Enhancing Image Annotation by Integrating Concept Ontology and Text-based Bayesian Learning Model

Rui Shi[1], Chin-Hui Lee[2] and Tat-Seng Chua[1]
[1] School of Computing, National University of Singapore, Singapore 117543
[2] School of ECE, Georgia Institute of Technology, Atlanta, GA 30332, USA
{shirui,chuats}@comp.nus.edu.sg, chl@ece.gatech.edu

ABSTRACT
Automatic image annotation (AIA) has been a hot research topic in recent years since it can be used to support concept-based image retrieval. However, most existing AIA models depend heavily on the availability of a large number of labeled training samples, which require significant human labeling efforts. In this paper, we propose a novel learning framework which integrates text-based Bayesian model (TBM) and concept ontology to effectively expand the training set of each concept class without the need of additional human labeling efforts or collecting additional training images from other data sources. The basic idea lies in exploiting the text information from training set to provide additional effective annotations for training images so that training data for each concept class can be augmented. In this study we employ Bayesian Hierarchical Multinomial Mixture Models (BHMMMs) as our baseline AIA model. By combining additional annotations obtained from TBM into each concept class in the training phase, the performance of BHMMMs can be significantly improved on Corel image dataset with 263 testing concepts as compared to the state-of-the-art AIA models under the same experimental configurations.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]:– *Retrieval Models; Search Process;* I.5.4 [**Applications**]:– *Computer Vision*

General Terms: Algorithms, Human Factors, Performance

Keywords: Automatic Image Annotation, Mixture Model, MLE, MAP

1. INTRODUCTION
With the increasing usage of Internet and digital image processing techniques, more and more images are now available on the World Wide Web. Effective tools to automatically index images are thus essential to support new applications in image retrieval. To facilitate concept-based image retrieval, annotating images with semantic concepts has become an intensive research topic in recent years. Traditionally, automatic image annotation (AIA) refers to the process of automatically labeling images with a predefined set of words representing image semantics. The annotated images can then be retrieved using concept-based search. In this paper, we loosely use the terms *annotation*, *concept* and *word* interchangeably to denote text annotations of images.

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

Most approaches to AIA can be divided into two categories. The approaches in the first category focus on finding joint probabilities of images and concepts. Cross-media relevance model [7], continuous relevance model [8] and multiple Bernoullli relevance model [5] are examples in this category. The AIA approaches in the second category [2, 3, 6, 11, 12], are formulated as a multi-class classification problem, and mixture model [2, 3, 11, 12] has been demonstrated as an effective way to cover large variations in images by simply increasing the number of mixture components with the same probabilistic formulation.

However, most existing AIA approaches depend heavily on the availability of a large number of training samples. One approach to tackle this problem is to manually annotate more unlabeled images. But labeling images manually is a tedious labor effort that is not scalable to a large amount of images. To reduce human labeling efforts, some approaches [13, 14] employ active learning scheme to select a small number of unlabeled images for users to label, and then use the labeled images as additional training samples. In [4], a bootstrapping framework is proposed to annotate images on the web, and combine the labeled web images into training set to re-estimate the model parameters. In essence, the approach in [4] employs an iterative framework to collect more training samples from other data sources. But as different human annotators have their own semantic understandings for similar images, there often exist inconsistent semantics among different data sources. Thus resorting to external data sources does not always work well.


(a) boats. canal. dock. sky.	(b) boats. water. buildings.

Figure 1. Two images with incomplete annotations

Based on the above analysis, these existing approaches neglect the available text information from training images and ignore the fact that most image collections often come with few and incomplete annotations. Figure 1 shows the original annotations of two images coming from the Corel image corpus. Obviously the possible missing annotations for the image in Fig. (1a) could be 'water' and 'buildings', and those for the image in Fig. (1b) could be 'dock' and 'sky'. Thus an easy way to acquire more training samples is to expand the original image annotations, which effectively results in more training images for each concept class within the same overall training set.

