proposed.

As mentioned in Section 3.1, top-ranked images in the search result can be roughly treated as the visual representation of the search concept. From another view, the visual consistence among these resulting images reflects the semantic uniqueness of the query keyword to some extent. For example, word 'jaguar' may represent an animal, a car or a plane, i.e. it is not a specific word. When it is used as a query, the search images from Google

include have different semantics. Those images have varied colors, shape or texture, thus are not visually consistent.

When two words with the conjunctive operator (AND) [1] are submitted as one query, the returned images are those indexed by both words. If the words are not semantically related to each other, they are unlikely to be associated with the same image together. Accordingly, the search results may be very noisy and not visually consistent. Based on this consideration, the visual consistence of the search results might be a good indicator of the correlation between words.

We use the variances of visual features to describe the visual consistence. The less variance of visual features corresponds to be more consistent on the visual appearance. In addition, various images usually require different types of visual properties to characterize their underlying visual appearances. To get an adaptive measure, we propose a new one called Dynamic Partial Variance (DPV). From studies of cognitive psychology [7], human infers overall similarity based on the aspects that are similar among the compared objects, rather than based on the dissimilar ones. Similarly, the DPV focuses on the features with low variances and activates different features for different image sets. Assuming the variances of each dimensional feature among images in set S are ordered as $\mathrm{var}_1(S) \le \mathrm{var}_2(S) \le \cdots \le \mathrm{var}_d(S)$, the *DPV* is defined as:

$$DPV(S) = \frac{1}{l}\sum_{i=1}^{l<d} \mathrm{var}_i(S) \tag{12}$$

where d is the dimension of the visual feature, and l is the number of similar aspects activated in the measure.

To make the DPVs of the resulting images given word-pair queries comparable to each other, we normalize the values according to the semantic uniqueness of each single word, i.e., the DPV of the resulting images given a single-word query. Given two words x and y, the correlation based on the visual consistence is calculated as follows:

Step 1: We submit 'x', 'y', 'x y' to Google image searcher, and obtain three sets of images, denoted as S_x, S_y, S_{xy}, which include top M (usually $M=20$) resulting images respectively.

Step 2: For each image, multi-modal visual features are extracted to characterize various visual properties of images.

Step 3: For each set, the values of DPV are calculated according to Eq. 12.

Step 4: The correlation between x and y is given as:

$$K_{CCS}(x,y) = \exp(-\sigma_2 \cdot \frac{DPV(S_{xy})}{\min\{DPV(S_x), DPV(S_y)\}}) \tag{13}$$

where $\sigma_2 > 0$ is a smoothing parameter.

- **Combined Word Correlation**

Now we get two types of word correlations with different characteristics. To make both relations complement each other, our proposal is to unify them in a linear form after they are normalized into [0, 1]. The linear combination is given as:

$$S_{WWR} = \varepsilon K_{SCS} + (1-\varepsilon) K_{CCS} \tag{14}$$

where $0<\varepsilon<1$, and experientially $\varepsilon=0.5$ in our implementation. Better performance is expected when more sophisticated combinations are used, which is our future work.

Obtaining both WIR and WWR, we can rank keywords in the lexicon for an un-annotated image by Eq. 6, and those top ranked keywords will be selected as final annotations for the image.

4. EXPERIMENTS

We will evaluate the proposed DCMRM on two different situations. The first is done on a high-quality training set i.e., the standard Corel dataset. The goal is to demonstrate that the proposed framework can get better results when the good training information is available. The other is based on a web dataset without any manual label information. It aims to prove that the framework can be applicable without any training knowledge and also achieve the promising performance.

4.1 Experiment Design

Corel Dataset: It is publicly available and widely used in current image annotation work. The dataset contains 5,000 images. Each image is segmented into 1-10 regions. A 36-dimensional feature is extracted from each region, which includes color, texture and area features as in [4]. Each image is annotated with 1-5 words. The total number of words is 371. The dataset is divided into two parts: 4,500 images for training and rest 500 for test.

