A Learning-based Approach for Annotating Large On-Line Image Collection

HuaMin Feng and Tat-Seng Chua
School of Computing, National University of Singapore
Email: { fenghm, chuats}@comp.nus.edu.sg

Abstract

Several recent works attempt to automatically annotate image collection by exploiting the links between visual information provided by segmented image features and semantic concepts provided by associated text. The main limitation of such approaches, however, is that semantically meaningful segmentation is in general unavailable. This paper proposes a novel statistical learning-based approach to overcome this problem. We employ two different segmentation methods to segment the image into two sets of regions and learn the association between each set of regions with text concepts. Given a new image, the idea is to first employ a greedy strategy to annotate the image with concepts derived from different sets of overlapping and possibly conflicting regions. We then incorporate a decision model to disambiguate the concepts learned using the visual features of the overlapping regions. Experiments on a mid-sized image collection demonstrate that the use of our disambiguation approach could improve the performance of the system by about 12-16% on average in terms of F_1 measures as compared to system that uses only one segmentation method.

1. Introduction

Effective techniques are needed to model and search the content of large digital image/video libraries. One popular technique is query-by-example (QBE), in which users provide visual examples of the contents that they want to seek, and the system retrieves images on the basis of similarity in content features such as the color, texture etc. Such low-level content-based retrieval schemes, however, have the obvious limitation that it is hard to retrieve images based on abstract concepts. Since most users wish to search for images in term of semantic concepts rather than visual contents [1], work on image/video retrieval research has begun to shift from QBE to query-by-keyword (QBK). QBK allows users to search for images by specifying their own query in terms of a limited vocabulary of semantic concepts [2]. The problem with such approach is the need to annotate images with semantic concepts accurately and completely. The traditional approach is to manually annotate images, which is very tedious and time-consuming. Hence, it is desirable to automatically assign semantic concepts (keywords) to images.

Several techniques for automatically assigning keywords [3-5] and semantic search [6, 7] of image databases have been proposed. Mori et. al. [3] proposed an approach to perform "image-to-word transformation based on dividing and vector- quantizing images with words". They assumed that each image in the training set associates with several keywords. They divided the image into fixed-size blocks and each block inherits the whole set of keywords associated with the image. Blocks are then clustered based on vector quantization, and the clusters are used in turn to predict the keywords for new images. The advantage of this approach is that it does not need to perform image segmentation. However, due to the use of fixed size blocks, one object within the image may be divided into several blocks or worse still, a block may cover several objects. Thus the extracted block feature vector is unable to represent the object, and hence the accuracy with this approach tends to be low.

Barnard and Forsyth [4] utilized a statistical model to associate image regions explicitly with words. They used Blobworld [8] to produce the segmented regions within the image. The system works by modeling the statistics of word and region feature occurrence and co-occurrence. The learned region-word probabilities are then used to associate words with regions in new images. The major problem with this approach is that it requires accurate and meaningful segmentation, which is not generally available.

To tackle the segmentation problem, Chang et al. [5] proposed a content-based soft annotation for multimodal image retrieval using bayes point machine (BPM). Each training image is manually assigned a concept term from the lexicon, and the visual content of the whole image is modeled using a color and texture feature vector (144-dimisension). BPM or SVM is then used to train a classifier for each concept to determine the confidence score of assigning the concept to the image. For a new image, the system chooses those concept terms with high confidence scores. However, due to the use of global image features, although this approach is good for general classes of objects such as forest, sky etc., it might not be able to recognize concepts for specific objects very well.

Another approach to overcome the segmentation problem is proposed by Wang and Li [9]. They assigned

Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

a textual description of concepts for an image collection and employed a 2-D multi-resolution HMM to capture the cross blocks and cross resolution dependencies between blocks for the entire image collection. Given a new image, the feature vector of the image is compared with the trained models, and statistically significant terms are extracted to annotate the image. However, because of fixed-size block (4×4), this approach might inherit the same problems as in [3].

