Region Based Image Annotation Through Multiple-Instance Learning

Changbo Yang, Ming Dong and Farshad Fotouhi
Department of Computer Science
Wayne State University
Detroit, MI 48202
cbyang, mdong, fotouhi@cs.wayne.edu

ABSTRACT

In an annotated image database, keywords are usually associated with images instead of individual regions, which poses a major challenge for any region based image annotation algorithm. In this paper, we propose to learn the correspondence between image regions and keywords through Multiple-Instance Learning (MIL). After a representative image region has been learned for a given keyword, we consider image annotation as a problem of image classification, in which each keyword is treated as a distinct class label. The classification problem is then addressed using the Bayesian framework. The proposed image annotation method is evaluated on an image database with $5,000$ images.

Categories and Subject Descriptors
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis-object recognition, H.2.8 [**Database Management**]: Database Applications - image databases.

General Terms:
Algorithms, Measurement, Experimentation

Keywords: Automatic image annotation, Multiple-Instance Learning

1. INTRODUCTION

Conventional content-based image retrieval (CBIR) systems require the user to retrieve images based on low-level image attributes such as color, texture, etc. Ideally, users would prefer querying an image database by performing semantic querying without a need to know the contents of the images in the database. Semantics can be represented more accurately by using words than by using low-level features. Consequently, image annotation has received extensive attention recently [1]. However, manual annotation is expensive and does not scale well when the volume of data is very large. A computer program that can perform automatic semantic annotation of images is highly desired to handle the massive digital image resources.

MM'05, November 6–11, 2005, Singapore.

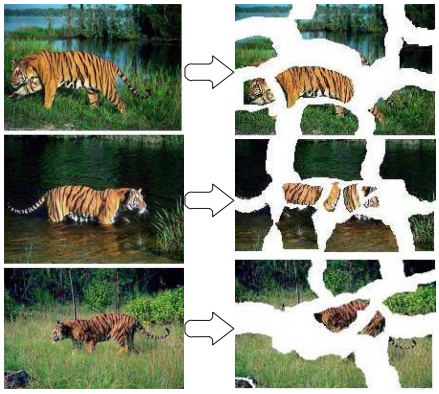

Figure 1: Three sample images of "tiger" (left column) and their segmented regions (right column). A large number of irrelevant noisy regions, such as "grass", "water", and "bush", exist in the training set for keyword "tiger".

An image contains several regions and each region may have different contents and represent different semantic meaning, it is intuitive to divide an image into regions and extract visual features from each region in the first step of automatic image annotation. A statistical model is then learnt from a set of annotated training images to link image regions to keywords and produce the annotation for a testing image. However, a major hurdle remains in the aforementioned learning infrastructure. With few exceptions, the annotation information for a training image is available only at the concept level, but NOT at the content level [2]. In other words, keywords are associated with images instead of individual regions. For example, the left column of Figure 1 shows three images of "tiger" and the right column shows the corresponding image regions segmented using Normalized-cuts [3]. To find the correct correspondence between an image region and the keyword "tiger", a learner must be able to differentiate "tiger" regions from other noisy regions at the first place.

In this paper, we propose to learn the correspondence between image regions and keywords through Multiple-Instance Learning (MIL) [4]. After a representative image region has been selected by MIL for a given keyword, we consider image annotation as a problem of image classification. The classification problem is then addressed using the Bayesian framework. The major contributions of this work include:

- Even though MIL has been previously used for image categorization and retrieval [5], this is the first paper that applies MIL to region based image annotation. The proposed image annotation algorithm is able to learn an explicit correspondence between image regions and keywords, while other existing annotation models either can not or only find an implicit correspondence [6].

- We model image annotation as a problem of image classification under Bayesian framework. This modelling compares favorably with other clustering based annotation approaches, whose performance is strongly influenced by the quality of clustering [7].

2. RELATED WORKS

Starting from a training set of annotated images, many statistical learning models have been proposed in the literature to associate region based visual features with semantic concepts (keywords). Mori et al. [8] developed a co-occurrence model, in which they looked at the co-occurrence of keywords with image regions created using a grid. Duygulu et al. [6] proposed to describe images using a vocabulary of blobs. Each image is generated by using a certain number of these blobs. The Translation Model assumes that image annotation can be viewed as the task of translating from a vocabulary of blobs to a vocabulary of words. Given a set of annotated training images. To our best knowledge, the machine translation algorithm is the only existing model that integrates correspondence learning into image annotation. More recently, Jeon et al. [9] introduced a cross-media relevance model that learns the joint distribution of a set of regions (blobs) and a set of keywords rather than the correspondence between a single region (blob) and a single keyword. The readers are referred to [1] for a comprehensive review on this topic.