In order to present a fair comparison on the dataset with other related work, we perform our implementations by searching on the training dataset instead of the web. In addition, all comparative experiments are carried on the dataset using visual features of segmented regions rather than ones of rectangle regions as in [5].

Web Dataset: It is built to test the applicability of the proposed algorithm. We select 120 queries from 300 top queries in our collected search log, and submit them to Google image searcher. For each query, 100 top-ranked images are crawled and their corresponding web pages are also downloaded. With an HTML parser which depends on DOM-tree structure, the textual information of each image is extracted. It includes the words in title, URL, ALT tag, anchor text and surrounding text. Those words occurring less than 30 times are filtered out. At last, we obtain a dataset of 12,000 images as the test images and 2,020 words as a pre-defined lexicon. Both images and words show great diversity. Additionally, a 64-dimensional visual feature vector for each image is extracted, including 44-dimensional color correlogram, 14-dimensional texture moment, and 6-dimensional color moment.

4.2 Evaluation on Corel Dataset

In this subsection, we use the Corel dataset to evaluate the performance of the dual cross-media relevance model (DCMRM) for image annotation. Similar to previous works, the quality of automatic image annotation is measured 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

[1] By default, Google returns pages that include all of your search words. There is no need to include "AND" between words.

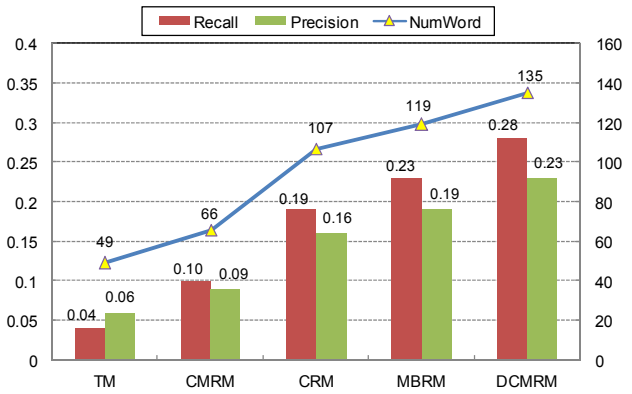

Figure 4. Image annotation performance comparison on all 260 words of Corel dataset: The bars present the average Precision & Recall according to the left axis and the line denotes the NumWord according to the right axis.

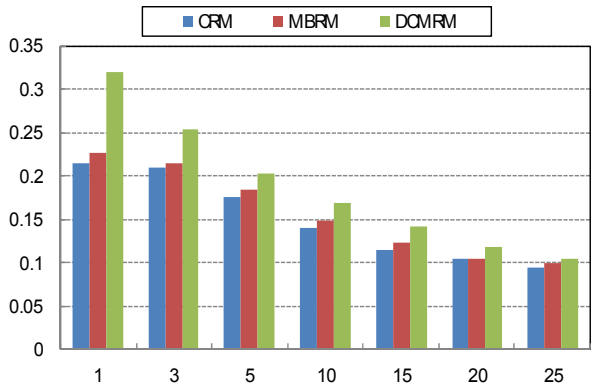

Figure 5. Average Precision at *m* retrieved images comparison on all 260 words of Corel Dataset

number of retrieved images is denoted as N_s, and the number of truly related images in test set is denoted as N_r. Then the precision and recall are computed as follows:

$$precision(w) = \frac{N_c}{N_s}, \qquad recall(w) = \frac{N_c}{N_r} \tag{15}$$

We compute the average precision and recall over all the words in test images (260 words) to evaluate the performance. In addition, we give another 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. This metric is important because a biased model can also achieve high precision and recall values by only performing extremely well on a small number of common words.

Fig. 4 shows the results derived on the test set, where the recall and precision are averaged over 260 words. We also present the results of other related work under the same experiment setting. Specifically, we consider: TM [4], CMRM [9], CRM [14], and MBRM [5]. For all the methods, the annotation length for each image is set to be 5.

