Automatic Image Annotation Using Maximum Entropy Model

Wei Li and Maosong Sun

State Key Lab of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University,
Beijing 100084, China
wei.lee04@gmail.com, sms@mail.tsinghua.edu.cn

Abstract. Automatic image annotation is a newly developed and promising technique to provide semantic image retrieval via text descriptions. It concerns a process of automatically labeling the image contents with a pre-defined set of keywords which are exploited to represent the image semantics. A Maximum Entropy Model-based approach to the task of automatic image annotation is proposed in this paper. In the phase of training, a basic visual vocabulary consisting of blob-tokens to describe the image content is generated at first; then the statistical relationship is modeled between the blob-tokens and keywords by a Maximum Entropy Model constructed from the training set of labeled images. In the phase of annotation, for an unlabeled image, the most likely associated keywords are predicted in terms of the blob-token set extracted from the given image. We carried out experiments on a medium-sized image collection with about 5000 images from Corel Photo CDs. The experimental results demonstrated that the annotation performance of this method outperforms some traditional annotation methods by about 8% in mean precision, showing a potential of the Maximum Entropy Model in the task of automatic image annotation.

1 Introduction

Last decade has witnessed an explosive growth of multimedia information such as images and videos. However, we can't access to or make use of the relevant information more leisurely unless it is organized so as to provide efficient browsing and querying. As a result, an important functionality of next generation multimedia information management system will undoubtedly be the search and retrieval of images and videos on the basis of visual content.

In order to fulfill this "intelligent" multimedia search engines on the world-wide-web, content-based image retrieval techniques have been studied intensively during the past few years. Through the sustained efforts, a variety of state-of-the-art methods employing the query-by-example (QBE) paradigm have been well established. By this we mean that queries are images and the targets are also images. In this manner, visual similarity is computed between user-provided image and database images based on the low-level visual features such as color, texture, shape and spatial relationships. However, two important problems still remain. First, due to the limitation of object recognition and image understanding, semantics-based image segmentation algorithm

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 34 – 45, 2005.

is unavailable, so segmented region may not correspond to users' query object. Second, visual similarity is not semantic similarity which means that low-level features are easily extracted and measured, but from the users' point of view, they are non-intuitive. It is not easy to use them to formulate the user's needs. We encounter a so-called semantic gap here. Typically the starting point of the retrieval process is the high-level query from users. So extracting image semantics based on the low-level visual features is an essential step. As we know, semantic information can be represented more accurately by using keywords than by using low-level visual features. Therefore, building relationship between associated text and low-level image features is considered to an effective solution to capture the image semantics. By means of this hidden relationship, images can be retrieved by using textual descriptions, which is also called query-by-keyword (QBK) paradigm. Furthermore, textual queries are a desirable choice for semantic image retrieval which can resort to the powerful text-based retrieval techniques. The key to image retrieval using textual queries is image annotation. But most images are not annotated and manually annotating images is a time-consuming, error-prone and subjective process. So, automatic image annotation is the subject of much ongoing research. Its main goal is to assign descriptive words to whole images based on the low-level perceptual features, which has been recognized as a promising technique for bridging the semantic gap between low-level image features and high-level semantic concepts.

Given a training set of images labeled with text (e.g. keywords, captions) that describe the image content, many statistical models have been proposed by researchers to construct the relation between keywords and image features. For example, co-occurrence model, translation model and relevance-language model. By exploiting text and image feature co-occurrence statistics, these methods can extract hidden semantics from images, and have been proven successful in constructing a nice framework for the domain of automatic image annotation and retrieval.

In this paper, we propose a novel approach for the task of automatic image annotation using Maximum Entropy Model. Though Maximum Entropy method has been successfully applied to a wide range of application such as machine translation, it is not much used in computer vision domain, especially in image auto annotation.