In this paper, we study the problem of how to expand the original image annotations effectively by utilizing text

information from the training set. Due to the few or incomplete original annotations, there is one key issue related to our problem, namely, the accurate parameter estimation especially when the number of training samples is small. To tackle this issue, we incorporate prior knowledge into the concept ontology, and propose a text-based Bayesian learning model called TBM to characterize the ontology structure and estimate the parameters of the concept mixture models. This facilitates a statistical combination of the likelihood function of the training data and the prior density of the concept parameters into a well-defined posterior density whose parameters can now be estimated via a maximum a posteriori (MAP) criterion.

By combining additional annotations obtained from TBM into each concept class in the training phase, experimental results on the Corel image dataset with 263 testing concepts show that we can significantly improve the performance of our baseline BHMMM model both in terms of mean precision and recall measures. This indicates that most additional annotations are positive reinforcements for the concept classes.

The main contributions of this paper are summarized as follows:

- We propose a novel learning framework to tackle the problem of effectively increasing the training set for concept classes by utilizing text information.
- We propose a text-based Bayesian model (TBM) to characterize the ontology structure and estimate the parameters of concept mixture models accurately.

The rest of the paper is organized as follows. In Section 2, we give a brief introduction to our proposed framework. In Section 3 we discuss key issues in TBM. Through a series of experiments, we analyze the effectiveness of the proposed leaning framework and TBM in Section 4. Finally Section 5 concludes our findings.

2. System Framework

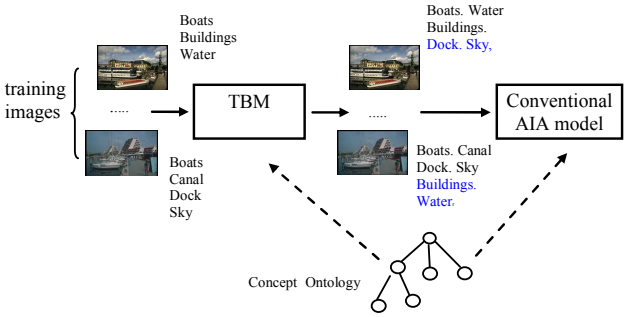


Figure 2. The proposed framework

As shown in Figure 2, we propose a novel framework to expand the image annotations and acquire image samples for concept classes. Given the annotated images in training phase, TBM is first used to expand the original image annotations, and then the images with expanded annotations are taken as the new set of training samples for the conventional AIA model. Here the conventional AIA refers to associating visual features to text annotations. Obviously our proposed framework is general, and other models can be also used to expand the annotations or perform the conventional AIA.

In this paper we employ Bayesian Hierarchical Multinomial Mixture Models (BHMMMs) [12] as our baseline to perform conventional AIA, since BHMMM is one of the state-of-the-art conventional AIA models. A concept ontology derived from WordNet [1, 9, 12] is used to model the concept relationships and estimate the hyperparameters of BHMMMs and TBMs. We will explain the hyperparameters in section 3. The pipeline of our proposed framework is as follows:

1) Given training set of images, estimate the parameters of TBMs by MAP criterion.
2) For any concept c_i, expand the annotations of images related to c_i by TBMs.
3) Generate a rank list of images for c_i based on their likelihoods.
4) Expand training set of c_i by combining top N images.
5) Estimate the parameters of BHMMM for c_i by combining the additional and original training set.

In step 2), since the size of the whole training corpus is often very large, given a concept, we don't perform TBM in the whole training set to expand the annotations of all the training images. Instead, TBM is performed only on the set of images which are related to the given concept. The set of related concepts includes three parts: i) the closest hypernym concept; ii) the co-occurred concepts in the training corpus; and iii) the co-occurred concepts from its sibling concepts.

3. TEXT-BASED BAYESIAN MODEL

In the training set a given image I has been labeled by some text annotations, and can be represented by a concept vector, $I = (m_1, m_2, \ldots, m_Q)$, where Q is the total number of the predefined concepts, and m_q ($1 \le q \le Q$) denotes the observed count of the q^{th} concept in image I. We use D_i ($I_t \in D_i$) to denote a collection of independent training images for concept class c_i.