In this paper, we propose a novel statistical learning-based approach to overcome the above problem. Instead of using a segmentation method, which has all the segmentation problems, we consider two different segmentation methods to segment the image into two independent sets of regions. We separately learn the association between concepts of regions derived from two different techniques. The idea is to first employ a greedy strategy to annotate the image using the different sets of overlapping and possibly conflicting segmented regions. We then incorporate a decision model to disambiguate the concepts learned using the visual features of the overlapping segmented regions. Experiments on a mid-sized image collection (with about 5,000 images from photoCD and Web) demonstrate that the use of our disambiguation approach could improve the performance of the system by about 12-16% on average in terms of F_1 measures as compared to system that uses only one segmentation method.

The paper is organized as follows. Section 2 presents the overview of our approach. Section 3 describes the process of associating semantic concepts and regions for images. Section 4 presents the annotation of image and disambiguation of concepts using a decision model. Experimental results and discussion are given in Section 5. Finally, the conclusion and our future work are discussed in Section 6.

2. Overview of our Approach

Until now, concept annotation approaches for images rely on only one method. The method is based on either fixed-size block segmentation, region segmentation or whole image. As is well known, image segmentation is a very difficult and challenging task because of the flexibility in light intensity, and structure and uncertainty in regions. Therefore, the main limitation of region segmentation approaches is the quality of image segmentation, which tends to be unstable and unreliable. Fixed-size block method also has similar problem because the fixed-size block segmentation tends to be arbitrary.

Due to human perception differences, the same image may be perceived as different set of objects by different users. Similarly, different image segmentation methods may produce different regions with most not representing true regions. Therefore any learning approach such as GMM, HMM, etc., that learns from just one segmentation method will produce biased results.

Thus in this work, we propose a novel approach to learn semantic concepts from multiple and overlapping regions and use a decision model to arbitrate among different learned concepts. To realize this idea, we employ two image segmentation methods, one is Blobworld from Berkeley [8], the other is JSEG from UCSB [10]. These two methods are based on different techniques and underlying assumptions. Blobworld segments image into regions by fitting a mixture of Gaussians to the pixel distribution in a joint color-texture-position feature space. On the hand, JSEG performs color quantization follow by spatial segmentation, and uses a region growing method to extract the final set of segments. As a result of their differences, the segments produced by the two methods are often different and are conflict with each other. Thus when learning concepts based on the conflicting regions, it may produce ambiguous concepts. This, however, gives us additional evidence to arbitrate the conflicting results. We can employ a decision making model to learn from the context (conflicting regions) to disambiguate the concepts. Our approach is to some extend inspired by the framework for Chinese named entity extraction[11] in natural language processing where multiple methods are used to extract Chinese named entities, and a decision tree is employed to disambiguate the conflicting named entities.

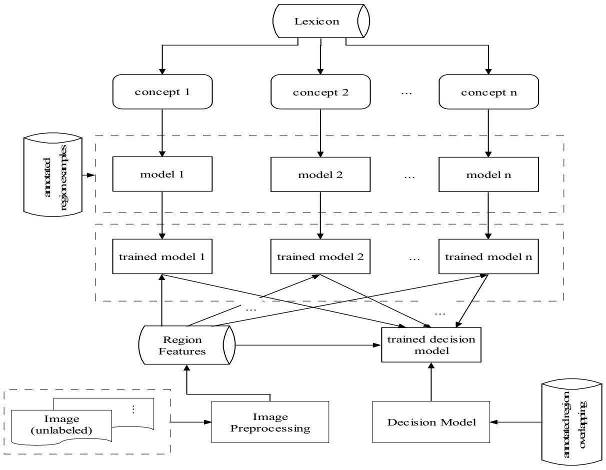


Figure 1 the Overview of our learning-based approach

Figure 1 gives the overview of our approach. The approach consists of three function blocks, i.e., image preprocessing, model training/testing, and decision model for image annotation. Image preprocessing module is responsible for the segmentation and feature extraction of regions. Model training deals with the training of concept classifiers to assign concept(s) to image region. Decision model aims to arbitrate among the ambiguous concepts to derive at the final annotation for new images.

The details of image preprocessing module are briefly outlined in Section 2.1. The rest of the modules are presented in Sections 3-4.

2.1 Image Preprocessing

Figure 2 presents the main function blocks for image preprocessing. First, we use two or more segmentation methods to segment each the image into multiple set of regions, and store the results separately for each method. Due to the uncertainty of segmented regions, the regions obtained from the two methods may conflict with each other. Second, in order to derive the relationship between regions obtained from different methods, we compute the overlap of a region produced by one method with all the regions obtained by the other method. Third, we extract region features for each region set and setup the region correlation between the two sets of regions. We store all the features of each region and corresponding regions in a conflicting(overlapping) correlation matrix, M_c, for latter use.