This paper is organized as follows: Section 2 presents related work. Section 3 describes the representation of labeled and unlabeled images, gives a brief introduction to Maximum Entropy Model and then details how to use it for automatically annotating unlabeled images. Section 4 demonstrates our experimental results. Section 5 presents conclusions and a comment for future work.

2 Related Work

Recently, many statistical models have been proposed for automatic image annotation and retrieval. The work of associating keywords with low-level visual features can be addressed from two different perspectives.

2.1 Annotation by Keyword Propagation

This kind of approach usually formulates the process of automatic image annotation as one of supervised classification problems. With respect to this method, accurate annotation information is demanded. That is to say, given a set of training images labeled with semantic keywords, detailed labeling information should be provided. For example, from training samples, we can know which keyword corresponds to which image region or what kind of concept class describes a whole-image. So each or a set of annotated keyword can be considered as an independent concept class, followed by training each class model with manually labeled images, then the model is applied to classify each unlabeled image into a relevant concept class, and finally producing annotation by propagating the corresponding class words to unlabeled images.

Wang and Li [8] introduced a 2-D multi- resolution HMM model to automate linguistic indexing of images. Clusters of fixed-size blocks at multiple resolution and the relationships between these clusters is summarized both across and within the resolutions. To annotate the unlabeled image, words of the highest likelihood is selected based on the comparison between feature vectors of new image and the trained concept models. Chang et al [5] proposed content-based soft annotation (CBSA) for providing images with semantic labels using (BPM) Bayesian Point Machine. Starting with labeling a small set of training images, an ensemble of binary classifier for each keyword is then trained for predicting label membership for images. Each image is assigned one keyword vector, with each keyword in the vector assigned a confidence factor. In the process of annotation, words with high confidence are considered to be the most likely descriptive words for the new images. The main practical problem with this kind of approaches is that a large labeled training corpus is needed. Moreover, during the training and application stages, the training set is fixed and not incremented. Thus if a new domain is introduced, new labeled examples must be provided to ensure the effectiveness of such classifiers.

2.2 Annotation by Statistical Inference

More recently, there have been some efforts to solve this problem in a more general way. The second approach takes a different strategy which focuses on discovering the statistical links between visual features and words using unsupervised learning methods. During training, a roughly labeled image datasets is provided where a set of semantic labels is assigned to a whole image, but the word-to-region information is hidden in the space of image features and keywords. So an unsupervised learning algorithm is usually adopted to estimate the joint probability distribution of words and image features.

Mori et al [4] were the earliest to model the statistics using a co-occurrence probabilistic model, which predicate the correct probability of associating keywords by counting the co-occurrence of words with image regions generated using a fixed-size blocks. Blocks are vector quantized to form clusters which inherit the whole set of keywords assigned to each image. Then clusters are in turn used to predict the keywords for unlabeled images. The disadvantage is that the model is a little simple and the rough fixed-size blocks are unable to model objects effectively, leading to poor annotation accuracy. Instead of using fixed-size blocks, Barnard et al [1] performed Blobworld segmentation and Normalized cuts to produce semantic meaningful regions. They constructed a hierarchical model via EM algorithm. This model combines both asymmetric clustering model which maps words and image regions into clusters and symmetric clustering model which models the joint distribution of words and regions. Duygulu et al [2] proposed a translation model to map keywords to individual image regions. First, image regions are created by using a segmentation algorithm. For each region, visual features are extracted and then blob-tokens are generated by clustering the features for each region across whole image datasets. Each image can be represented by a certain number of these blob-tokens. Their Translation Model uses machine translation model Iof IBM to annotate a test set of images based on a large number of annotated training images. Another approach using cross-media relevance models (CMRM) was introduced by Jeon et al [3]. They assumed that this could be viewed as analogous to the cross-lingual retrieval problem and a set of keywords $\{w_1, w_2, ..., w_n\}$ is related to the set of blob-tokens $\{b_1, b_2, ..., b_n\}$, rather than one-to-one correspondence between the blob-tokens and keywords. Here the joint distribution of blob-tokens and words was learned from a training set of annotated images to perform both automatic image annotation and ranked retrieval. Jeon et al [9] introduced using Maximum Entropy to model the fixed-size block and keywords, which gives us a good hint to implement it differently. Lavrenko et al [11] extended the cross-media relevance model using actual continuous-valued features extracted from image regions. This method avoids the clustering and constructing the discrete visual vocabulary stage.