Considering text annotations, each labeled training image is a text document represented by a concept vector. So we formulate the task of expanding annotations as a multi-class text classification problem. As pointed out in [10] for multi-topic text classification, these text documents often form mixtures of multiple topics, which makes mixture models suitable for representing complex class-conditional probability density function for classification, and maximum likelihood estimation (MLE) is used to estimate the model parameters. Given a total of K text mixture components, the observed vector I from concept class c_i is assumed to have the following probability:

$$p(I_c \mid \eta_i) = \sum_{k=1}^{K} \beta_{i,k}\, p(I_c \mid \chi_{i,k}) \qquad (1)$$

where $\eta_i = \{\beta_{i,1},\ldots,\beta_{i,K},\chi_{i,1},\ldots,\chi_{i,K}\}$ is the parameter set for the text mixture model, including mixture weight set $\{\beta_{i,j}\}_{k=1}^{K}$ ($\sum_{j=1}^{K}\beta_{i,j}=1$), mixture parameter set $\Gamma_i = \{\chi_{i,k}\}_{k=1}^{K}$, and $p(I_c \mid \chi_{i,k})$ is the k^{th} mixture component to characterize the class distribution. We use multinomial distribution to model each mixture component as in [10]. Here each parameter $\chi_{i,j,q}$ in $\chi_{i,j}$ can be interpreted as the average distribution of q^{th} concept for images belonging to j^{th} mixture component of the i^{th} concept class. Here we call the Eq.(1) as text mixture model (TMM).

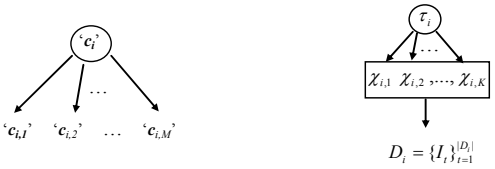
However, the major shortcoming of TMM is that there are usually too many parameters to be estimated but not enough training images due to few and incomplete annotations. Furthermore, as pointed out in [15], smoothing of the maximum likelihood estimation is extremely important when the number of training samples is small. One way to enhance the ML estimates for Eq.(1) is to incorporate prior knowledge into modeling by assuming the mixture parameters in $\chi_{i,k}$ as random variables with a joint prior density $p_0(\chi_{i,k} \mid \tau_i)$ with parameters τ_i (often referred to as *hyperparameters*). Thus the posterior probability of observing the training set can be evaluated as:

$$p(\eta_i \mid D_i) \propto \{\prod_{t=1}^{|D_i|}\sum_{k=1}^{K}[\beta_{i,k}p(I_t \mid \chi_{i,k})]\} * p_0(\Gamma_i \mid \tau_i) \quad (2)$$

In contrast to conventional ML estimation, we can impose a maximum a posterior (MAP) criterion to estimate the parameters as follows:

$$\overline{\eta}_i^{map} = \arg\max_{\eta_i} \log\{\prod_{t=1}^{|D_i|}\sum_{k=1}^{K}[\beta_{i,k}p(I_t \mid \chi_{i,k})]\} * p_0(\Gamma_i \mid \tau_i) \quad (3)$$

In [8, 15], the hyperparameter can be interpreted as an extra count or prior observation count to smooth the average distribution of every word in concept classes, and is simply assumed to be $\tau_i = (\mu p(c_1), \mu p(c_2), ..., \mu p(c_Q))$. Here μ is an empirical constant, and $p(c_q)$ is the relative frequency of observing the word c_q in the whole training set. However, such a choice for hyperparameter ignores the concept dependency. For example, if we want to estimate the hyperparameters for the 'tiger' class, then 'prior observation counts' of 'street', 'buildings' should be lower. But if we want to estimate the hyperparameter of 'city', then 'prior observation counts' of 'street', 'buildings' should be higher.



