# ImageSaker: A Semantic-based Image Retrieval System Refining with Concept Model

Ke Gao[1,2], Jian-xin Zhou[1,2], Shou-xun Lin[1], Yong-dong Zhang[1],and Sheng Tang[1]

*1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100080*

*2 Graduate University of the Chinese Academy of Sciences,Beijing,China,100039*

*{kegao, zhoujianxin, sxlin, zhyd, ts}ict.ac.cn*

**Abstract**

In this demonstration, a two-level system for semantic-based image retrieval is proposed. To overcome the shortcoming of the traditional retrieval system, we present a novel method which can provide effective retrieval result in a short time. Firstly, it uses surrounding text to get a related candidate image set. Secondly, a semantic network is used to map the keyword to one of concept models which describe the statistical character of semantic relevant images. Afterwards, the system refines the small image set using the model to get more accurate retrieval result. In order to train concept models, we propose an improved method based on SVM (Support Vector Machine). Experiments show that the proposed method is effective for WWW image retrieval.

*Keywords:* Image Retrieval, Concept Model, Semantic Network

## 1. Introduction

As large collections of images are available to the public, an efficient image annotation and retrieval system is highly desired. Although traditional text-based search engines offer reasonable recall, they are unable to express the visual feature exactly. Therefore, the last decade has witnessed the rapid development of content-based image retrieval technology [1, 2]. However, there exist an enormous gap between visual feature and high-level semantic information. Thereby, more and more scholars attempt to combine textual evidence and visual feature. This research explores the use of machine learning approaches to automatically annotate images based on a predefined list of concepts, and then the annotated concepts can be used as the basis to support keyword-based image retrieval [3].

Most adopted methods include supervised and unsupervised classification of images. Supervised classification needs labeled image set while for unsupervised classification, there isn't any prior knowledge. Simple concepts such as city, landscape, and forest have been detected through supervised classification with high accuracy [4]. Using Gaussian mixture models to learn concepts from user's feedback and form a dynamically changing image database is discussed in [5]. As SVM is a fast-growing field within pattern recognition, a novel system which uses dynamic ensemble of SVM classifiers has been proposed in [6]. The main limitation of above learning-based approaches is that for effective learning, a large set of labeled training samples is needed. And the predefined list of concepts is too limited to satisfy the practical application. To solve these problems, ImageSaker is designed to retrieval WWW images by refining text-based retrieval result with concept models.

The main contribution of this research is twofold. First, we proposed a two-level method to combine both the text annotations and visual contents of the images. Because the first level can get a small candidate image set quickly, and the second level refines the candidate set at semantic level, this method can ensure the retrieval speed and precision at the same time. Second, we present an improved method based on SVM to get the concept models, which aim to capture users' query intention, and only need a small set of training samples.

The paper is organized as follows. In section 2, we give the overview of the system. Text-based image retrieval is described in section 3. Section 4 discusses the concept-based image refining. In Section 5, we present our empirical results, and finally, Section 6 draws the conclusions.

## 2. System Design

ImageSaker supports QBK (Query by keywords) as the interaction method which is convenient and can describe users' query intention at semantic level. In order to combine low-level visual features and high-level semantic information, and improve retrieval precision in acceptable period of time, the system adopts a two-level process.

As illustrated by Figure1, the system consists of two sub-systems: text-based image retrieval and concept-based image refining. In the first sub-system, the texts around the WWW images are analyzed to get the keywords of every image in the image database. Although the keywords may not be exact, they can provide effective indexes and the first round related candidate image set. Most images of the candidate set are semantic related with the query word while the size of the image set is much smaller than the primary database. Then in the second sub-system, a semantic network is used to map the keyword to one of our concept models which have been trained off-line in advance. Since the concept models are trained based on image visual features, the refining process give prominence to visual character compared with the text-based retrieval round, so more accurate retrieval result could be expected.