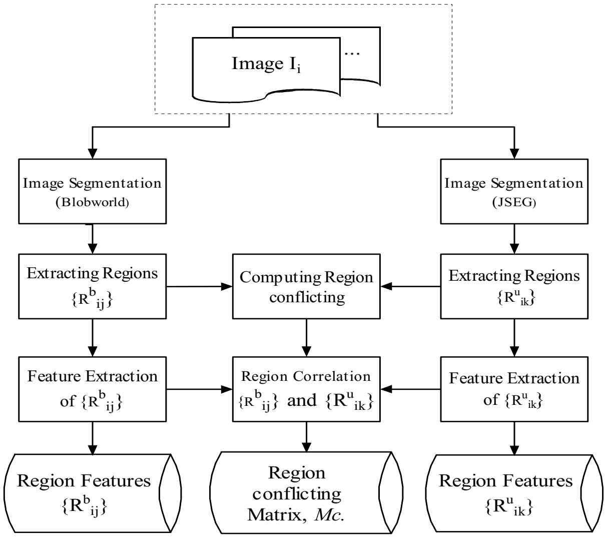


Image I_i
...
Image Segmentation (Blobworld)
Image Segmentation (JSEG)
Extracting Regions $\{R^b_{ij}\}$
Computing Region conflicting
Extracting Regions $\{R^u_{ik}\}$
Feature Extraction of $\{R^b_{ij}\}$
Region Correlation $\{R^b_{ij}\}$ and $\{R^u_{ik}\}$
Feature Extraction of $\{R^u_{ik}\}$
Region Features $\{R^b_{ij}\}$
Region conflicting Matrix, *Mc.*
Region Features $\{R^u_{ik}\}$

Figure 2 the function flowchart of image preprocessing

3. Learning Semantic Concepts for Image Collection

Given a set of training images together with the segmentation of these images using two different segmentation methods, we have

$$\text{Image} I_i \xrightarrow{\text{segmentation}} R^b_{I_i} = \{r^b_{ij}\}_{j=1..N_b} \;\&\&\; R^u_{I_i} = \{r^u_{ik}\}_{k=1..N_u} \qquad (1)$$

where $\{r^b_{ij}\}$, $\{r^u_{ik}\}$ denote the set of segmented regions generated by using the Blobworld [8] and UCSB JSEG[10] methods, respectively.

To train the association between regions and concepts, and vice versa, we need to solve two problems. First, we need to associate a region with a set of concept terms. Second, we need to train a classifier to learn the association between region and concepts so that given a new region, the classifier is able to assign an appropriate concept.

To tackle the first problem, we need to come up with a list of concepts and assign the concepts to known regions in the training images. According to human perception and knowledge, semantic concepts can be divided into two types. One is the atomic concept, which corresponds to the basic object that can not be further subdivided. The atomic concept tends to correspond to a homogeneous region. Another type of concept is the complex concepts, which can be abstract concepts or any other composite concepts made up of several atomic concepts. An example of the composite concept is the plane-taking-off, which can be inferred from atomic concepts such as the plane, sky, audio sound of taking-off etc. Lexicon used to denote concepts can be organized in a hierarchical structure based on the WorldNet [12] or other schemes. For example, MPEG-7 provides a number of classification scheme (CS), such as the Genre CS, which provides a hierarchy of genre categories for classifying multimedia content[13]. Alternately, more extensive classification systems can be used such as the Library of Congress Thesaurus of Graphical material (TGM)[14], which provides a set of categories for cataloging photographs and other types of graphical documents.

For this work, we focus only on atomic concepts organized as a set. Other form of complex concepts and organization scheme will be investigated later. An example of concepts is given in Table 1. The set of concepts is collected in a lexicon, *Lc*. We next manually annotate the two set of regions in the training images separately. Each region is assigned only one atomic concept from the lexicon. This is reasonable as most regions generated are homogeneous based on either color and/or texture, and tend to cover one or part of an atomic concept.