Most existing annotation methods share the same two-step procedure in tacking this problem: (1) clustering image regions to region clusters; and (2) finding joint probability of concepts and region clusters. Since these approaches rely on clustering as the basis for image annotation, the performance of annotation is strongly influenced by the quality of clustering [10]. Currently, most approaches perform region clustering based on visual features and suffer from the semantic gap. Here we model image annotation as a problem of image classification under Bayesian framework [11]. With the aids of the annotation assigned to training images, we are able to identify relevant images and irrelevant images for a given keyword by imposing strict semantic constrains on the data.

Another annotation performance bottleneck is the lack of annotation at the content level in the training images. The training image set usually does not provide explicit correspondence between keywords and regions – the keywords are associated with images instead of individual regions. To address this problem, we formulate a MIL based image annotation model that is able to infer the correspondence between regions and keywords in its learning process. Moreover, the images with same semantic meaning usually share some common low-level features. For example, mountains may share similar shape features. For a particular concept, besides the representative image region, a set of features that can best describe the concept is also determined.

3. REGION BASED IMAGE ANNOTATION THROUGH MIL

3.1 Bayesian Framework

In this section, we show how to address the image annotation as a classification problem under the Bayesian framework.

Let J denote the testing set of un-annotated images, and let T denote the training collection of annotated images. Each testing image $I \in J$ is represented by its regions' visual feature $r = \{r_1 \dots r_m\}$, and each training image $I \in T$ is represented by both a set of regions' visual feature $r = \{r_1 \dots r_m\}$ and a word list $w = \{w_1 \dots w_n\}$, where r_j $(j = 1 \dots m)$ is the set of visual features for region j and w_i $(i = 1 \dots n)$ is the i^{th} word in the vocabulary V.

If we treat each word w_i as a distinct class label, the annotation problem can be formulated as a supervised classification problem under Bayesian framework [11]. Since a testing image is divided into many regions, if at least one region in I has the semantic meaning of w_i, we can annotate the image by w_i. For image classification, the testing image I is classified to class $\hat{w}$ based on the *maximum a posteriori* (MAP) criterion as follows,

$$\begin{aligned} \hat{w} &= arg \max_{w_i \in V} \{P(w_i|I)\} \\ &= arg \max_{w_i \in V} \{\max_{r_j \in I} P(r_j|w_i)P(w_i)\} \end{aligned} \quad (1)$$

To estimate the posterior probabilities in Equation 1, let R_i denote the set of regions extracted from the training images in T_i and let r_i^* denote a region vector in R_i that best describes the concept of w_i. Also let α_i^* denote a feature weight vector that indicates the importance of each dimension in the feature space for the concept w_i. The class conditional probability of a region r_j given word w_i can then be calculated as follows,

$$P(r_j|w_i) = \exp\left(-\sum_k \alpha_{ik}^*(r_{jk} - r_{ik}^*)^2\right) \quad (2)$$

where r_{jk} and r_{ik}^* is the k^{th} dimension in the vector r_j and r_i^* respectively, and α_{ik}^* is the feature weight associated to the k^{th} feature. Finally we could select top N keywords with the largest posterior probabilities, i.e., the highest degree of confidence, as the annotation for image I.

One of the key steps in Bayesian classification is to select the most representative image region r_i^* and corresponding feature weight vector α_i^* for keyword w_i. As mentioned in Section 2, there are a large amount of irrelevant noisy regions in R_i, which poses a great challenge on any learning algorithm. In the next Section, we describe how MIL could be used to predict the representative region r^* and the feature weight vector α_i^* from the training data.

3.2 MIL for Region Selection

Multiple-instance learning[4] is a variation of supervised learning, where the task is to learn a concept given positive and negative bags of instances. Each bag may contain many instances, but a bag is labelled positive even if only one of the instances in it falls within the concept. A bag is labelled negative only if all the instances in it are negative.

In region based image annotation, each region is an *instance*, and the set of regions that comes from the same image can be treated as a *bag*. We annotate an image by

keyword w_i if at least one region in the image has the semantic meaning of w_i. For example, the first image in Figure 1 is annotated with keyword "tiger" and segmented to 10 regions. These 10 regions consist a positive bag for "tiger". In this positive bag, there are only 2 positive instances because only 2 regions are actually relevant to "tiger". Given an image labelled by keyword w_i, we can expect that at least one region will correspond to w_i even if segmentation may not be perfect. Hence, the image annotation problem is in essence identical to MIL setting.

One way to solve MIL problem is to examine the distribution of these instances, and look for an instance that is close to all instances in the positive bags and far from those from negative bags. In other words, we should search for a point where there is a high Diverse Density (DD) of positive instances.

