Improve Web Image Retrieval by Refining Image Annotations

Peng Huang, Jiajun Bu*, Chun Chen, Kangmiao Liu, and Guang Qiu

College of Computer Science, Zhejiang University, Hangzhou, China
{hangp,bjj,chenc,lkm,qiuguang}@zju.edu.cn

Abstract. Automatic image annotation techniques are proposed for overcoming the so-called semantic-gap between image low-level feature and high-level concept in content-based image retrieval systems. Due to the limitations of techniques, current state-of-the-art automatic image annotation models still produce some irrelevant concepts to image semantics, which are an obstacle to getting high-quality image retrieval. In this paper we focus on improving image annotation to facilitate web image retrieval. The novelty of our work is to use both WordNet and textual information in web documents to refine original coarse annotations produced by the classic Continuous Relevance Model (CRM). Each keyword in annotations is associated with a certain weight, and larger the weight is, more related to image semantics the corresponding concept is. The experimental results show that the refined annotations improve image retrieval to some extent, compared to the original coarse annotations.

1 Introduction

Currently web image retrieval systems usually fall into two main categories: text-based and content-based. In text-based systems, images are first annotated with texts which are produced by people manually, or extracted from image surroundings automatically, then text retrieval techniques are used to performed image retrieval. However, annotating image manually is tedious, time-consuming and subjective, while annotating images automatically with surroundings often involves terms irrelevant to image semantics unavoidably. Content-based image retrieval (CBIR) systems automatically index and images by their low-level visual features. However, it is flawed in the following ways. Firstly, the user query must be provided in the form of a draft of the desired image. Secondly, the images with similar low-level features may have different contents [8]. Finally, there is so-called semantic gap between image low-level features and high-level concepts.

Recently, many approaches have been proposed to automatically annotate images with keywords [7,2,6]. Automatic image annotation is a promising methodology for image retrieval. However it is still in its infancy and is not sophisticated enough to extract perfect semantic concepts according to image low-level features, often producing noisy keywords irrelevant to image semantics. Noisy concepts may be an obstacle to getting high-quality image retrieval. In this paper we propose a novel approach to improve image retrieval, which utilizes *coherence* between coarse concepts and *relatedness* between concepts and web textual information to refine image annotations. The

[*] Corresponding author.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 430–435, 2008.

refinement is based on two basic assumptions. One assumption is that concepts contained in an image should be semantically related to each other. Another assumption is intuitive that the observation of some specific terms in web documents should increase the belief of certain semantically similar concept. For example, if ‘tiger’ is included in web documents, the annotated concept ‘tiger’ should be more credible than before the observation. In this paper image is annotated as follows: first we use the classic annotation model CRM [6] to associate an image with a set of keywords (coarse concepts); then these coarse concepts are associated with weights which are calculated from *coherence* and *relatedness* using ontological lexicon WordNet [3]. The model proposed by Jin et al. [5] also uses WordNet to improve image annotations, but there are two important differences comparing to our work. First, Jin et al. only take into account *coherence* to remove noisy concepts, while we use extra text in web documents. Second, Jin et al. focused on eliminating ‘noisy’ concepts, while we focus on applying these weighted concepts to improving the ranking of image retrieval results.

The remainder of this paper is organized as follows. Section 2 briefly introduces the classic image annotation model, CRM. Section 3 describes how to refine coarse concepts produced by CRM in details, together with a brief introduction to an image retrieval prototype. Section 4 presents experimental results and some discussions. The last section concludes this paper plus some ideas for future work.

