Co-training for search-based automatic image annotation

Yufeng Zhao¹, Yao Zhao¹, Zhenfeng Zhu¹
¹Institute of Information Science, Beijing Jiaotong University
Beijing, China. 100044
snowmanzhao@163.com

ABSTRACT: Search-based automatic image annotation is an effective technology to enhance the performance of annotation. By integrating the co-training technique, this paper addresses on a novel scheme for search-based image annotation, in which two classifiers can contribute to each other during the training phase. Since each classifier can select some most confident images to enhance the generalization ability of the other one, the co-training learning algorithm is triggered out for automatically mining more and more relevant images, which improve the annotation performance greatly. To characterize the various contribution of each relevant image, the probability output of the classification is taken as the corresponding weight. Moreover, the histogram of retrieved keywords is proposed to re-rank the final reliability of the keywords to be annotated, which can also guarantee the scalability property of annotation to some extent. With the promoted search precision on the basis of co-training strategy, the experimental results demonstrate the improvement of annotation performance.

Categories and Subject Descriptors
I4.10 [Image Representation]; I.4.2 [Segmentation]

General Terms
Online learning, recursive learning algorithm

Keywords: Co-training, Automatic image annotation, SVM, Bayesian classifier

Received 19 April 2007; Revised and accepted 26 Oct. 2007

1. Introduction

In content-based image retrieval (CBIR), it is well known that the similarity of the visual features provides less semantic meaningful information. Therefore, automatic image annotation, which is to explore some keywords to describe the image content, has been the key issue for the further development of CBIR. A lot of annotation algorithms have been proposed [1-9], which testify that the technique of automatic image annotation can boost the quality of image retrieval greatly. Generally, the referred algorithms in [1-7] pay emphases on learning the association probability model between the content of image and the keyword. Barnard and Duygulu et al [1] bring forward the co-occurrence model by utilizing machine translation method, in which they treat the image annotation as a process of translation from "visual language" to "text language" and collect the co-occurrence information by the estimating the translation probability. In addition, a cross-media relevance model (CMRM) [2] is promoted via the continuous-space relevance model (CRM) [3] and the multiple Bernoulli relevance model (BMRM) [4]. Some more sophisticated graph models have also been applied to the task of image annotation [5, 6]. Another way to focus on automatic image annotation is based on classification approach [7].

Recently, motivated by the search technology, a data-driven annotation approach turns up to be effective [8, 9]. Given a query image and a labeled keyword, X. J. Wang et al [8] apply the search result cluster (SRC) algorithm into a three-layer annotation model. In [9], an improvement on [8] is made by C. Wang et al, who propose a scalable search-based approach to annotate the web personal images only provided with a query image. One of advantages of search-based image annotation is to avoid a complex supervised learning since it is learned from the labeled textual keywords of retrieved images, which have been labeled accurately. Furthermore, the annotation is taken on a high-level semantic property and is suitable for a scalable image database. However, the average precision of annotation [8, 9] is unsatisfactory because the retrieved labeled images are involved in too many non-relevant images, which lead to incomplete or improper annotation results. Obviously, the retrieval effectiveness is a crucial element in search-based annotation approach to some extent, and directly influences the performance of image annotation. Hence, to give higher retrieval accuracy under current technique, it is prone to applying a semi-supervised learning fashion into the search-based image annotation procedure. Thus, we propose a novel annotation framework based on the co-training strategy, the goal of which is to exploit more relevant images with related to the unlabeled image for annotating. Based on learning two independent classifiers on their additional training set, each classifier can select some most confident images to enhance the generalization ability of the other one. Since each classifier is optimized gradually in a co-training manner, more and more relevant images can be explored automatically during the training process, which directly boosts the annotation performance. In addition, each relevant image makes different contribution for the final annotation result hence a corresponding weight is assigned to it, which is given by the probability output of the corresponding classification. Furthermore, to decide the final reliability of keywords to be annotated, the histogram of retrieved keywords is proposed to guarantee the scalability of annotation. With the promoted search precision based on the co-training strategy, the experimental analyses demonstrate the improvement of annotation performance.