3 The Implementation of Automatic Annotation Model

3.1 The Hierarchical Framework of Automatic Annotation and Retrieval

The following Fig. 1 shows the framework for automatic image annotation and keyword-based image retrieval. Given a training dataset of images labeled with keywords. First, we segment a whole image into a collection of sub-images, followed by extracting a set of low-level visual features to form a feature vector to describe the visual content of each region. Second, a visual vocabulary of blob-tokens is generated by clustering all the regions across the whole dataset so that each image can be represented by a number of blob-tokens from a finite set of visual symbols. Third, both textual information and visual information is provided to train the Maximum Entropy model, and the learned model is then applied to automatically generate keywords to describe the semantic content of an unlabeled image based on the low-level features. Consequently, both the users' information needs and the semantic content of images can be represented by textual information, which can resort to the powerful text IR techniques to implement this cross-media retrieval, suggesting the importance of textual information in semantics-based image retrieval.

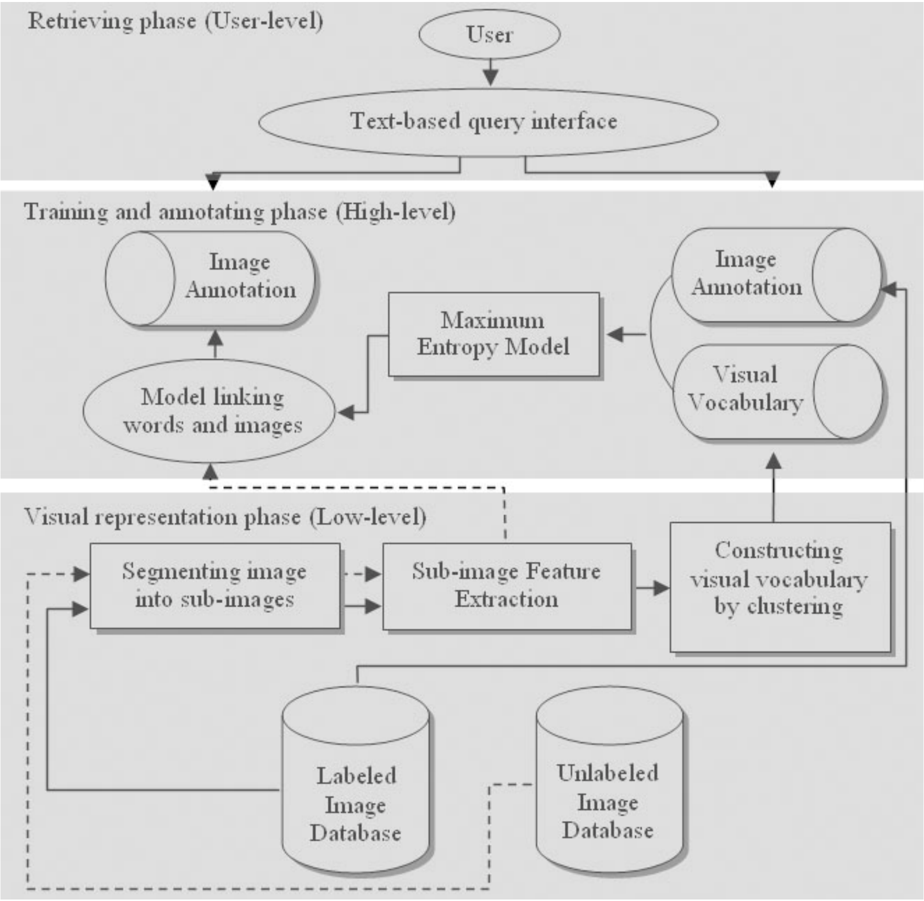

Retrieving phase (User-level)
User
Text-based query interface
Training and annotating phase (High-level)
Image Annotation
Maximum Entropy Model
Model linking words and images
Image Annotation
Visual Vocabulary
Visual representation phase (Low-level)
Segmenting image into sub-images
Sub-image Feature Extraction
Constructing visual vocabulary by clustering
Labeled Image Database
Unlabeled Image Database

Fig. 1. Hierarchical Framework of Automatic Annotation and Retrieval

⟶ learning correlations between blob-tokens and textual annotations

- - -▶ applying correlations to generate annotations for unlabeled images

3.2 Image Representation and Pre-processing

A central issue in content-based image annotation and retrieval is how to describe the visual information in a way compatible with human visual perception. But until now, no general framework is proposed. For different tasks and goals, different low-level features are used to describe and analyze the visual content of images. On the whole, there are two kinds of interesting open questions remain unresolved. First, what feature sets should be selected to be the most expressive for any image region. Second, how blob-tokens can be generated, that is to say, how can one create such a visual vocabulary of blob-tokens to represent