(a) Sub-tree of multi-level concept ontology (b) TBM

Figure 3. An illustration of the proposed TBM

To better model the concept dependencies, we derive concept ontology through WordNet as in [1, 12]. Thus we propose a text-based Bayesian learning model (TBM) to characterize the concept ontology structure. The basic idea behind the proposed TBM is that the mixtures from the most dependent concepts share the same set of hyperparameters and these concept mixture models are constrained by a common prior density parameterized by this set of hyperparameters. This is reasonable since given a concept, say, 'oahu', the image annotations from its sibling concepts (say, 'kauai' and 'maui') are often related and can be used as prior knowledge as well. Thus the hyperparameters can be interpreted as the shared prior knowledge among the dependent concepts. Fig. (3a) shows a sub-tree of a multi-level concept ontology in which the concepts ($c_{i,1}$, $c_{i,2}$, … , $c_{i,M}$) are derived from their parent node, labeled 'c_i'. As shown in Fig. (3b), the mixtures from concepts $c_{i,1}$, $c_{i,2}$, … , $c_{i,M}$ share the same set of hyperparameters, τ_i.

Thus based on the Eq. (2) and (3), TBM needs to address three key issues, namely i) the definition of the prior density, p_0 ; ii) the specification of the hyperparameters based on concept ontology, τ_i ; and iii) MAP estimation of the mixture model parameters, $\overline{\eta}_i^{map}$. We employ the approaches similar to that used in [12] to address these key issues, and thus we do not repeat the technical details in this paper.

4. TESTING SETUP & EXPERIMENTAL RESULTS

Following [7, 12], we conduct our experiments on the same Corel CD data set, consisting of 4500 images for training, and 500 images for testing. In this corpus, there are 371 concepts in the training set but only 263 such concepts appear in the testing set, with each image assigned 1-5 concepts. Based on the same concept ontology as in [12], we have a total of 513 concepts (Q=513). If a non-leaf concept node in the concept ontology doesn't belong to the concept set in Corel CD corpus, then its training set will consist of all the images from its child nodes. As with the previous studies, the performance is evaluated by comparing the set of generated annotations with the actual annotations specified in the testing set. We assign a set of five top concepts to each test image based on their likelihoods from the conventional AIA model.

4.1 Performance evaluation on whole training set

In terms of the results in [12], without expanding the annotations by TBM and TMM, BHMMMs with J=25 (where J is the number of the mixtures) achieved the best performances of conventional AIA in terms of averaging precision, recall as compared to one of the state-of-the-art AIA models, HC [11]. Thus we take BHMMMs with J=25 as our baseline.

First, we want to verify the effectiveness of our proposed framework and TBM. We used BHMMMs (J=25) as the conventional AIA. In the step 4 of our pipeline in Section 2, we picked the top N=5% (the mean number of the increased training samples for each testing concept class is about 7 images) or top N=10% (the mean number of the increased training samples for each testing concept class is about 15 images) of the additional samples.

Table 1. Performances of TMM and TBM

Models	HC [11]	BHMMM [12]	TMM (top5%)	TMM (top10%)	TBM (top5%)	TBM (top10%)
# concepts (recall>0)	93	122	134	143	152	153
Mean Per-concept metrics on all 263 concepts on the Corel dataset						
mPrecision	0.100	0.142	0.143	0.145	0.152	0.156
mRecall	0.176	0.225	0.278	0.301	0.330	0.341

We tabulate the performance of TBM, TMM, HC [11] and BHMMMs (J=25) [12] in Table 1. We derive the following observations from Table 1. (a) The use of additional training examples derived from TMM and TBM is beneficial, since all the performances of TMM- and TBM models are better than those of BHMMM and HC. This also indicates that text information is important and effective to expand the original annotations. (b) As compared with TMM (top 5% and 10%), TBM achieved even better performance in both mean precision and recall measures. In particular, TBM (top 10%) achieves the best performance. Some examples of the additional annotations

or training samples derived from TBM are shown in the Appendix.

4.2 Performance evaluation with small set of samples

This section analyzes the effect of our proposed framework with TBM when the number of original training images is small. We selected a subset of 132 testing concepts in Section 4.1 in which the number of training examples in each class is less than 21.