The paper is organized by the following sections. Section 2 introduces the framework of co-training based automatic image annotation and its mathematics expression. The proposed co-training algorithm is dealt with in section 3. The automatic annotation procedure is shown in section 4. In section 5, the performance of our approach is evaluated on the experimental dataset. At last, we conclude the paper and give the future work.
2. The framework of co-training based automatic image annotation

Recently, on the assumption that the keywords of the relevant labeled images are available for annotating the unlabeled image, search-based automatic image annotation is proposed from the viewpoint of search and mining process in [8, 9]. Given an unlabeled image, some relevant images are retrieved from the labeled image database firstly, which is usually based on the low-level visual feature. Then, the annotation is conformed to mining a few keywords \( w^* \) from the labeled keywords of retrieved images for best representing the concept of the relevant image set. The procedure can be reformulated as:

\[
p(w|I_q) = \sum_{I_r \in \Psi_q} p(w|I_r) \cdot p(I_r|I_q)
\]

where \( \Psi_q \) is a set of relevant labeled images to the unlabeled image \( I_q \). \( p(I_r|I_q) \) simulates the search process and \( p(w|I_r) \) denotes the posterior probability of the keyword \( w \) annotating \( I_r \). Obviously, the annotation results can be more satisfactory if the retrieved relevant image set is more accurate to some extent.

As mentioned in [8, 9], the final annotation results are determined by the initial round of retrieval images, in which many non-relevant images are involved into the relevant image set \( \Psi_q \). Hence, much improper labeled keywords are used to annotate the unlabeled image so that the performance of annotation is poor. Based on the current retrieval technology, the main aim of this paper is to make sure that the elements \( \Psi_q \) are as relevant to the unlabeled image as possible, for which the co-training strategy is imported into the task of search-based image annotation in a semi-supervised learning manner.

Due to the higher accuracy and the better performance of prediction for unknown data, the co-training technique has been utilized to resolve many complex computer vision problems. In [10], the co-training technique is applied to obtaining high performance of text retrieval. In Raskutti’s work [11], the co-training scheme is implemented by taking two kinds of feature view. Another way for co-training technology is to form multi-view of features by randomly splitting in original feature space as suggested by Chan et al [12]. As in [13], two independent classifiers are imported into the co-training procedure and thus can be boosted for each other. For our case, the co-training strategy is triggered out for collecting as much relevant images as possible to annotate the unlabeled query image, which is a promising solution to improve the quality of search-based image annotation. Figure 1 illustrates the framework of co-training based image annotation, which contains two main parts: the co-training stage and the annotation stage.

During the co-training stage, the underlying Bayesian classifier and the probabilistic SVM classifier are learnt on the initial training set selected from the first round of retrieval results firstly. Then, by selecting some most confident images to predict for each other, each classifier is re-trained with the additional training set. The procedure is repeated until the maximum time is achieved. During the annotation stage, the final relevant image set \( \Psi_q \) is gained via fusing the corresponding training sets of two classifiers, on which several most representative keywords are mined to characterize it. It is obvious that the annotation contains the high-level semantic property. Since each relevant image takes on different significance, the corresponding weight is concerned and is given by the corresponding output of classification. To capture more reliable annotation quality, the histogram of keywords is proposed to re-rank the keyword list to be annotated, which guarantees the scalable annotation.

3. The co-training scheme

In this paper, the purpose of the co-training algorithm is to mine more and more relevant images for annotating the unlabeled images. In the co-training scheme, the Bayesian classifier and the probabilistic SVM classifier are taken as two underlying classifiers, which can also contribute to each other during the training phase. Provided an unlabeled image, the initial retrieval results are gained according to CBIR technique, from which \( m \) most top-ranked images are chosen as the relevant images on the assumption of most top-ranked images being the relevant data. With \( n \) non-relevant images selected from the remaining database randomly, the initial training set of Bayesian classifier is achieved, which is the same with that of probabilistic SVM classifier. Then, on the base of the corresponding training set, two classifiers are trained independently. Each classifier selects some most confident images to predict for the other classifier, which are added into the other training set for forming an additional training set. Thirdly, two classifiers are re-trained with their additional training set given by the other, and the process repeats until achieving the maximum times. In addition, since the reliability
of each relevant image takes on various contributions to the final annotation list, a weight is assigned, which is developed by the corresponding probabilistic output of its classification. In particular, the initial weight of relevant image is defined as their similarity to the unlabeled image. The pseudo-code of co-training scheme is shown in Table 1.