From Fig. 4, it is clear that the four relevance-model-based approaches are better than the translation model approach (TM) (with 2-5 times). So we present following comparisons focusing on the relevance model based methods. The proposed DCMRM achieves the best performance in the comparison. Compared with MBRM, it gains 22%, 20% and 13% on Recall, Precision and NumWord respectively. And the corresponding gains are 47%, 44% and 26% respectively compared with CRM. Obviously, the more gains can be obtained when compared with CMRM. We can see that DCMRM achieves encouraging improvements on all three measures, in which the average recall is greatly improved. This demonstrates that the proposed framework combined with WWR and WIR presents a more effective description of the joint distribution between images and words. Moreover, it gives fair consideration to each word and relatively enhances the probability of correctly annotating with the rare word, which brings more improvement on the average recall.

In the following, we analyze the annotation performance from the view of image retrieval. Indeed, when one wants to find one type

Table 1. Performance comparison on four query sets

Query length	1 word	2 words	3 words	4 words
Number of Queries	179	386	176	24
P@5				
CRM	0.248	0.190	0.189	0.233
MBRM	0.255	0.203	0.219	0.251
DCMRM	0.285	0.207	0.223	0.281

of images, she or he would like to find the most related ones, i.e., the ranking order of images annotated with certain keyword is important in practice. Then, we should consider another measure to evaluate the performance of these annotation models. Concretely, given a query word, the system returns all the images that are annotated with the query word, and ranks them according to the probabilities that the word can annotate these images. The precision for the top *m* retrieved images by each word, denoted as "P@*m*", is calculated and finally the average precision over all the test words is used to evaluate the retrieval performance.

Fig. 5 presents the performance comparison among CRM, MBRM and DCMRM using the measure of "P@*m*", in which *m* is set to be different values. The comparison demonstrates that DCMRM outperforms other two related work. Compared with CRM, the gains of DSCMRM are 49%, 21%, 16%, 20%, 23%, 13%, and 11% on the average P@1, P@3, P@5, P@10, P@15, P@20, and P@25 respectively. Similar results can be derived when compared with MBRM. This indicates that the ranking order obtained from our proposed method is much better. Furthermore, we observe that the improvements in the precision at less retrieved images are more impressive. Thus we can see that the proposed method is preferable to a casual user in that she or he would like to search for a few relevant items without looking at too much junk.

Additionally, to simulate the habit of common users, we also test the ranking effectiveness of the method using the queries with multiple words as in [14]. Here, we use four sets of queries. The query sets are constructed with 1-, 2-, 3- and 4-word combinations which should occur at least twice in the testing set. An image is considered relevant to a given query if its manual

annotation contains all of the query words. Table 1 shows the details of the average precisions at top 5 retrieved images using four query sets. Obviously, the proposed DCMRM achieves the best performance in the comparison with CRM and MBRM. This is particularly encouraging because the results are obtained over a large number of queries.

4.3 Evaluation on Web Dataset

Generally, web images have extensive semantics and large variation on visual content. The auto-annotation for web image becomes a great challenge work. In the experiment, we use the web dataset mentioned in Section 4.1 to evaluate the practicality of the propose method. Because those web images are very diverse, the annotation length is set to 10. Because the acquisition of the ground truth is too expensive, we evaluate the performance from the view of image retrieval, i.e. *P@m* used in Section 4.2. In the experimental evaluation, we randomly select 100 query words in the lexicon which are listed in Table 2, and manually label the relevance of resulting images. Then we report the average *P@m* over the 100 words to evaluate the performance.

Because the absence of ground truth for the web dataset makes other annotation models unfeasible, we have no way to present a comparison with those models mentioned above. In Fig. 6, we present the performance of DCMRM on different numbers of retrieved images. Although much larger diversity exists in those web images, DCMRM achieves considerable performance. Specially, the precision at first 5 returned images is about 0.24. That is, at least one image is correct among top five images. Even if top 100 images are concerned, the precision is larger than 0.15. Considering that all the information are obtained from web and no human intervene is provided, the performance is really remarkable.