The problem of finding the global maximum DD is difficult, especially when the number of dimension is large. It is usually solved by using a gradient ascent method with multiple starting points. Alternatively, Point-wise Diverse Density (PWDD) algorithm [4] attempts to find an instance from each positive bag which is likely to be a "true" positive instance, i.e., the instance with largest DD. As to image annotation, *PWDD is particularly useful because it can return the most representative region for a keyword, which make it possible to explicitly observe which region corresponds to which keyword and evaluate the learning performance visually.* Other existing annotation methods, such as Machine Translation model [6], use EM algorithm to find the correspondence of regions and keywords. However, EM algorithm cannot obtain an explicit mapping between regions and words. The learned correspondence can only be evaluated by the annotation results, which are affected by many other factors.

Finding the optimum feature vector by hill-climbing for all the instances in the positive bags is a very time-consuming process. In this paper, we employ a sequential search algorithm to find t and α with highest DD instead of taking an iterative optimization approach. The underlying reason is that we want an algorithm that scales well for large image databases. The sequence of steps in Figure 2 illustrates our algorithm in detail.

for each word w_i
1. Collect the training images with w_i as the positive bags B^+, remaining images as negative bags B^-.
2. With equal feature weight, find instance B_j with highest Diverse Density.
3. Find the scaling vector α_i^* that maximizes the DD of B_j by gradient search.
4. Use this scaling α_i^* to find the instance with highest DD and output as r_i^*

end for

Figure 2: Representative region and feature vector selection by a sequential PWDD algorithm.

4. EXPERIMENT AND RESULTS

The data set used in this paper is same as the data set used in the experiment of [6]. There are $5,000$ images from 50 Corel Photo CDs in this data set. We use $4,500$ images as training set and the remaining 500 images as testing set. Images are segmented using Normalized cut [3]. Only regions larger than a threshold are used, each image is typically represented by $5-10$ regions sorted by region size. Each region is represented as a 30 dimensional vector, including region color, region average orientation energy, region size and location and so on. The vocabulary contains 371 different words.

The proposed sequential PWDD algorithm can learn a representative region for each keyword, which makes it possible to explicitly show the correspondence between image regions and words and evaluate the learning outcome. Because the training data does not contain content level annotation, it is hard to check the correspondence canonically for a large image database; instead, each region must be viewed by hand to tell whether the representative region is correct for a given keyword. Inevitably, this evaluation is somewhat subjective.

We visually evaluate the selected regions for all 371 keywords and show 21 example words and their corresponding representative regions in Table 1. Some major findings are summarized as follows,

1. Regions associated with high frequency keywords are more likely to be correctly predicted than those of rarely occurring words. The underlying reason is that some concepts are too complicated to be described by just a few samples. The performance of PWDD algorithm is reduced by insufficient relevant instances for these concepts. With the existence of a large number of negative regions, the algorithm tries to find the "farthest" point to all negative regions, which may be irrelevant to the target concept. For instance, the representative regions of low frequency keywords, such as "glass", "sheep", "gate" and "fish", are no better than random guesses.
2. The words with similar semantic meaning usually share the same representative region. For example, the selected region is same for words "horses" and "foals". This is because those words share the similar set of relevant and irrelevant images. Consequently, the PWDD algorithm may return the same region as the maximum Diverse Density instance.
3. The concepts with prominent visual features, such as "tiger" and "sky", are usually well learned by MIL, while some complicated concepts with diverse visual characters such as "house" and "people" are hard to be predicted.

After the selection of the representative region for each keyword by PWDD, the posterior probability of each candidate word given a testing image is computed. Top five keywords are selected based on the degree of confidence as the final annotation for that image.

To evaluate the performance of the proposed annotation algorithm, we compare the automatic annotation with the previously removed manual annotation for each testing image.

The entire vocabulary contains 371 words, and most of them do not appear frequently in the training data set. To demonstrate the effect of region selection on image annotation, we select 70 most frequently used keywords and conduct our first experiment. The average recall and precision of a set of keywords with correct representative regions (by