Here, the SVM classifier with RBF kernel function is applied to the semi-supervised learning algorithm. Since our focus is not on the selection of the parameters for SVM, the regulating constant $C$ and the variance of the RBF kernel are set to 500 and 0.4 respectively. For the Bayesian classifier, Gaussian distribution is used to characterize the relevant images and the uniform distribution is utilized to represent the non-relevant images. With the end of co-training procedure, the main work is focus on how to fuse two additional training sets ($T^K_B$ and $T^K_S$) to generate the final relevant set $\Psi_q$ of the unlabeled image $I_q$. Here, to simplify computation, $\Psi_q$ can be built by the union of relevant images in $T^K_B$ and $T^K_S$. With a few labeled images, each classifier can accurately find some most confident images to be added into the other training set, which means the co-operative training for each other. With the expanded training set, more and more relevant images are collected, which are helpful for annotating the unlabeled image.

### 4. Automatic annotation for images

For the obtained relevant image set $\Psi_q$, the next work emphasizes on exploiting some best representative keywords to describe it, which is implemented by computing the posterior probability $p(w|I_q)$ of each keyword annotating unlabeled image $I_q$ as shown in equation (1). Taking it as the criterion, the final annotation can be determined automatically. To simplify the computation, the part $p(w|I_q)$ in equation (1) is defined as:

$$p(w|I_q) = \begin{cases} 1 & \text{if } w \text{ appearing in } I_q \\ 0 & \text{else} \end{cases} \quad (2)$$

The part $p(I_q|w)$ in equation (1) is concerned on the different similarity of each retrieved image to the unlabeled image, which is just the corresponding weight of each relevant image and is given by the probabilistic output of its classification.

For the keyword itself, the reliability property to be annotated is different, which is higher if the appearing times in the textual information $\Psi_q$ of. Thus, the histogram of keywords from $\Psi_q$ is investigated to represent the annotation reliability, which is defined as reliability factor. By the reliability factor, the annotation list is re-ranked, which guarantees the scalability property of annotation approach. The final ranking function can be $f(w)$ defined as:

$$f(w) = h(w) \cdot p(w|I_q) \quad (3)$$

where $h(w)$ is the reliability factor and is equal to the frequency of the keyword $w$ appearing in $\Psi_q$. According to the output of $f(w)$, two annotation strategies can be applied. When the output is more than the fore-defined experimental value, the keyword is chosen as the annotation result, which takes on the scalable property of annotation. Another way is to determine the final annotation results according to the five most top-ranked posterior probabilities. The pseudo-code of automatic image annotation algorithm can be found in Table 2.
From the automatic image annotation process, it is obvious that the results are completely mined from the labeled textual keywords of the relevant images, which takes on the semantic level information. Furthermore, the annotation quality can be less influenced although the lexicon is enlarged, which is more suitable for the lager image database.

5. Experimental results and analyses

We test the proposed algorithm on the Corel dataset from Barnard et al. [1], which is extensively used as basic comparative data for recent researches in image annotation. The experimental data set comprises 5,000 images, in which 4,500 images are used as training set and the remaining 500 images as testing set. For each image, 36-dimensional visual features, which characterize the color histograms in HSV color feature space, are extracted for characterizing its global information. In addition, each image is annotated with 1 to 5 keywords, and totally 374 keywords have been used in annotations, which are composed of the vocabulary. To evaluate the annotation performance, we also measure the image annotation results by using the average annotation recall and precision defined in [4].

\[
\text{recall} = \frac{B}{A} \quad \text{and} \quad \text{precision} = \frac{B}{C}
\]

where \(A\) is the number of images automatically annotated with a given word in the top 10 returned word list; \(B\) is the number of images correctly annotated with that word in the top 10 returned word list; and \(C\) is the number of images having that word in ground truth annotation.