each image in the collection using a number of symbols from this finite set? In our method, we carried out these following two steps: First, segment images into sub-images, Second, extract appropriate features for any sub-images, cluster similar regions by k-means and then use the centroid in each cluster as a blob-token. The first step can be employed by either using a segmentation algorithm to produce semantically meaningful units or partitioning the image into fixed-size rectangular grids. Both methods have pros and cons, a general purpose segmentation algorithm may produce semantic regions, but due to the limitation in computer vision and image processing, there are also the problems of erroneous and unreliable region segmentation. The advantage of regular grids is that is does not need to perform complex image segmentation and is easy to be conducted. However, due to rough fixed-size rectangular grids, the extracted blocks are unable to model objects effectively, leading to poor annotation accuracy in our experiment.

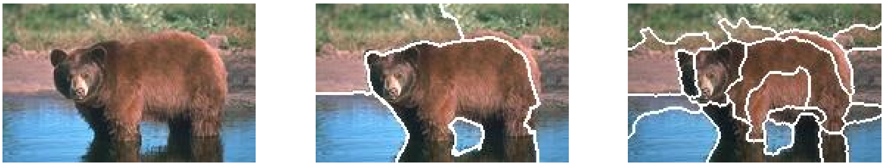


Fig. 2. Segmentation Results using Normalized cuts and JSEG

In this paper, we segment images into a number of meaningful regions using Normalized cuts [6] against using JSEG. Because the JSEG is only focusing on local features and their consistencies, but Ncuts aims at extracting the global impression of an image data. So Ncuts may get a better segmentation result than JSEG. Fig. 2 shows segmentation result using Normalized cuts and JSEG respectively, the left is the original image, the mid and the right are the segmentation result using Ncuts and JSEG respectively. After segmentation, each image region is described by a feature vector formed by HSV histograms and Gabor filters. Similar regions will be grouped together based on k-means clustering to form the visual vocabulary of blob-tokens. Too much clusters may cause data sparseness and too few can not converge. Then each of the labeled and unlabeled images can be described by a number of blob-tokens, instead of the continuous-valued feature vectors. So we can avoid the image data modeling in a high-dimensional and complex feature space.

3.3 The Annotation Strategy Based on Maximum Entropy

Maximum Entropy Model is a general purpose machine learning and classification framework whose main goal is to account for the behavior of a discrete-valued random process. Given a random process whose output value *y* may be influenced by some specific contextual information *x*, such a model is a method of estimating the conditional probability.