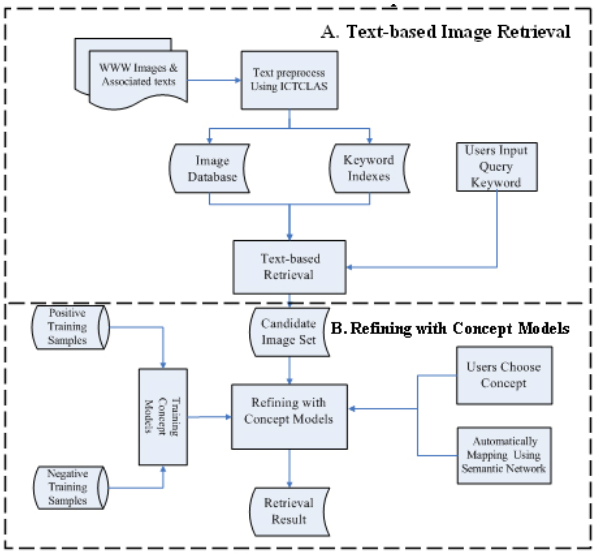


Fig. 1 System Design Frame

## 3. Text-based Image Retrieval

To satisfy practical application, textual evidences are very important for WWW image retrieval. In this paper, textual evidences are extracted from image file name, page title and alternate text.

ImageSaker uses Chinese lexical analyzer ICTCLAS [7] to analyses these collected texts. ICTCLAS is free software and available at the Open Platform of Chinese NLP (www.nlp.org.cn). ICTCLAS can achieve 97.58% in segmentation precision and the lexical analyses results are reliable.

When users input a query word, a small candidate image set with about 100 pictures will be created, which is much fewer than the original image resource. The first round process can be implemented very rapidly. However, considering the inherent meaning difference of many words and phrases, the concept-based image refining is introduced as the second round process.

## 4. Refining with concept models

In our system, we use two steps to implement the image refining: training concept models using an improved SVM and mapping query keywords to concept models using a semantic network.

### *4.1. Visual feature extraction*

Selecting suitable features is as important as designing an effective learning algorithm for Image retrieval. In our system, we intend to choose features that are representative and simple enough for on-line processing and real-time retrieval.

In order to understand user's query intention, object/region based retrieval technology is necessary. However, most of image segmentation methods such as "blobs" are too complex and time consuming to be used for commercial application. Furthermore, accurate segmentation is unnecessary for image retrieval, which is not an accurate matching task. Accordingly, as illustrated by Figure 2, we adopt a simple but effective segmentation method.

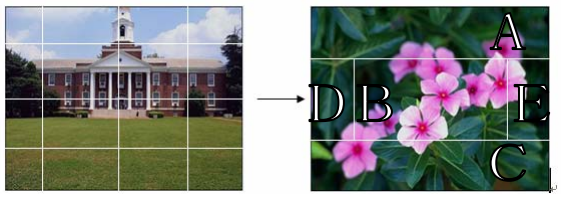


Fig. 2 Block-based image segmentation

Each image is divided equally into 4*4 rectangular pieces, and they are assembled like figure 2. There are five blocks for each image, and only three of them will be calculated. Because D and E blocks often describe the left and right limited sides of an image, which usually don't implicate useful semantic information like other blocks.

A block represents the up side of the image, which often contains things such as "blue sky". C block represents the bottom side, which often means "green grass" or "brown land". B block always has the main object of this image, and attracts most attention. After quantizing the HSV color space, ImageSaker extracts 3 color moments for A and C blocks, and extracts 32-bin histogram for B block. Moreover, 3 color moments are also taken into account to depict global color feature. As for texture feature, we adopt edge histogram descriptor which was proposed by Dong Kwon Park [8]. Experiments have shown that our visual feature extraction method is timesaving and useful.