Given the set of regions and its annotated concepts, the next problem is to train the classifiers to associate the regions with the concepts. i.e., given the region features, we want to predict the text concepts to be used to annotate the region. One popular technique that can be used to accomplish this task is the Support vector machine (SVM). SVM in general requires fewer parameters and assumptions and its discriminant result is in general satisfactory if appropriate features are selected.

For effective training, however, we need to derive a good representation of visual contents of the regions. There are many features that can be extracted from regions, including color histogram (and its transformation), structure information (such as the edge direction histogram), shape structure, etc. In our research, we compared many combinations and found that the combination of the following 9 features gives the best results. The feature set is: normalized region area, contrast, anisotropy, normalized boundary ratio, normalized mass center of region, RGB moments, La*b* moments, Color histogram, Shape structure. These 9 features capture both the shapes and contents of the regions.

In summary, we denote the annotation of regions of the training image M as:

$$\text{ImageI}_{\text{i}} \rightarrow \begin{matrix} [r_{ij}^{b} = \{C_{ij}^{b}, f_{ij}^{b}\}]_{j=1,..,N_b},\ r_{ij}^{b} \in R_{I_i}^{b}\ \& C_{ij}^{b} \in L_C \\ [r_{ik}^{u} = \{C_{ik}^{u}, f_{ik}^{u}\}]_{k=1,..,N_u}, r_{ik}^{u} \in R_{I_i}^{u}\ \& C_{ik}^{u} \in L_C \end{matrix} \qquad (2)$$

where f_{ij}^{b} and f_{ik}^{u} denote the feature set derived for regions r_{ij}^{b} and r_{ik}^{u}, respectively.

We are now ready to train the classifier using SVM. The purpose of the classifier is to annotate a region r_n with the concept, where $r_n \in \{r_{ij}^{b}, r_{ik}^{u}\}$. That is, we want to train two separate SVM models as follows:

$$\begin{matrix} [r_n^{b} = \{C_n^{b}, f_n^{b}\}]^{b} \rightarrow SVM^{\,b} \\ [r_n^{u} = \{C_n^{u}, f_n^{u}\}]^{u} \rightarrow SVM^{\,u} \end{matrix} \qquad (3)$$

We will consider the use of normal SVM, which assigns one concept to a region, and probabilistic SVM, which can be tuned to return multiple concepts per region. Here normal SVM refers to the traditional hard margin SVM, probabilistic SVM refers to the soft margin SVM, which we denote it as soft decision binary model.

3.1 Normal SVM Model

Here we adopt the traditional binary SVM [15], which will assign a single concept to a region with a binary confident value of 0 and 1. As is well known, the SVM model's performance is to some extent dependent on the choice of kernel function and its parameters. With our model, we use radial basis function (RBF) as the kernel function for SVM model. Thus the parameters such as cost constant (Cc) and gamma need to be selected carefully. Based on the training set, we selected different Cc's and gamma parameters for different concept models. Further details of the selection of SVM model parameters can be found in [5].

3.2 Soft Decision Binary SVM Model

The soft decision binary SVM is, in fact, a probabilistic SVM model. We first get the normal SVM model and then maps SVM decision boundary to probabilistic space via the first order sigma function. We also adopt the radial basis function (RBF) as the kernel and select an appropriate set of parameters for the cost constant (Cc) and gamma parameter depending on the concept model.

We formally define the soft binary SVM model of concepts as follows:

<u>Step 1</u>: Given set of annotated regions: $r_n=\{C_n, f_n\}$, *n=1,…, total number of regions, where* $r_n \in \{r_{ij}^{b}, r_{ik}^{u}\}$. We train the SVM models, SVM$_{\text{i}}^{\text{b}}$ and SVM$_{\text{i}}^{\text{u}}$, respectively for each concept c$_{\text{i}}$ for the two set of regions generated by Blobworld and JSEG segmentation methods.

<u>Step 2</u>: Given the trained SVM models, we derive the optimal set of parameters α and β for the sigma function, which maps the decision value{-1,1} to the probabilistic space [0,1]. The sigma function is of the form

$$p(x) = 1 \Big/ (1 + e^{\alpha x + \beta}) \qquad (4)$$

where each concept corresponds to one group of α and β.