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{j=1}^{k} \alpha_j^{f_j(x,y)} \tag{1}$$

In the process of annotation, images are segmented using normalized cuts, every image region is represented by a feature vector consisting of HSV color histogram and the Gabor filters, and then a basic visual vocabulary containing 500 blob-tokens is generated by k-means clustering. Finally, each segmented region is assigned to the label of its closest blob-token. Thus the complex visual contents of images can be represented by a number of blob-tokens. Due to the imbalanced distribution of keywords frequency and the data sparseness problem, the size of the pre-defined keyword vocabulary is reduced from 1728 to 121 keywords, by keeping only the keywords appearing more than 30 times in the training dataset.

We use a series of feature function $f_{\text{FC, Label}}(b_i, w_j)$ to model the co-occurrence statistics of blob-tokens b_i and keywords w_j, where FC denote the context of feature constraints for each blob-token. The following example represents the co-occurrence of the blob-token b_* and the keyword "water" in an image *I*.

$$f_{\text{FC}_{\text{w}},\ \text{water}}(b_i, w_j) = \begin{cases} 1 & if\ w_j == 'water'\ and\ FC_w(b_i) == true \\ 0 & otherwise \end{cases} \quad (2)$$

If blob-token b_i satisfies the context of feature constraints and keyword "water" also occurs in image *I*. In other words, if the color and texture feature components are coordinated with the semantic label 'water', and then the value of the feature function is 1, otherwise 0.

The following Fig. 3 shows the annotation procedure that using MaxEnt captures the hidden relationship between blob-tokens and keywords from a roughly labeled training image sets.

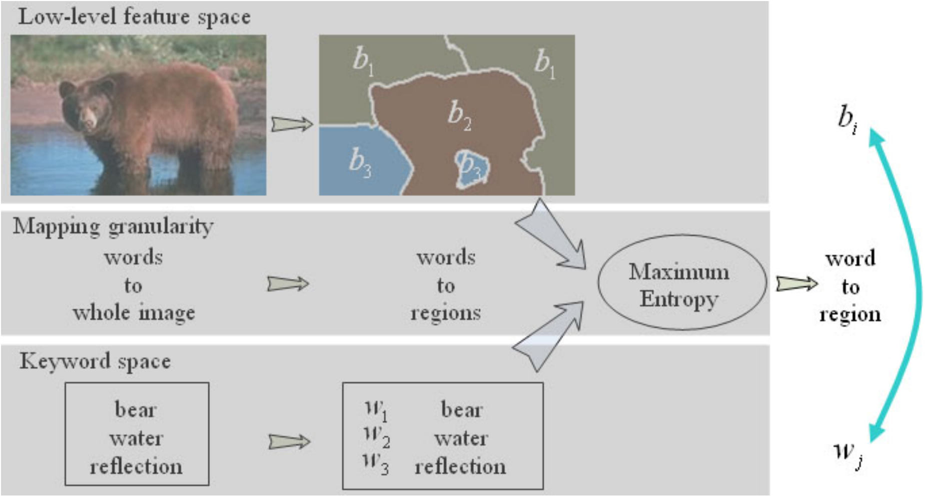


Fig. 3. Learning the statistics of blob-tokens and words

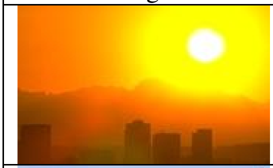
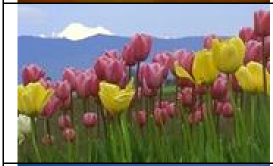
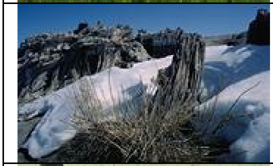
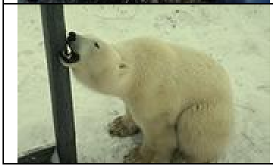
In the recent past, many models for automatic image annotation are limited by the scope of the representation. In particular, they fail to exploit the context in the images and words. It is the context in which an image region is placed that gives it meaningful interpretation.