2 Continuous Relevance Model

Let V be the annotation vocabulary, T be the training set, J be an element of T. J is partitioned into a set of fixed-size small regions $\mathbf{r}_J = \{r_1, \ldots, r_n\}$, along with corresponding annotation $\mathbf{w}_J = \{w_1, \ldots, w_m\}$ where $w_i \in V$. The Continuous Relevance Model [6] (CRM) assumes that generating J is based on three distinct probability distributions. First, the set of annotation words $\mathbf{w}_J$ is a result of $|V|$ independent samples from underlying multinomial distribution $P_V(\cdot|J)$. Second, each image region r is a sample of a real-valued feature vector g using a kernel-based probability density function $P_G(\cdot|J)$. Finally, the rectangular region r is produced according to some unknown distribution conditioned on g, so $\mathbf{r}_J$ are produced from a corresponding set of vectors $g_1 \ldots g_n$ according to a process $P_R(r_i|g_i)$ which is independent of J. Now let $\mathbf{r}_A = \{r_1, \ldots, r_n\}$ be the feature vectors of certain image A, which is not in the training set T. Similarly, let $\mathbf{w}_B$ be some arbitrary subset of V ($|\mathbf{w}_B| = m$). Then we like to model $P(\mathbf{r}_A, \mathbf{w}_B)$, the joint probability of observing an image defined by $\mathbf{r}_A$ together with annotation words $\mathbf{w}_B$. The observation of $\{\mathbf{r}_A, \mathbf{w}_B\}$ can be supposed to come from the same process that generated one of the image J^* in the training set T. Formally, the probability of a joint observation $\{\mathbf{r}_A, \mathbf{w}_B\}$ is:

$$P(\mathbf{r}_A, \mathbf{w}_B) = \sum_{J \in T} P_T(J) \cdot \prod_{b=1}^{m} P_V(w_b|J) \times \prod_{a=1}^{n} \int_{\mathbb{R}^k} P_R(r_a|g_a) P_G(g_a|J) dg_a \quad (1)$$

So given a new image we can split it into regions $\mathbf{r}_A$, compute feature vectors $g_1 \ldots g_n$ for each region and then use formula 1 to determine what subset of vocabulary w^* is the most likely to co-occur with the set of feature vectors:

$$\mathbf{w}^* = arg \max_{\mathbf{w} \in V} \frac{P(\mathbf{r}_A, \mathbf{w})}{P(\mathbf{r}_A)} \tag{2}$$

Here we only give a brief introduction to CRM, and for details please refer to [6].

3 Refining Image Annotations

In previous section, we have described how to use CRM to assign an a coarse concept sequence $(c_1, \ldots, c_T)$ to image. However, some concepts are possible noisy or incorrect with respect to image semantics. In what follows we will describe how to distinguish these 'noisy' concepts from others by using the notions of *coherence* and *relatedness* based on WordNet. The notion of *coherence* assumes that concepts in annotations should be semantically similar each other, while *relatedness* refers to the semantic similarity between image annotations and terms in web documents.

3.1 Measuring *Coherence* and *Relatedness*

The JCN algorithm [4] is adopted here to measure the similarity between words (concepts) due to its effectiveness, in which the similarity measure of two concepts 'c_1' and 'c_2' is based on the notations of Information Content (*IC*) and concept-distance, defined as:

$$sim_{jcn}(c_1, c_2) = \frac{1}{dist_{jcn}(c_1, c_2)} = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2))} \tag{3}$$

where $IC(c) = -logP(c)$ and $P(c)$ is the probability of encountering an instance of concept 'c' in WordNet; $lcs(c_1, c_2)$ is the lowest common sub-summer that subsumes both concepts 'c_1' and 'c_2'. Note that all measures are normalized so that they fall within a 0-1 range. For simplicity, normalization factor is omitted, simply assuming that $0 \le sim_{jcn}(c_i, c_j) \le 1$.

Let $C = (c_1, \ldots, c_T)$ be the coarse concepts produced by CRM, $D = (d_1, \ldots, d_n)$ be the terms in page title and image surroundings etc., then the measure of *coherence* of concept c_i is defined as:

$$a_i = \frac{1}{\eta_1} \sum_{j=1 \wedge j \neq i}^{T} sim_{jcn}(c_j, c_i), \quad \eta_1 = \sum_{i=1}^{T} \sum_{j=1 \wedge j \neq i}^{T} sim_{jcn}(c_j, c_i) \tag{4}$$

where η_1 is normalization factor. Similarly, the measure of increased belief for a concept c_i according to the relatedness between concept c_i and textual information D is defined as:

$$b_i = \frac{1}{\eta_2} \sum_{j=1}^{n} sim_{jcn}(d_j, c_i), \quad \eta_2 = \sum_{i=1}^{T} \sum_{j=1}^{n} sim_{jcn}(d_j, c_i) \tag{5}$$

Now we get two variables, a_i and b_i, as the measure of the importance of concept c_i to the semantics of an image. Note that $\sum a_i$and $\sum b_i$ are both 1, that is to say, a_i

and b_i can be regarded as two independent probability distributions of the quantified importance of concept c_i. We combine these two factors linearly as follows:

$$s(c_i) = (a_i + b_i)/2 \tag{6}$$

where $s(c_i)$ is the final score associated with concept c_i. The larger $s(c_i)$ is, the more important it is to the semantics of corresponding image.