### *4.2. Concept Model Training*

It's acknowledged that the nature of CBIR (content-based image retrieval) is to search the relevant or similar images based on low-level visual features, which implies that relevant images have similar visual features. So it is possible to cluster or classify images according to low-level visual features. Here, "concept" means some core phrases which have been selected to represent some semantic class, and images with the same concept have similar visual features. For instance, "forest" often has green as its global color, while "electronic production" often has silver gray as its background, and their texture are much more different. For each concept, a corresponding concept model is trained using an improved method based on SVM. Concept models describe the statistical character of visual features extracted from a set of images which are semantic relevant. They are designed to capture user' query intention automatically at semantic level.

#### *4.2.1. An improved method based on SVM*

Support vector machine is a well-known pattern classification method [9]. For pattern classification, SVM has a good generalization performance without domain knowledge of the problems. Being based on the structural risk minimization principle and capacity concept with pure combinatorial definitions, the quality and complexity of the SVM solution does not depend directly on the dimensionality of the input space. In the classical SVM approach, many support values are zero, and only those non-zero values correspond to the so called support vectors.

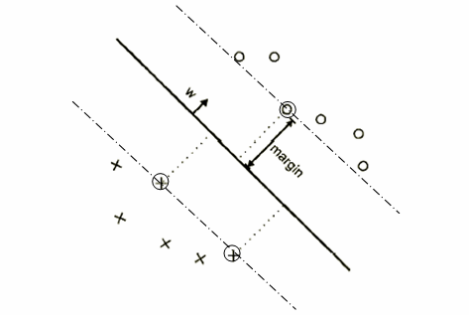


Fig.3 support vectors in feature space

Our principle behind this algorithm is: the resulting decision function of an SVM depends only on its support vectors. In other words, training an SVM on the support vectors alone results in the same decision function as training on the whole data set. Because of this, selecting appropriate training samples are very important, especially those images which are precisely corresponding to support vectors in visual feature space. They will affect the refining precision greatly.

In previous work, the training samples are often selected only by manual work, which is time-consuming, costly and subjective. In order to select the most informative images as training samples, and learn a boundary that best separates the image dataset, Simon Tong [10] had presented an active learning method for image retrieval. In this process, at each feedback round, the system selects twenty images to ask the user to label as "relevant" or "not relevant" with respect to the query concept. It then uses the labeled instances to successively refine the concept boundary. However, this method needs iteration for many times and also fails to tell the users which samples are most likely corresponding to support vectors.

In our system, we use a simple clustering method to cursorily simulate the distribution of all the samples in the image dataset, and choose training samples within those images distributing near the boundary. We believe that these samples have better discriminability and would result in a more exact distinguish function than using randomly chosen samples.

*4.2.2. Training Concept Models*

K-means clustering is used to instruct the choice of training samples in our system. Samples all over the dataset are clustered according their distance from each center, and only those samples distributing near the boundary between every two classes are sorted out to be labeled with “relevant” or “not relevant”. Flow chart of this part is shown in Figure.4.

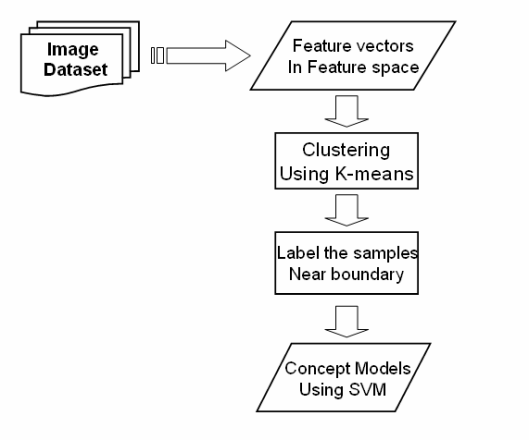


Fig.4 an improved method based on SVM

We train a set of probabilistic SVM classifiers by using the set of labeled training samples, and the output probability of each image is used to reset the sequence of these candidate images. That’s the way we refine text-based retrieval result.

In our system, we pay more attention to the close-up images, which are considered to contain more details. In other words, images which have a lot of objects are seldom taken into account.