In our annotation procedure, each annotated word is predicted independently by the Maximum Entropy Model, word correlations are not taken into consideration. However, correlations between annotated words are essentially important in predicting relevant text descriptions. For example, the words “trees” and “grass” are more likely to co-occur than the words “trees” and “computers”. In order to generate appropriate annotations, a simple language model is developed that takes the word-correlation information into account, and then the textual description is determined not only by the model linking keywords and blob-tokens but also by the word-to-word correlation. We simply count the co-occurrence information between words in the predefined textual set to produce a simple word correlation model to improve the annotation accuracy.

4 Experiments and Analysis

We carried out experiments using a mid-sized image collection, comprising about 5,000 images from Corel Stock Photo CDs, 4500 images for training and 500 for testing. The following table 1 shows the results of automatic image annotation using Maximum Entropy.

Table 1. Automatic image annotation results

Images	Original Annotation	Automatic Annotation
	sun city sky mountain	Sun sky mountain clouds
	flowers tulips mountain sky	Flowers sky trees grass
	tufa snow sky grass	snow sky grass stone
	polar bear snow post	bear snow sky rocks

For our training datasets, the visual vocabulary and the pre-defined textual set contain 500 blob-tokens and 121 keywords respectively, so the number of the training pairs (b_i, w_j) is 60500. After the procedure of feature selection, only 9550 pairs left. For model parameters estimation, there are a few algorithms including Generalized Iterative Scaling and Improved Iterative Scaling which are widely used. Here we use Limited Memory Variable Metric method which has been proved effective for Maximum Entropy Model [10]. Finally, we can get the model linking blob-tokens and keywords, and then the trained model $p(y|x)$ is applied to predict textual annotations $\{w_1, w_2, \ldots, w_n\}$ given an unseen image formed by $\{b_1, b_2, \ldots, b_m\}$.

To further verify the feasibility and effectiveness of Maximum Entropy model, we have implemented the co-occurrence model as one of the baselines whose conditional probability $p(w_j|b_i)$ can be estimated as follows:

$$p(w_j|b_i) = \frac{p(b_i|w_j)p(w_i)}{\sum_{k=1}^{N} p(b_i|w_k)p(w_i)} \approx \frac{(m_{ij}/n_j)(n_j/N)}{\sum_{k=1}^{N}(m_{ik}/n_k)(n_k/N)} = \frac{m_{ij}}{\sum_{k=1}^{N} m_{ik}} = \frac{m_{ij}}{M_i} \quad (3)$$

Where m_{ij} denote the co-occurrence of b_i and w_j, n_j denote the occurring number of w_j in the total *N* words.

The following Fig. 4 shows the some of the retrieval results using the keyword '*water*' as a textual query.

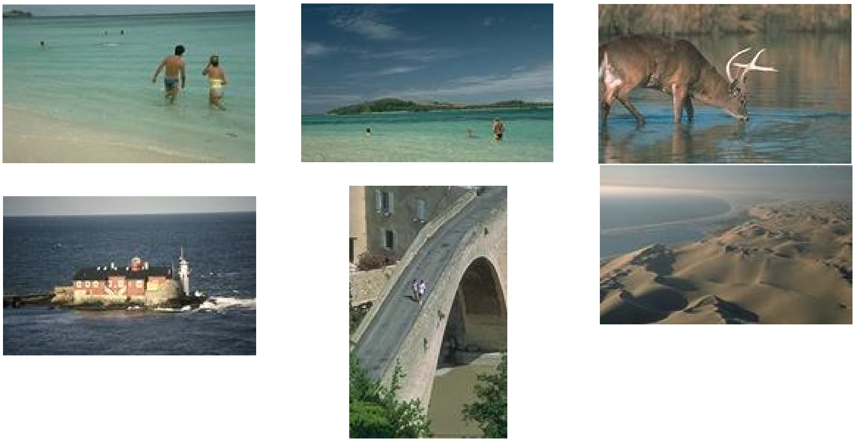


Fig. 4. Some of retrieved images using '*water*' as a query

The following Fig. 5 and Fig. 6 show the precision and recall of using a se of high-frequency keywords as user queries. We implemented two statistical models to link blob-tokens and keywords.