3.2 Retrieval Prototype

In the rest of this paper, we refer to $C = \{c_1, \ldots, c_T\}$ as the annotation-set of image *I*, and $Q = \{q_1, \ldots, q_m\}$ as the query. Let *n* be the number of index terms in the system, k_i be a generic index term. $K = \{k_1, \ldots, k_n\}$ is the set of all index terms. A weight $w_{i,c} \geq 0$ is associated with each index term k_i of a annotation-set *C*. For an index term which does not appear in the annotation-set *C*, $w_{i,c} = 0$. With the annotation-set *C* associated, an index term vector $\overrightarrow{c}$ is represented by $\overrightarrow{c} = (w_{1,c}, w_{2,c}, \ldots, w_{n,c})$. Similarly, let $w_{i,q}$ be the weight associated with the pair (k_i, Q) where $w_{i,q} \geq 0$, and $\overrightarrow{q} = (w_{i,q}, \ldots, w_{n,q})$, then we rank the retrieved images as follows:

$$rank(I,Q)=rank(\overrightarrow{c},\overrightarrow{q})=\frac{\overrightarrow{c}\cdot\overrightarrow{q}}{|\overrightarrow{c}|\times|\overrightarrow{q}|}=\frac{\sum_{i=1}^{n}(w_{i,c}\times w_{i,q})}{\sqrt{\sum_{i=1}^{n}(w_{i,c})^2}\times\sqrt{\sum_{i=1}^{n}(w_{i,c})^2}} \tag{7}$$

where the specifications of $w_{i,c}$ and $w_{i,q}$ are as follows::

$$w_{i,c} = \begin{cases} s(k_i), \; k_i \in C \wedge s(k_i) \geq \beta \\ 0, \quad \text{otherwise.} \end{cases}, \; w_{i,q} = \begin{cases} 1, \; k_i \in Q \\ 0, \; \text{otherwise.} \end{cases} \tag{8}$$

where *C* is the annotation-set for image *I* produced by CRM, $s(k_i)$ is the scoring function in formula 6, and β is a threshold for filtering noisy concepts whose scores are below it. In our experiment β was set be 0.1 empirically. For the sake of comparison, we implemented another probability based ranking strategy like in [6]: given a text query Q, we get a conditional probability $P(Q|J)$ for image J according to formula 1, then retrieved images are ranked according to $P(Q|J)$. In short, we use two ranking strategies in this paper, one is the proposed vector-based ranking (VIR) and another is probability-based ranking (PIR).

4 Experiment and Results

The training data set is the Corel Image Dataset, consisting of 5000 images from 50 Stock Photo CDs. Each image is partitioned into 4×6 regions. These images are annotated with words drawn from a vocabulary of size 374, denoted by *V*. In addition, we have previously downloaded 10,000 web pages from the WWW accompanied with images. These images are used as test set in the experiments. We selected top 25 frequent terms in training set as test queries. The CRM was used to annotate each web image with up to 5 keywords. These keywords were used as image indexes for image retrieval.

We used *precision* and *recall* metrics to evaluate the image retrieval results. Given a query *Q* and a set *R* of relevant images for a query *Q*, we obtained a set *A* of relevant