Table 2. Performance summary on the concept classes with small number of training examples

Models	BHMMM (J=25)	TMM (top10%)	TBM (top10%)
# of concepts (recall>0)	25	50	57
Mean Per-concept metrics on all 132 concepts on the Corel dataset (# of original training samples <=21)			
mPrecision	0.059	0.071	0.090
mRecall	0.106	0.264	0.333

Table 2 compares three models, BHMMMs (J=25), TMM (top10%) and TBM (top10%). Obviously TBM achieves the best performances. This indicates again that our proposed framework and TBM are effective in acquiring more training samples when the number of training samples is small.

5. CONCLUSION

In this paper, we proposed a novel framework by integrating concept ontology and text-based Bayesian learning model to tackle the problem of effectively expand the training set for each concept class without the need of additional human labeling efforts or collecting additional training images from other data sources. Experimental results on the Corel image dataset show that our proposed framework and TBM can effectively expand the original annotations. In future work, we will study how to utilize the text and visual information from training set to expand the annotations, and leverage on the vast amount of information on the Web to improve AIA performance.

6. REFERENCES

[1] K. Barnard, P. Duygulu and D. Forsyth, "Clustering Art", In *Proc. Of IEEE Computer Vision and Pattern Recognition*, 2001.

[2] G. Carneiro and N. Vasconcelos, "Formulating Semantic Image Annotation as a Supervised Learning Problem", *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[3] J. P. Fan, H. Z. Luo and Y. L. Gao, "Learning the Semantics of Images by Using Unlabeled Samples", *Proceedings CVPR*, 2005.

[4] H.M. Feng. R. Shi and T.S. Chua, "A Bootstrapping Framework for Annotating and Retrieving WWW Images", In *ACM Multimedia'04*, pp. 960-967, New York, 2004.

[5] S.L. Feng, R. Manmatha and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation", *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04.

[6] S. Gao, D.-H. Wang and C.-H. Lee, "Automatic Image Annotation through Multi-Topic Text Categorization", *Proc. ICASSP*, Toulouse, France, May 2006.

[7] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models", *Proc. of the 26th ACM SIGIR*, 2003.

[8] V. Lavrenko, R. Manmatha and J. Jeon, "A Model for Learning the Semantics of Pictures", *NIPS*, 2003.

[9] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to WordNet: an on-line lexical database", *Intl. Jour. Of Lexicography*, pp. 235-244, 1990.

[10] J. Novovicova and A. Malik, "Application of Multinomial Mixture Model to Text Classification", *Pattern Recognition and Image Analysis*, pp. 646-653, 2003.

[11] M. Srikanth, J. Varner, M. Bowden and D. Moldovan, "Exploiting Ontologies for Automatic Image Annotation", *Proceedings of the 28th ACM SIGIR*, 2005.

[12] R. Shi, T.S. Chua, C.H. Lee and S. Gao, "Bayesian Learning of Hierarchical Multinomial Mixture Models of Concepts for Automatic Image Annotation", In *Proc. of CIVR'06*, pp. 102-112, Arizona, United States, 2006.

[13] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval", In *ACM Multimedia'01*, pp. 107-118, Ottawa, Canada, 2001.

[14] R. Yan, and A.G. Hauptmann, "Multi-class Active Learning for Video Semantic Feature Extraction", In *Proc. of ICME'04*, pp. 69-72, 2004.

[15] C.X. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval", *SIGIR'01*, 2001.

APPENDIX: Examples of top additional training samples obtained from TBM for 6 concept classes, where blue words denote the concept class, and black words denote the original annotations.

jet	vendor	coast
		
plane. prop. runway	**people. market. tree.**	**boats. harbor. sky. water.**
		
formation. sky. plane. prop.	**people. market.**	**beach. sand. sky. water.**
antlers	**athlete**	**railroad**
		
bulls. elk. field. grass.	**people. water. swimmers. pool.**	**locomotive. road. train.**
		
field. moose. tree.	**people. swimmers.**	**locomotive. tracks. train. snow.**