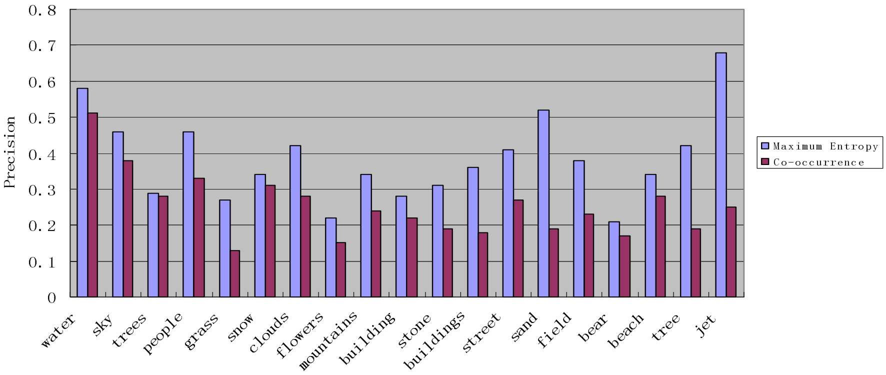


Fig. 5. Precision of retrieval using some high-frequency keywords

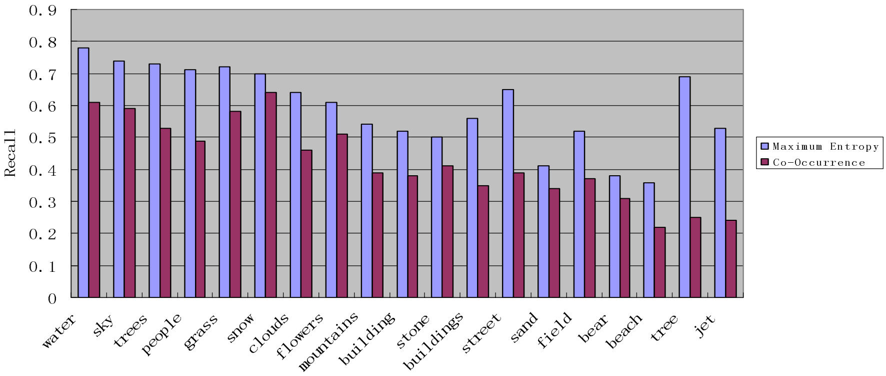


Fig. 6. Recall of retrieval using some high-frequency keywords

The annotation accuracy is evaluated by using precision and recall indirectly. After posing a keyword query for images, the measure of precision and recall can be defined as follows:

$$precision = \frac{A}{A+B} \qquad recall = \frac{A}{A+C} \qquad (4)$$

Where *A* denote the number of relevant images retrieved, *B* denote the number of irrelevant images retrieved, *C* denote the number of relevant images not retrieved in the image datasets, and images whose labels containing the query keyword is considered relevant, otherwise irrelevant.

Table 2. Experimental results with average precision and mean

Method	Mean precision	Mean recall
Co-occurrence	0.11	0.18
Maximum Entropy	0.17	0.25