Table 2. 100 query words in the experiment on Web dataset.

airport	cookies	heart	planet	spider
apple	cross	horse	puppy	squirrel
ball	dodge	jaguar	rainbow	star
balloon	dog	kitchen	ring	sun
beach	dolphin	kite	river	sunset
bear	donkey	laptop	rose	sword
bike	door	leaf	Saturn	tattoo
bird	dragon	lion	sauna	tiger
boat	eagle	man	shark	tower
Buddha	earth	map	ship	tree
butterfly	Eiffel	mars	shirt	truck
camera	feet	mobile	shoe	tulip
cape	fish	monkey	shower	turtle
car	flag	motorcycle	skateboard	volcano
cat	ford	mountain	sky	water
chopper	forest	mouth	skull	wedding
cloud	frog	ocean	sky	whale
coast	fruit	panda	snake	wind
colosseum	garden	pizza	snow	wing
computer	gun	planet	sock	wolf

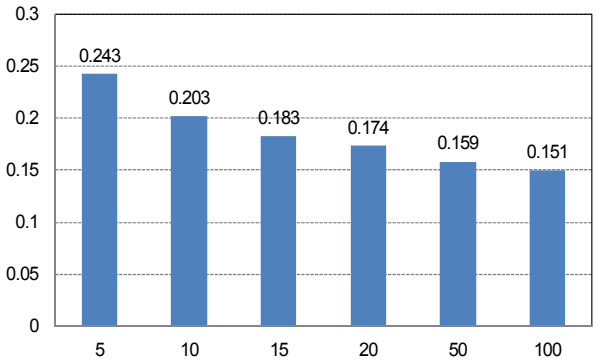

Figure 6. Average P@*m* (*m* = 5, 10, 15, 20, 50, and 100) on randomly selected 100 words of Web dataset

Some examples are presented in Fig. 7. Each column corresponds to one query and its top 5 retrieved images. It can be seen that some retrieved images with the same semantics are possible to be visually inconsistent. We attribute this good performance to the word-oriented characteristic of DCMRM.

In addition, it is worth mentioning that the performance of the proposed method can be further improved by a well-defined lexicon. In our implementation, the lexicon is simply built with all parsed textual information of images without any manual direction. As listed in Table 2, some unsuitable keywords, e.g. "flag", "jaguar", with diverse visual representations or rich semantics, are also collected in the lexicon. Thus such a crude lexicon may influence the performance of the proposed method.

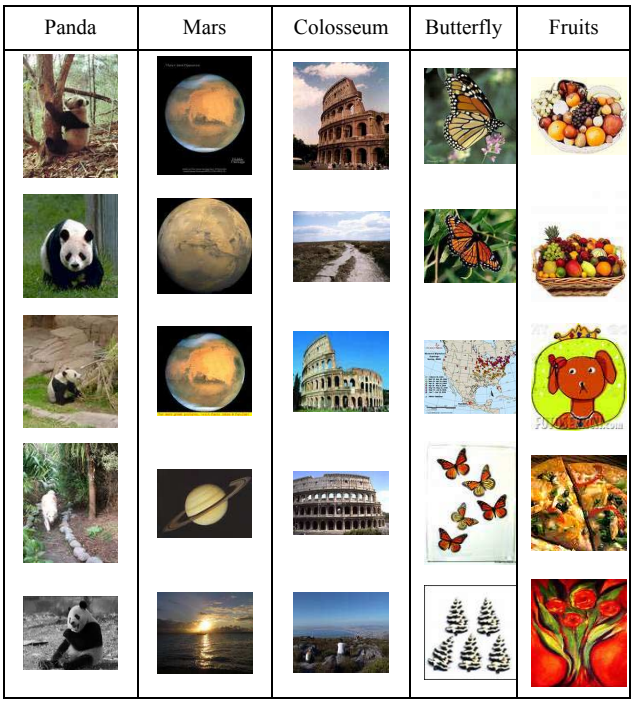

Figure 7. Top 5 retrieved images by some query examples

Table 3. Performance comparison on WWR: The candidate annotations are provided by MBRM.