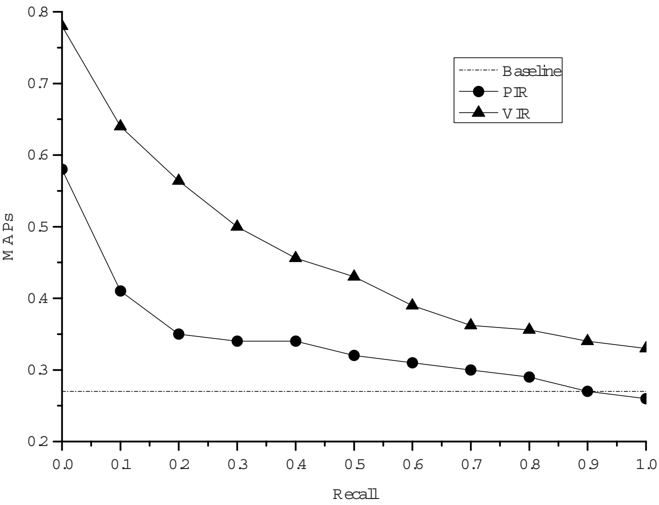

Fig. 1. MAPs of 25 queries using PIR and VIR ranking strategies plus Baseline

images for *Q* through one of above two ranking strategies, then precision is $|A \cap R|/|A|$, and recall is $|A \cap R|/|R|$. To determine the set *R* of relevant images to each of the 25 test queries, we adopted a strategy in [1]: For each test query, we ran out two ranking strategies above. The 40 highest ranked images returned by each of the two ranking strategies were pooled into a set of unique images and then classified by volunteers as relevant or irrelevant with respect to the query term. At the same time, the byproduct of the construction of R is a small set of images which have been labeled be relevant or irrelevant to certain query term by human, denoted by D_R. So we evaluated CRM over D_R as the baseline. Note that all images in $|D_R|$ was associated with the query word, so the precision of CRM over D_R is the number of images correctly annotated with a given word, divided by $|D_R|$. Additionally, it is evident that the recall of CRM over D_R is 100%. We calculated the mean precision for 25 queries and obtained 27%. It was used as baseline and depicted in figure 1.

Usually we want to evaluate average precision at given recall levels. The standard 11-point average precision curve is used for this purpose. It plots precisions at 0 percent, 10 percent, . . . , 100 percent of recalls. The mean average precision (*MAP*) is the arithmetic mean of average precision calculated over all queries (here 25) at some specific percent recall. Note that the results of CRM was a straight line, rather than a curve, because they were annotation accuracies rather than retrieval accuracies. The results were depicted in figure 1. As indicated earlier, the baseline is the mean precision of CRM over D_w for 25 query terms w. The curve for baseline in figure 1 reveals the weakness of automatic image annotation technique in image retrieval task without any ranking strategy, only 27 percent precision. In contrast, both of PIR and VIR ranking strategies improved image retrieval. Furthermore, the performance of the proposed approach (VIR) is overall superior to PIR owing to the removal of some noisy concepts and more reasonable weights associated with concepts. Because some noisy concepts were removed, the final precision at recall level 100 percent of VIR is above that of PIR. Especially, in our experiment some retrieved images by using PIR ranking strategy would never be retrieved by using VIR ranking strategy because the correct annotation keyword, i.e. query term, was accidentally removed as noise. For this situation, we simply removed this image from the image pool. This should not affect our final conclusions, since this happened seldom.

5 Conclusions and Future Work

Due to the limitations of current techniques, image annotations have a poor performance in image retrieval systems. To mitigate this problem, we propose an model which scores each annotated concept using semantic similarity measure based on knowledge-based lexicon WordNet. The experimental results show that the precision is improved to some extent. Moreover, after re-ranking, most correctly annotated images are associated with higher rank. In real life it is reasonable since users often are interest to a first couple of retrieval results. However, some problems still need be further researched. The experimental training data has a limited size of vocabulary, so the annotation results have a low coverage over total keyword space. In addition, the evaluation of image retrieval is conducted on a small set of retrieved results using only top 25 terms, since judging relevancy/irrelevancy to test queries requires substantive human endeavors. A wider evaluation on larger data set will be carried out in future work

References

1. Coelho, T.A.S., Calado, P.P., Souza, L.V., Ribeiro-Neto, B.: Image retrieval using multiple evidence ranking. Image 16(4), 408–417 (2004)
2. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European Conference on Computer Vision-Part IV, pp. 97–112 (2002)
3. Fellbaum, C.: Wordnet: an electronic lexical database. MIT Press, Cambridge (1998)
4. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics, pp. 19–33 (1997)
5. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 706–715 (2005)
6. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proceedings of Advance in Neutral Information Processing (2003)
7. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: First International Workshop on Multimedia Intelligent Storage and Retrieval Management (1999)
8. Sheikholeslami, G.C., Zhang, W.A., Syst, C., Jose, C.A.S.: Semquery: semantic clustering and querying on heterogeneous features for visual data. Knowledge and Data Engineering, IEEE Transactions on 14(5), 988–1002 (2002)