With soft decision binary SVM model, it is possible to derive the confidence vector of concepts for each region. The confidence vector (CV) has the form:

$$\text{CV=(conf}_1\text{, conf}_2\text{, …, conf}_{\text{Nc}}) \qquad (5)$$

where *Nc* denotes the total number of concepts in lexicon *Lc*, and conf$_{\text{i}}$ $\in [0,1]$ denotes confidence of i$^{\text{th}}$ concept for one region (of r_{ij}^{b} or r_{ik}^{u}).

With the confidence vector, it is convenient to link and control the choice of concepts to region. Due to the unreliable of region segmentation method, a single concept may not be appropriate to describe the contents of a region especially when the region is inappropriately segmented to cover more than one object. Thus, with the use of confidence vector, we can easily choose one or more concepts for a region depending on the strategies we adopt. In this research, we adopt two strategies as follows:

<u>Strategy 1</u>: One region corresponds to only one semantic concept, *C*. That is, we only choose the concept with the highest confidence value as follows:

$$C = \underset{i}{\text{argmax}} \{conf_i \,|\, conf_i \in \text{CV}\} \qquad (6)$$

Strategy 2: Assign one or more semantic concept $\{Cs\}$ per region. We choose those concepts whose confidence values are larger than some predefined threshold τ. Formally we have:

$$\{\mathrm{Cs}\}=\begin{cases}\{concept_i \mid conf_i \geq \tau, conf_i \in \mathrm{CV}\}, \text{if } \exists \mathrm{conf} \geq \tau \\ \{concept_i \mid \underset{i}{\operatorname{argmax}}\{conf_i \mid conf_i \in \mathrm{CV}\}\}, \text{if } \forall \mathrm{conf} < \tau\end{cases} \qquad (7)$$

There are two further cases that should be considered for Strategy 2. One is that when there are many concepts with confidence values larger than τ. In this case, we simply choose the top 4 concepts in terms of confidence values. The other case is when there is no concept with confidence value larger than τ. In this case, we choose the concept with the highest confidence value as in Strategy 1.

4. The Annotation of Images

Given a new image, we use the two segmentation methods to generate two sets of regions. We then used the trained classifiers to associate one or more concepts to each region. As different regions generated by different methods are quite different, it is likely that two overlapping regions are assigned different and conflicting concepts. Fortunately, region concept is not independent of each other. While some concepts may occur simultaneously within the context, others may not. Thus we need to make use of the context to disambiguate the annotated concepts as discussed in Section 2.

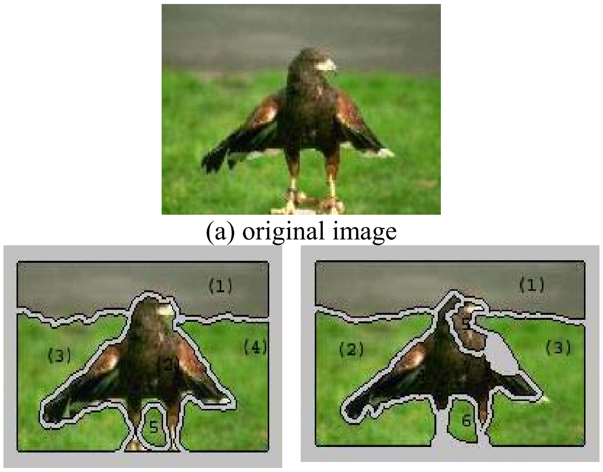
Figure 3 shows one image segmented using the two different segmentation methods. Due to the image segmentation error, the regions found are not strictly semantically based. Some segmented regions may be parts of objects, background or foreground, or may cover multiple objects. The numbers shown in the brackets in Figure 3 gives the sequence number of the regions found in descending order of size. We ignore those regions whose sizes are smaller than a predefined threshold of 1% of the original image size. It is obvious that some regions from the two segmentation methods overlap with each other and we expect these overlapping regions to have higher coherence. Thus when evaluating the semantic concept for the target region, we must consider its context, which can be inferred from the visual attributes and semantic concepts of the overlapping regions.

To take the context of regions into consideration, we must first be able to evaluate the relationship between different overlapping regions, and from these derive appropriate features for use in a decision model to arbitrate ambiguous concepts.