Model	Precision	Recall	F1	NumWord
MBRM	0.192	0.231	0.209	119
+WNC	0.170	0.219	0.191	115
+SC	0.186	0.243	0.211	125
+SCS	0.195	0.239	0.215	124
+CCS	0.208	0.238	0.221	122
+ SCS-CCS	0.216	0.245	0.230	124
+ SC-SCS-CCS	0.215	0.251	0.232	125

4.4 Comparison among Word Correlations

To ensure good semantic relevance over annotations for each image, the annotation refinement based on a good word correlation is quite necessary. In this section, we make a comparison among different word correlations on the Corel dataset, in which WordNet-based correlation (WNC) [10], statistical correlation in training set (SC) [18], and both proposed word correlations: statistical correlation by search (SCS) and visual consistence based correlation by search (CCS), are considered.

To make an objective analysis, we apply the MBRM to decide the candidate annotations and present the performance comparison among MBRM, MBRM+WNC, MBRM+SC, MBRM+SCS, MBRM+CCS, MBRM+(SCS-CCS), and MBRM+(SC-SCS-CCS) as listed in Table 3. Compared with MBRM, the gains on F1-measure are 1.4% (MBRM+SC), 3.4% (MBRM+SCS), 6.3% (MBRM+CCS), 10.6% (MBRM+(SCS-CCS)), and 11.5% (MBRM+(SC-SCS-CCS)) respectively, but MBRM+WNC gets a poorer performance.

To analyze the comparison in detail, we can find some useful observations. First, the statistical correlation by co-occurrence, i.e., SC, gains obvious improvement on the measure of NumWord, while it losses a little on the average precision. This demonstrates that the correlation is capable of connecting more words through the statistical information, but the connections cannot ensure the relatedness on the semantic level. Second, SCS and CCS achieve more improvements synthetically and the later seems to be better. This is because that both search-based correlations are in the web context and accordingly provide the word correlations from a more general and reasonable level. Additionally, CCS is estimated by an adaptive measurement, i.e. DPV, so as to be more robust to web noise. Third, the WNC shows the worst performance. Specially, the WordNet-based correlation takes a negative role through the annotation refinement. There are 49 out of 371 words in the Corel dataset that either does not exist in WordNet lexicon or have no available relations with other words in WordNet structure. Thus the sparse relation largely weakens the effect of this measure. Finally, the combination of (SCS-CCS) shows comparable performance with the combination of (SC-SCS-CCS). Both of them give a relatively precise and comprehensive representation of word semantic relatedness, which shares the advantages from each single correlation. To make the correlation independent on certain dataset, especially the case for the web images, we select the combination of (SCS-CCS) in our implementation.

5. CONCLUTIONS AND FUTURE WORK

In this paper, a novel annotation framework, the dual cross-media relevance model (DCMRM) is proposed. In contrast to traditional relevance models, which compute the joint probability by the expectation over training images, DCMRM calculates the expectation over words in a pre-defined lexicon. This duality of images and words enables much potential in image annotation. The commercial search engines and their well-developed indexing are explored to estimate the critical components in DCMRM. The search-based WIR and the combined WWR within the web context are unified to perform image annotation effectively. Thus, the dependence on the training dataset, which is necessary to most traditional methods, is relieved by using web search techniques in DCMRM.

The experiments are conducted on both the Corel dataset and the web image dataset. The superior performance on the Corel dataset demonstrates that the proposed DCMRM is more preferable to annotate images when the same training data are provided. In addition, the experiments on the tough web dataset show that DCMRM is potentially applicable for the web image annotation, while other models are unfeasible because the required labeling information with high-quality is not ready for the web dataset.

In our future work, the construction of a good lexicon for image annotation is expected to enhance the performance of the proposed model. Some advanced techniques in the fields of natural language processing, image analysis, and data mining can be applied to collect more keywords that are appropriate for image annotation. Furthermore, we will try more other web search techniques to refine the proposed framework and make the large-scale image annotation more possible and effective.