The above experimental results in table 2 show that our method outperforms the Co-occurrence model [4] in the average precision and recall. Since our model uses the blob-tokens to represent the contents of the image regions and converts the task of automatic image annotation to a process of translating information from visual language (blob-tokens) to textual language (keywords). So Maximum Entropy Model is a natural and effective choice for our task, which has been successfully applied to the dyadic data in which observations are made from two finite sets of objects. But disadvantages also exist. There are two fold problems to be considered. First, since Maximum Entropy is constrained by the equation $p(f) = \tilde{p}(f)$, which assumes that the expected value of output of the stochastic model should be the same as the expected value of the training sample. However, due to the unbalanced distribution of keywords frequency in the training subset of Corel data, this assumption will lead to an undesirable problem that common words with high frequency are usually associated with too many irrelevant blob-tokens, whereas uncommon words with low frequency have little change to be selected as annotations for any image regions, consider word "sun" and "apple" , since both words may be related to regions with "red" color and "round" shape, but it is difficult to make a decision between the word "sun" and "apple". However, since "sun" is a common word as compared to "apple" in the lexical set, the word "sun" will definitely used as the annotation for these kind of regions. To address this kind of problems, our future work will mainly focus on the more sophisticated language model to improve the statistics between image features and keywords. Second, the effects of segmentation may also affect the annotation performance. As we know, semantic image segmentation algorithm is a challenging and complex problem, current segmentation algorithm based on the low-level visual features may break up the objects in the images, that is to say, segmented regions do not definitely correspond to semantic objects or semantic concepts, which may cause the Maximum Entropy Model to derive a wrong decision given an unseen image.

5 Conclusion and Future Work

In this paper, we propose a novel approach for automatic image annotation and retrieval using Maximum Entropy Model. Compared to other traditional classical methods, the proposed model gets better annotation and retrieval results. But three main challenges are still remain:

1) Semantically meaningful segmentation algorithm is still not available, so the segmented region may not correspond to a semantic object and region features are insufficient to describe the image semantics.
2) The basic visual vocabulary construction using k-means is only based on the visual features, which may lead to the fact that two different semantic objects with similar visual features fall into the same blob-token. This may degrade the annotation quality.
3) Our annotation task mainly depend on the trained model linking image features and keywords, the spatial context information of image regions and the word correlations are not fully taken into consideration.

In the future, more work should be done on image segmentation techniques, clustering algorithms, appropriate feature extraction and contextual information between regions and words to improve the annotation accuracy and retrieval performance.

Acknowledgements

We would like to express our deepest gratitude to Kobus Barnard and J.Wang for making their image datasets available. This research is supported by the National Natural Science Foundation of China under grant number 60321002 and the National 863 Project of China under grant number 2001AA114210-03, and the ALVIS Project co-sponsored by EU PF6 and NSFC.

References

1. K. Barnard, P. Dyugulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. Journal of Machine Learning Research, 3: 1107-1135, 2003.
2. P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Ojbect recognition as machine translation: Learning a lexicon fro a fixed image vocabulary. In Seventh European Conf. on Computer Vision, 97-112, 2002.
3. J. Jeon, V. Lavrenko and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th intl. SIGIR Conf, 119-126, 2003.
4. Y. Mori, H. Takahashi, and R. Oka, Image-to-word transformation based on dividing and vector quantizing images with words. First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.
5. Edward Chang, Kingshy Goh, Gerard Sychay and Gang Wu. CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuts and Systems for Video Technology Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Descriptions, 13(1): 26-38, 2003.
6. J. shi and J. Malik. Normalized cuts and image segmentation. IEEE Transactions On Pattern Analysis and Machine Intelligence, 22(8): 888-905, 2000.
7. A. Berger, S. Pietra and V. Pietra. A maximum entropy approach to natural language processing. In Computational Linguistics, 39-71, 1996.
8. J. Li and J. A. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on PAMI, 25(10): 175-1088, 2003.
9. Jiwoon Jeon, R. Manmatha. Using maximum entropy for automatic image annotation. In proceedings of third international conference on image and video retrieval, 24-31, 2004.
10. Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of the 6th Workshop on Computational Language Learning, 2003.
11. V. Lavrenko, R. Manmatha and J. Jeon. A model for learning the semantics of pictures. In Proceedings of the 16th Annual Conference on Neural Information Processing Systems, 2004.