5.1. Comparison with other annotation models

Table 1 shows the effectiveness and the promise of our co-training based annotation model, which is compared with MBRM [4] and SBIA [9]. MBRM focuses on learning the association probability model between the content of image and the keyword via a supervised learning algorithm. SBIA is one of search-based annotation models and is implemented on the common dataset.

For all keywords in the lexicon, the systematic evaluation results are reported in table 3, which illustrates the search-based automatic image annotation is better than the previous MBRM. Meanwhile, the average recall and average precision of our annotation model outperforms that of SBIA greatly. It is believed that the retrieved non-relevant images are more in SBIA than that in our model, which can verify that the co-training strategy can exploit as much relevant images as possible for annotating the unlabeled image. Moreover, by using the reliability factor to re-rank the annotation list, the proposed model can improve the quality to some extent. Both of the above mentioned advantages make contribution to the greater performance.

5.2. Annotation performance on different learning algorithm

For exploring more relevant images, many learning algorithms can be applied into the search-based annotation procedure such as the co-training algorithm, the self-training algorithm \(k\)-\(k\)-NN and algorithm. To show the good performance of co-training algorithm is better than the others, this experiment is designed, in which the self-training process is based on the underlying SVM classifier. The method of \(k\)-\(k\)-NN is to retrieve \(k\) top-ranked images to annotate the unlabeled image. The curve of average precision on different number of lexicon is shown in figure 2 according to the three learning algorithm.

From the figure, we can find out that the co-training algorithm is better than the others. Compared with the self-training algorithm, co-training algorithm can mine more relevant images from the labeled database since two classifiers can co-operative training for each other. However, the method of \(k\)-\(k\)-NN cannot use the unknown data, which leads to the relevant image set containing more non-relevant images and unavailable keywords. In addition, the annotation performance is less influenced when the lexicon is enlarged, which can lead to an unlimited lexicon.

5.3. Evaluation of reliability factor

As shown in the equation (3), the task of annotation is determined automatically, in which the reliability factor extremely makes an impact on the final annotation results. To demonstrate the fact, one keyword is regarded as one query, and the average retrieval precision for 10 keywords is shown in Figure 3.

<table>
<thead>
<tr>
<th>Models</th>
<th>MBRM</th>
<th>SBIA</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>#words with recall ( \geq 0 )</td>
<td>122</td>
<td>153</td>
<td>221</td>
</tr>
<tr>
<td>Results on all 374 keywords</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Per-word Recall</td>
<td>0.25</td>
<td>0.33</td>
<td>0.387</td>
</tr>
<tr>
<td>Average Per-word Precision</td>
<td>0.24</td>
<td>0.27</td>
<td>0.350</td>
</tr>
</tbody>
</table>

Table 3. The performance comparison among different models
From the figure 3, we can find that the annotation is improved when the reliability factor is taken into the ranking function. The reason is that the reliability factor can leverage the weight of the keywords in retrieved images. In particular, the reliability factor can enhance the weight of the useful keyword, which is important to annotate the unlabeled image.

Figure 3. The average retrieval precision for 10 keywords

6. Conclusions
In this paper, we have developed a new search-based annotation model, in which a co-training learning algorithm is carried out to explore more relevant images for annotating the unlabeled image. By expanding accurately relevant image set, the quality of annotation can be boosted greatly. The main merit of co-training based annotation approach is to make the annotation adopting more semantic information. Besides, the reliability factor is considered to re-rank the annotation list, in which the textual property of keywords is also taken into account. The experimental results have tested its effectiveness. In addition, the co-training method is also suitable for the scalable image database, such as the web images, which is our future work.

Acknowledgments
This work was supported in part by NSFC (No. 06020307No. 90604032), 973 National Basic Research Project (2006CB303104), Program for New Century Excellent Talents in University, Open foundation of National Laboratory of Pattern Recognition, Research Foundation of Beijing Jiaotong University (No.2005SS005, No. 2005SM013).

References