***4.2. Mapping Keywords to Concept Models***

How to comprehend user’s intention exactly? In other words, which concept model should be chosen to refine the candidate image set when the user input a query keyword? ImageSaker provide two methods to deal with this problem. As shown is Figure 5, all of the concepts are shown on the interface of this retrieval system, and users could choose the appropriate one manually. If users only put up the query keyword without choosing any concept, the system will do the work automatically using a semantic network.


Fig. 5 Interface of ImageSaker

To improve retrieval precision, the establishment of a list of concepts is very important. In this paper, a limited set of “core” concepts is manually selected according to practical application.

The semantic network is represented by a set of keywords having links to the list of concepts which have been defined in advance. The links between the keywords and concepts provide structure for the network. On each link, the relevance degree between keyword and concept is represented as weights, which are assigned to each individual link according to experience fact. This representation is shown as Figure 6.

At present, the weights are manually labeled and changeless. In our future work, we will utilize the Internet and users’ feedback to modify the semantic network dynamically and automatically.

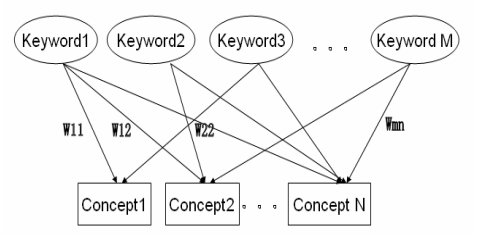


Fig.6 Semantic network

## 5. Experimental Results

Here, we take famous Chinese search engine www.baidu.com for an example to explain the necessary of the image refining.

As shown in figure 7, when we use text-based-only strategy to search for images about “手机”(mobile telephone), we may get a lot of pictures with vary different content. Consequently, if users can indicate the concept they are concerning with, for example, electronic production, and then the system will output close-up photos about new style mobile telephone, with other images being excluded.

Fig. 7 Retrieval result comparison of Baidu (left) and ImageSaker(right)

### *5.1 Test Data*

We use the Baidu image search function to gather the images. About 3,000 images with associated text evidences are downloaded and used in our evaluations. As shown is figure 3, there are 28 concepts in all to test the effectiveness of our novel method. These concepts are chosen carefully to ensure that they are frequently used in WWW image retrieval, and each of them represents distinct visual modality. We download at least 100 images for each concept to provide sufficient data for training and testing.

### *5.2 Experiment Evaluation*

To evaluate retrieval result, we choose 10 frequently used keywords and their corresponding concepts listed in Table 1 to compare the performance of Baidu and ImageSaker based on retrieval precision as shown in figure 6. To describe the images ranking sequence, we adopt average precision which is calculated with different scopes. The scope ranges from 10 to 100, and figure 7 shows the result.

The precision p and average precision $p_{ave}$ are defined as follows:

$$P = Num_{correct} / Num_{retrieved}$$

$$P_{ave} = \sum_{i=1}^{n} P_i / n$$

The correct images are determined artificially according to their image contents. The letter "n" stands for the total number of different concepts, and here n equals to 10.

Table 1. List of keywords and corresponding concepts

| |
|---|
| 1.手机（数码产品）2. 电冰箱（家用电器） 3.马尔代夫（海滨风光） 4.北京饭店（室内装潢） 5.喜马拉雅山（雪域风光） 6.AK47（枪械） 7.捷运（公共交通）8. 蒙古草原（田园风光）9.战斗机（空军）10.安徒生（人物特写） |

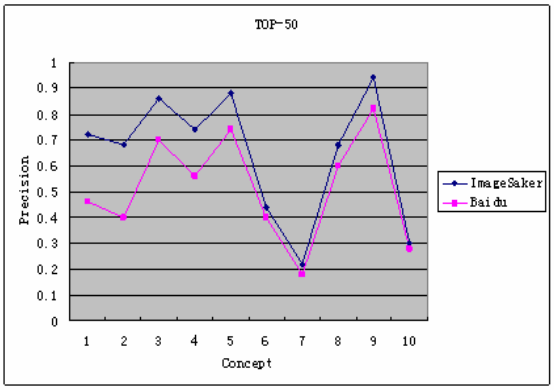


Fig.8 Retrieval result comparison of Baidu and ImageSakers with different concepts