6. ACKNOWLEDGEMENTS

The research was supported by National 863 Project (2006AA01Z315), National Natural Science Foundation of China (60121302), and Beijing Natural Science Foundation (4072025).

7. REFERENCES

[1] Cilibrasi, R., Vitanyi, P.M.B. *Automatic Extraction of Meaning from the Web*, Proc. IEEE International Symp. Information Theory, USA, 2006.

[2] Chang, E., et al.. *CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines*. CirSysVideo, 2003. 13(1): p. 26-38.

[3] Cusano, C., G. Ciocca, and R. Schettini. *Image Annotation Using SVM.* Proceedings of Internet Imaging, Vol. SPIE 5304. 2004.

[4] Duygulu, P., Barnard, K., Freitas, J., and Forsyth, D. *Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary*. Proceedings of the 7th European Conference on Computer Vision, London, UK, pp. 97 – 112, 2002.

[5] Feng S., Manmatha, R., and Laverenko V. *Multiple Bernoulli Relevance Models for Image and Video Annotation*.

IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.

[6] Fan, X., Xie, X., Li, Z., Li, M., and Ma, W.-Y. *Photo-to-Search: Using Multimodal Queries to Search the Web from Mobile Devices*. ACM SIGMM Workshop on MIR, 2005

[7] Goldstone R. *Similarity, Interactive Activation and Mapping.* Journal Experimental Psychology: Learning, Memory, and Cognition, 20, 1994, 3-28

[8] http://blog.searchenginewatch.com/blog/050809-200323

[9] Jeon, J., Lavrenko, V., Manmatha, R. *Automatic Image Annotation and Retrieval Using Cross-Media Relevance Model*. Proceedings of the 26th annual international ACM SIGIR, 2003.

[10] Jiang, J., Conrath, D. *Semantic Similarity based on Corpus Statistics and Lexical Taxonomy*. Proceedings on International Conference on Research in Computational Linguistics, 1997.

[11] Jin, R., Chai, J., Si, L., *Effective Automatic Image Annotation via a Coherent Language Model and Active Learning.* Proceedings of 12th Annual ACM International Conference on Multimedia, pp. 892-899, USA, 2004.

[12] Jin, Y., Khan, L., Wang, L. and Awad, M. *Image Annotations by Combining Multiple Evidence & WordNet*. Proceedings of the 13th Annual ACM International Conference on Multimedia, Singapore, 2005, pp. 706-715.

[13] Kang, F., Jin, R., and Sukthankar, R. *Correlated Label Propagation with Application to Multi-label Learning*. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, USA, 2006, pp. 1719-1726.

[14] Lavrenko, V., Manmatha, R., and Jeon, J. *A Model for Learning the Semantics of Pictures*. Proceedings of Advance in Neutral Information Processing, 2003.

[15] Li, J. and J.Z. Wang, *Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003. 25(19): p. 1075-1088.

[16] Miller, G. *WordNet: A Lexical Database for English.* Communications of the ACM, 1995.

[17] Salton, G. and Buckley, C.. *Term - Weighting Approaches in Automatic Text Retrieval*. Information Processing and Management, vol. 24, no. 5, pp. 513-523, 1988.

[18] Wang, C. H., Jing, F., Zhang, L., and Zhang, H. J., *Image Annotation Refinement Using Random Walk with Restarts*. Proceedings of the 14th Annual ACM International Conference on Multimedia, 2006, pp. 647-650.

[19] Wang, X., Zhang, L., Jing, F., Ma, W.Y. *AnnoSearch: Image Auto-Annotation by Search*. International Conference on Computer Vision and Pattern Recognition, New York, USA, June, 2006.

[20] Yeh, T., Tollmar, K., Darrell, T. *Searching the Web with Mobile Images for Location Recognition*. International Conference on Computer Vision and Pattern Recognition, 2004, pp. 76-81.

[21] Yanai, K. and Barnard K., *Image region entropy: A measure of 'visualness' of web images associated with one concept*. Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005.