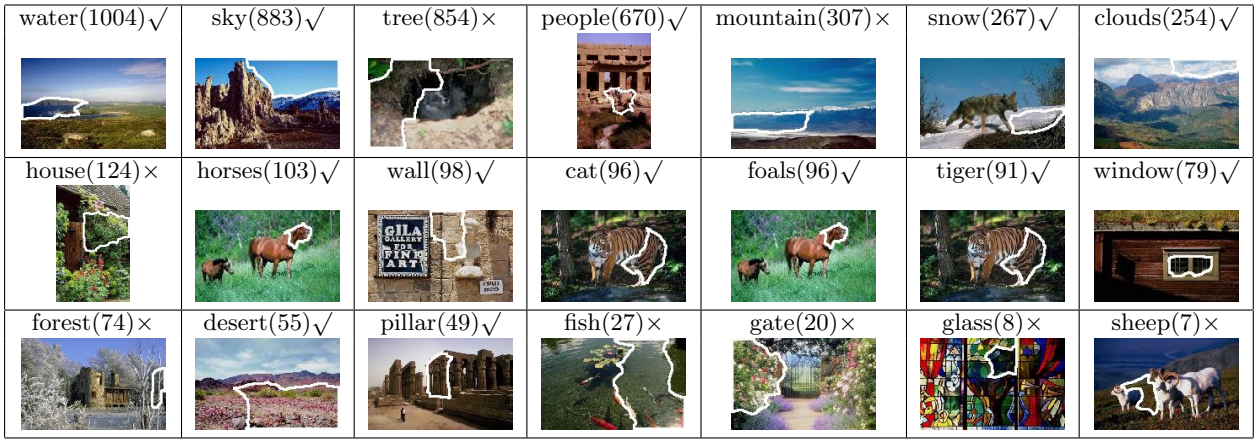

Table 1: Representative regions of 28 selected keywords sorted by the occurring frequency. The number in the bracket is the appearing times of the word in the training set. A "√" indicates correct prediction and a "X" denotes an incorrect one.

human judgement) are 0.54 and 0.49, respectively. While the average recall and precision of a set of keywords with wrong representative regions or with diverse visual characters (again, by human judgement) are 0.16 and 0.09, respectively.

It is clear that the annotation performance is strongly related to the correctness of region selection. The keywords with correct representative regions have much higher recall and precision than those with wrong representative regions. The results demonstrate that learning the correspondence between the regions and the words is crucial to the success of image annotation.

We perform our second experiment with all 371 words to make a direct comparison with the results obtained by Machine Translation model [6], in which EM algorithm is employed to find an implicit correspondence of regions and keywords. The results are summarized in Table 2. The average per-word recall and precision are reported for all 371 words and the best 49 keywords as [6] did. To evaluate the effect of feature selection in image annotation, we also implement a simplified version of PWDD algorithm with equal feature weights. The annotation results are also reported in the same Table. Table 2 clearly shows that significant improvements in annotation performance have been achieved by MIL model over Machine Translation model. In addition, the embedded feature selection in PWDD is helpful in improving the recall and precision of image annotation.

Approaches	All 371 Words		Best 49 Words	
	Avg. pr.	Avg. re.	Avg. pr.	Avg. re.
M.T.	0.04	0.06	0.20	0.34
PWDD	0.07	0.09	0.31	0.46
PD. wo. FS.	0.05	0.09	0.25	0.44

Table 2: Performance comparison of Machine Translation Model (M. T.), Point-wise Diverse Density Model (PWDD) and Point-wise Diverse Density Model without feature selection (PD. wo. FS.) on the task of automatic image annotation.

5. CONCLUSIONS

In this paper, we propose to learn the correspondence between image regions and keywords through the proposed sequential PWDD MIL algorithm. After a representative image region has been learned for a given keyword, we address image annotation as a classification problem under the Bayesian framework. We thoroughly analyzed the region selection results obtained from an image database with 5,000 images. Our finding suggests that the accuracy of representative region selection is crucial to the success of image annotation. Our results also show that the proposed approach outperforms other annotation models [6] in terms of annotation precision and recall.

6. REFERENCES

[1] K. Barnad, P. Duygulu, N. Fretias, D. Forsyth, D. Blei, and M. I. Jordan., "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[2] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *Proc. of ACM MM*, 2004, pp. 540–547.

[3] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," in *Proc. of IEEE CVPR 97*, Puerto Rico, 1997.

[4] O. Maron and T. Lozano-Perez, "A framework for multiple-instance learning," in *Proc. of NIPS*, 1997, vol. 10, pp. 570–576.

[5] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, pp. 913 –939, 2004.

[6] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation :learning a lexicon for a fixed image vocabulary," in *Proc. of ECCV*, 2002, vol. 4, pp. 97–112.

[7] W. Jin, R. Shi, and T. Chua, "A semi-naive bayesian method incorporating clustering with pair-wise constraints for auto image annotation," in *Proc. of ACM MM*, 2004, New York, NY, pp. 336–339.

[8] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. of MISRM*, 1999.

[9] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. of ACM SIGIR*, Toronto, Canada, 2003.

[10] C. Yang, M. Dong, and F. Fotouhi, "I^2A: an Interactive Image Annotation system," in *Proc. of IEEE ICME*, Amsterdam, the Netherlands, July 2005.

[11] C. Yang, M. Dong, and F. Fotouhi, "Image content annotation using bayesian framework and complement components analysis," in *Proc. of IEEE ICIP*, Genova, Italy, Sept 2005.