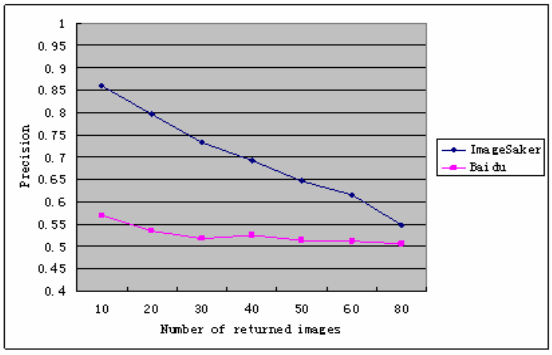


Fig. 9 Retrieval result comparison of Baidu and ImageSakers with different scopes

Experiments demonstrate that our two-level method could give reasonably accurate retrieval result. As the scope becomes larger, the precision of ImageSaker reduces regularly, which implicates that the most

semantic-relevant images are ranked to the top of the images list. Further analysis of the results reveals that the system performs well for concepts just like beach, sky and digital production, which tend to have coherent color and texture characteristics, and the texts around them are almost exact. The problems indicate that we need to develop better visual descriptor and adopt better feature space.

## 6. CONCLUSIONS

In this paper, we propose a new method for semantic-based image retrieval by refining text-based retrieval result with concept models, and a system named ImageSaker is implemented to validate our method.

To overcome the limitation of traditional text-based-only image retrieval methods, we present a two-level system. At the first level, associated WWW textual evidences are analyzed to get a small set of candidate images. At the second level, concept models are adopted to refine the candidate image set, and intend to capture user' query intention at semantic level.

Experiments show that in this system, textual evidences and visual features are combined effectively to achieve higher retrieval accuracy.

ImageSaker is available at http://soutu.ict.ac.cn.

## 7. ACKNOWLEDGEMENTS

We wish to thank Open Platform of Chinese NLP (www.nlp.org.cn) for their free software ICTCLAS. And we also would like to thank all of our team members for their generous support of this work.

This work is supported by the Key Project of Beijing Natural Science Foundation (4051004), and Beijing Science and Technology Planning Program of China (D0106008040291, Z0004024040231).

## References

[1] Ritendra Datta, Jia Li, and James Z. Wang. Content-Based Image Retrieval - Approaches and Trends. of the New Age. MIR'05,November 11-12,Singapore,2005

[2] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval: the end of the early years. IEEE trans. PAMI, vol. 22, pp. 1349 – 1380, December, 2000

[3] Huamin Feng, Rui Shi, and Tat-Seng Chua. A Bootstrapping Framework for Annotating and Retrieving WWW Images. MM'04, October 10-16, 2004

[4] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J.Zhang, Image Classification for Content-Based Indexing, IEEE Trans. Image Processing, 10(1):117−130, 2001

[5] A. Dong and B. Bhanu, Active Concept Learning for Image Retrieval in Dynamic databases, Proc. IEEE International Conference on Computer Vision, 2003

[6] B. Li, K.-S. Goh, and E. Y. Chang, Confidence-based Dynamic Ensemble for Image Annotation and Semantics Discovery, ACM Multimedia, 2003

[7] Zhang HP, Yu HK, Xiong DY, Liu Q. HHMM-Based Chinese lexical analyzer ICTCLAS. In: Proc. of the 2nd SigHan Workshop. 2003.

[8] DK Park, YS Jeon, CS Won, and S.-J. Park, Efficient use of local edge histogram descriptor, in Proc. ACM Workshop Standards, Inter-. operability, and Practice, Los Angeles, CA, Nov. 2000

[9] Vapnik V., The nature of statistical learning theory, Springer-Verlag, New-York, 1995

[10] Simon Tong, Edward Chang, Support Vector Machine Active Learning for Image Retrieval, MM'01, 2001