For each image M, we compute the degree of overlap between regions derived from Blobworld, $R_{I_i}^b = \{r_{ij}^b\}$, and from UCSB method, $R_{I_i}^u = \{r_{ik}^u\}$. As discussed in Section 2.1, we compute the overlap between every region $r_{ij}^b \in R_{I_i}^b$ and region $r_{ik}^u \in R_{I_i}^u$. We then normalize the overlap area by the size of image, i.e.,

$$M_{c_{jk}} = U_{r_{ij}^b, r_{ik}^u} = \frac{r_{ij}^b \cap r_{ik}^u}{|\mathrm{Im}age\ \mathrm{I_i}|} \qquad (8)$$

This information is stored in the Region Conflicting Matrix M_c, which encodes the overlaps between all the regions. We choose up to 4 largest overlapping regions around the target region as the context. For each target region, we derive a feature vector comprising the following attributes: normalized target region area, target concept frequency within the image, target region dominant color (5 attributes), and 12 attributes from the four overlapping regions. For each overlapping region, we include the normalized region area, the overlapping region ratio with respect to the target region, and the overlapping region's concept.



(a) original image

(b) region derived from Berkeley Blobworld (5 regions)

(c) region generated by UCSB JSEG (6 regions)

Figure 3: Image segmentation with two different segmentation methods.

We then use the decision tree package SEE5[16] to train a decision tree to evaluate the confidence level of the concepts for the target region. The following pseudo codes compute the confidence of concepts assigned to the target region using SEE5.

Pseudo Codes for assigning the confidence of concepts for target region based on context information:

```
Confidence_level=0.0;
  • For each target region of the image,
```

Step 1: *Find the overlapping regions around the target region from the Region Conflicting Matrix, M_c.*

Step 2: *Get target and overlapping regions' concepts from soft SVM classifiers. The number of concepts selected is dependent on the strategy adapted for soft SVM.*

Strategy 1: get one concept according to Equation (6);

Strategy 2: get one or more concepts according to Equation (7).

- *Construct the feature vector for the target region for the decision model;*
- *Employ the trained decision model to derive the confidence values of concepts for the target region.*
- *Return the confidence values for all related concepts for the target region.*

From the results of above decision model, we select the concepts with the highest confidence value above a threshold as the concept for the target region. We then union the concepts for all the regions from the entire image. The final result of union is the semantic concepts assigned to the image.

5. Experimental Results and discussion

In this section, we discuss the experiments and retrieval results using our framework.

5.1 The Image Datasets Used in the Experiments

We selected about 5,000 images from PhotoCD and the Web. For PhotoCD image collection, we choose all images except texture images. For images derived from the Web, we include scenic images, natural images, landscape etc. We preprocess all the images before learning, including changing the large size images into standard size of 192 by 128 (or 128 by 192) resolution depending on the original image size to save computation costs. We then employ the Blobworld and UCSB JSEG segmentation methods to generate two independent sets of regions for each image.

As described in Section 3, we focus only on the use of atomic concepts to annotate the region. For this research, we selected a list of about 20 atomic concepts as our lexicon as listed in Table 1. These concepts cover all major concepts found in our image collections. We randomly select about 400 images for training and use the rest for testing. For training images, we manually assign a concept from the lexicon, *Lc,* for each of the regions found. In case no appropriate concept can be found for a given region, we simply assign a "null" label. The training set is then used to train the SVM models and decision model.

Table 1 Concepts list

Sky, clouds, sun, animals(tiger, cow, dog, cat, rock), road, grass, plant, tree, waterfall, sea, river, lake, snow, food, fruits, people, beach, indoor, null, travel, vegetation, etc.

5.2 The Experimental Results

Table 2 presents our initial results. We show the results obtained using the soft SVM and decision models. For comparison, we also give the results obtained from the normal SVM with or without decision model. We adopt the recall, precision and F_1 measures, which are very common measures used in information retrieval.

Table 2 Initial Results on learning semantic concepts for images

SVM	**Mode**	**Automatically Checked Result (ACR)**			**Manually Checked Result (MCR)**			**Comment**
		Rec.	**Pre.**	**F1**	**Rec.**	**Pre.**	**F1**	
Soft Binary SVM (probabilistic binary SVM Classifier)	Mode 1	23.73	22.95	23.33	36.10	33.69	34.85	Blob only
		18.37	15.98	17.09	25.22	20.92	22.87	UCSB only
		21.05	19.47	20.23	30.66	27.31	28.88	Average
	Mode 2	45.00	26.49	33.35	50.45	39.03	44.02	With DT
	Mode 3	47.58	28.52	**35.35**	51.10	44.17	**47.38**	With DT
Normal SVM Classifier	Mode 4	30.43	32.24	31.31	37.40	39.52	31.31	Blob only
		20.93	27.1	23.62	26.97	34.28	30.19	UCSB only
		25.68	29.67	27.53	32.18	36.9	34.38	Average
	Mode 5	50.28	31.55	38.77	50.55	37.57	43.10	With DT

We present two sets of results, corresponding to "automatically checked Result (ACR)" and "manually checked Result (MCR)" respectively. ACR gives the learned results as compared with the ground truth of images, which are obtained from the original annotation provided by the image authors. ACR does not consider those learned relevant concepts which are not stated in the ground truth. In general, an image is assigned one or more keywords, corresponding to the main objects as perceived by the authors. For example, the image in Figure 3 is assigned the keyword "eagle" for the main object. However, it is also appropriate to include keywords such as "grass" and "mountain". Our automatic annotation method would probably be able to derive this information. This information, however, would be considered as incorrect if we based strictly on the original annotation of the image. This will give most automated approach a lower precision than it should be. MCR is designed to correct this problem. For MCR, we manually checked the learned concepts with the ground truth. If the concept(s) we learned is not in the ground truth, but the users evaluate that it is appropriate for the image, then we add the concepts learned into the ground truth. We use the updated ground truth to compute the performance measures for MCR.

Table 2 shows the results for 5 modes. Modes 1 to 3 are based on soft binary SVM with/without decision model(using decision tree, DT). Mode 4 and Mode 5 are based on normal SVM with/without decision model. Mode 1 and Mode 4 use only one of the segmentation methods without performing the decision process. The F_1 measures achievable are between 20-27% (average of 24%) for ACR and between 28-34% (average of 30%) for MCR. For Mode 2 and Mode 5, we incorporate multiple segmentation methods with decision model, but adopt the "one region, one concept" strategy. We see that we could achieve a higher F_1 measure of 33-38% (average of 36%) for ACR, and around 43-44% (43% on average) for MCR. This is significantly higher than method using only evidence derived from only one segmentation method.

Finally, for Mode 3, we adopt the "one region, one or more concepts" strategy and use the multi-segmentation methods with decision model. We could achieve an F_1 measure of over 35% for ACR, and 47% for MCR. Although the results of Mode 3 evaluated using ACR is lower than that achievable by Mode 5, it achieves the highest F_1 measure for MCR. These results clearly show that the use of multiple segmentation methods with decision model could significantly improve the performance of automatic annotation methods.

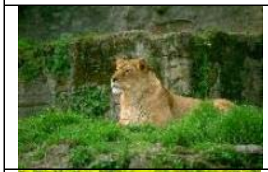
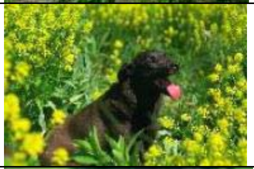
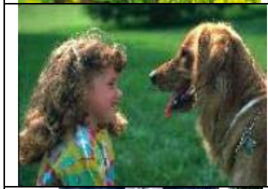
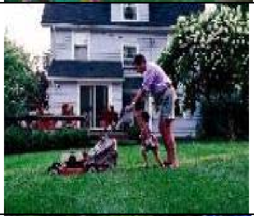
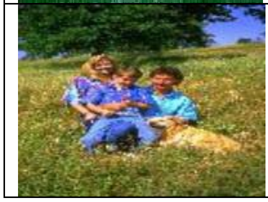
(1) Image	(2) Keywords	(3) QBK Queries
	Original: tiger, grass, rock Learned: animals, grass	Query with "animals, grass"
	Original: dog, plants Learned: animals, plant, grass	Query with "animals, grass"
	Original: girl, dog, grass Learned: people, animals, grass.	Query with "animals, grass"
	Original: people, plant, rock, vegetation Learned: people, travel, grass	Query with "people, grass"
	Original: people, plant, travel, animals Learned: people, travel, grass, sky	Query with "people, grass"

Figure 4 Examples of image annotation

5.3 Examples of Annotated Images

Figure 4 gives some examples images annotated using our approach. Column 2 of Figure 4 shows both the original annotation provided by the authors, as well as the annotation automatically learned by our system. The results show that our annotation scheme could give reasonably accurate and complete annotation. Note that as we support only "animal" as the general concept for all types of animals, specific animals such as "dog", "tiger" are tagged as "animal", which is considered correct here. Column 3 of Figure 4 gives the QBK queries that can be used to correctly retrieve the images.

6. Conclusion and Future work

Automatic annotation of images is a very challenging task. Current approaches rely on image content features extracted from segmented regions, fixed size blocks or whole images to learn the association between the visual features of images with associated concepts. Each approach has its limitations. For the image segmentation approach, because the segmentation technique is not mature, it is hard to derive semantically meaningful segmented regions to support effective learning. In this paper, we propose a novel learning-based approach to learn the semantic concepts for images. We employ multiple segmentation methods, instead of one, to derive different sets of overlapping segmented regions, and learn the association between concepts and regions independently. We then use the overlaps between regions and concepts as context in a decision model to disambiguate the concepts learned. The experiments on the middle size image collection (from PhotoCD and Web) demonstrate that our approach can greatly improve the performance of automatic annotation approach by over 12-16% on average in terms of F_1 measures compared to methods that use only one segmentation method.

Our approach is still in the early stage of research. Currently we are working on enhancing the following areas. First, instead of considering the use of two segmentation methods to extract different image content features, we will consider multiple methods in different categories such as the mixing of segmentation and fixed block size approaches etc. In addition, we could consider different learning methods in a co-training framework to improve annotation performance. Second, we need a better choice and structure of concepts in the Lexicon. We need to support not just atomic concepts but complex concepts. We also need to consider the relationships between concepts during the concept disambiguation process. Third, we need to derive better representation of context, not just in terms of overlapping regions, but in concepts as well.

Acknowledgement

The first author would like to thank the National University of Singapore for the provision of a research scholarship. The authors also would like to thank to Dr Chih-Jen Lin for sharing with us the LIBSVM library.

References

[1] John R.Smith and S-F Chang, VisualSeek: A Fully Automated Content-based Query System. In Proc Fourth Int Conf Multimedia, ACM 87-92 (1996).

[2] John R.Smilth, milind Naphade and Apostol (Paul) Natsev, Multimedia Semantic Indexing Using Model Vectors. ICME (2003).

[3] 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).

[4] K.Barnard and D.A.Forsyth, Learning the Semantics of Words and Pictures. IEEE International Conference on Computer Vision II, 408-415 (2001).

[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 Circuits and Systems for Video Technology Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description 13, 26-38 (2003).

[6] K.Barnard, P.Duygulu and D.Forsyth, Clustering Art. In Proc of IEEE Computer Vision and Pattern Recognition 434-441 (2001).

[7] S.Belongie, C.Carson, H.Greenspan and J.Malik, Recognition of Images in Large Databases Using a Learning Framework. Technical Report 07-939, UC Berkeley CS Tech Report 07-939, (1997).

[8] C.Carson, M.Thomas, S. B. , J.M.Hellerstein and J.Malik, BlobWorld: A System for Region-based Image Indexing and Retrieval. Int Conf Visual Info Sys (1999).

[9] James Z.Wang and Jia Li, Learning-based Linguistic Indexing of Pictures with 2-D MHHMs. The 10th ACM Int Conference on Multimedia 436-445 (2002).

[10] Y.Deng and B.S.Manjunath, Unsupervised Segmentation of Color-texture Regions in Images and video. IEEE Trans on Pattern Analysis and Machine Intelligence 23, 800-810 (2001).

[11] Tat-Seng Chua and Jimin Liu. Learning Pattern Rules for Chinese Named-entity Extraction. 411-418. 2002. Edmonton, Canada., AAAI'2002.

[12] Christiane Fellbaum, WordNet: an electronic lexical database, MIT Press, Cambridge, Mass 1997.

[13] José M.Martínez, Overview of mpeg-7 description tools, part 2. IEEE Multimedia 83-93 (2002).

[14] Prints and Photographs Division of Library of Congress. Thesaurus for Graphic Materials. http://www.loc.gov/rr/print/tgm2/toc.html . 2003.

[15] Vladimir Vapnik, The Nature of Statistical Learning Theory, Springer, New York 1995.

[16] Ross Quinlan. Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html . 2003.