A World Wide Web Based Image Search Engine Using Text and Image Content Features

Bo Luo, Xiaogang Wang and Xiaoou Tang

Department of Information Engineering

The Chinese University of Hong Kong

Shatin, NT, Hong Kong

{bluo1, xgwang1, xtang}@ie.cuhk.edu.hk

ABSTRACT

Using both text and image content features, a hybrid image retrieval system for Word Wide Web is developed in this paper. We first use a text-based image metasearch engine to retrieve images from the World Wide Web based on the text information on the image host pages to provide an initial image set. Because of the high-speed and low cost nature of the text-based approach, we can easily retrieve a broad coverage of images with a high recall rate and a relatively low precision. An image content based ordering is then performed on the initial image set. All the images are clustered into different folders based on the image content features. In addition, the images can be re-ranked by the content features according to the user feedback. Such a design makes it truly practical to use both text and image content for image retrieval over the Internet. Experimental results confirm the efficiency of the system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering, Relevance feedback, Retrieval models, Search process.*

General Terms

Design, Performance, Experimentation.

Keywords

Content based image retrieval, WWW search engine, meta-search engine, image search engine

1. INTRODUCTION

As the diversity and size of digital image collections grow exponentially, efficient image retrieval is becoming increasingly important. In general, current automatic image retrieval systems can be characterized into two categories: text-based and image content-based. For text-based image retrieval, the images are first annotated by text and then the text-based Database Management Systems are used to perform image retrieval. In this framework, manual image annotation is extremely laborious and the visual content of images are difficult to be described precisely by a limited set of text terms. To overcome these difficulties, content-based image retrieval systems index images by their visual content, such as color, shape, texture, etc. Some representative text-based and visual content-based image retrieval systems can be found in [1, 2, 3, 4, 5].

Most of the visual content based image retrieval systems are based on image database. The image resource in the database is limited and updated slowly. With the development of internet technology, the fast growing World Wide Web has become one of the most important sources of visual information. Efficient tools are needed to retrieve images from the Web. Comparing to image databases, the Web is an unlimited, immense repository of images, covering much broader resources, and is increasing at an astonishing speed continuously. However, the Web is also a completely open information system without a well-defined structure. Image retrieval from the World Wide Web has to overcome great difficulties concerning speed, storage, computational cost, and retrieval quality. Some content-based image search engines [6, 8, 9, 10] use Web crawlers to continuously traverse the Internet, collect images, and extract features from the images. However, given the unlimited data size, the demand on computational power, image transmission cost, and image storage quickly becomes a bottleneck for these systems. Practical network based image retrieval based on this approach is very difficult if not impossible to achieve. Therefore, keyword-based general WWW search engines [6, 7] seem more realistic for internet image retrieval. In fact, practically all the current commercial image search engines are based on text descriptions. The main problem with such a system is that the retrieved images have relatively low relevance to user's need.

In this paper, a hybrid image retrieval system for World Wide Web is proposed. The system performs image retrieval in two steps. In the first step, a text-based image metasearch engine retrieves images from the Web using the text information on the host page to provide an initial image set. Since the complexity and cost of text-based retrieval is much lower than that of CBIR, this step can be performed at a low computational cost, low storage requirement, and low transmission bandwidth, with near real-time speed. Then CBIR methods based on clustering and ranking with relevant feedback are applied to the initial image set to improve the relevance of the retrieval output. This approach combines the advantages of both the text-based and visual content-based approaches to achieve both high speed and high precision image retrieval. Experimental results show that our system provides an effective and efficient way to image retrieval from the World Wide Web.

The paper is organized as follows. In Section 2, a description on the text-based image metasearch engine is given. In Section 3, the image content based ordering on the initial image set is presented. The system implementation is presented in Section 4. Section 5 gives the experimental results on the system performance. Finally conclusions are drawn in Section 6.

2. TEXT CONTENT BASED RETRIEVAL

Unlike image retrieval from a fixed database, where each image is treated as an independent object, for image retrieval over the Web each image comes along with a host page, which contains a great deal of relevant information about the image. In general, for most of the images on the Web, their content is more or less related to the content of the host pages. For example, a photo of Mars is found more likely from a page talking about space and planets than from a page talking about pop-music. Therefore, we can use not only the image file names but also the page titles and text terms around the images to index and retrieve the images. This actually makes text-based image retrieval more efficient over the Internet than over a database since manual annotation is no longer required.

The text-based approach is also significantly more efficient than the CBIR system on the Internet, in terms of computational cost as well as image transmission and storage cost. For the CBIR system, it is next to impossible to transmit, store, and compute content features for the unlimited amount of images on the Internet. On the other hand, text document retrieval over the Internet has become a routine task for a commercial web search engine. Retrieving images based on text is even simpler since only the portion of the document around the image needs to be searched. Due to the low cost, text-based approach can retrieve significantly more relevant images over the Internet, therefore gives a much higher recall rate than the CBIR approach.

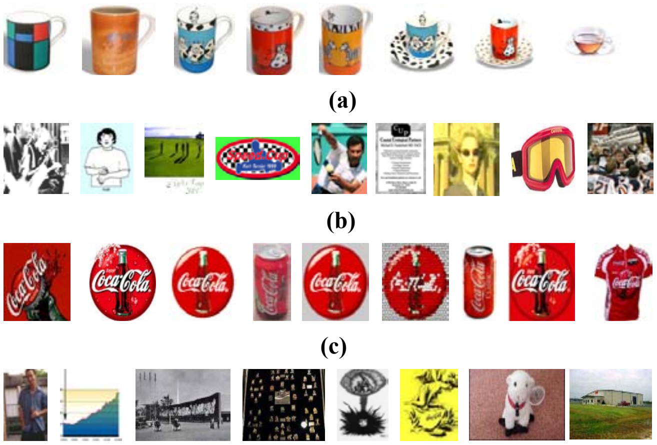
However, accompanying the relevant search results, there could be a large number of irrelevant search results, i.e. the precision of the text-based search can be low. In many situations, a few words cannot precisely describe the image content, and many words have multiple meanings. For example, the query term *sun* may retrieve photos of the Sun or the logos of SUN Microsystems company. Here, the definition of relevancy depends on the interest of the user. With a low-precision retrieval result, a user may soon lose patience flipping through dozens of pages of images that contain many irrelevant images.

To overcome the disadvantage of low-precision rate of the text-based approach, we propose to use the CBIR approach to re-filter the search results. Since relevant retrieved images tend to contain similar visual features, through visual content-based clustering or relevant feedback re-ranking, we can extract a high-precision set of relevant images from the text-based results. For the above example, images containing the Sun and images containing the SUN Microsystems logos should fall into different visual clusters. Since we only need to compute visual features for a limited set of images retrieved by the text-based query, the computation, transmission, and storage cost is very small.

Since the CBIR processing can only increase the precision of the search results at the expense of a relatively small drop of recall rate, it is to our advantage to obtain a high recall rate at the first step text-based retrieval. Toward this end, we use a meta-search engine for the text-based retrieval. Since each individual search engine use different strategy for image indexing and retrieval, the returned results can be quite different. To cover a wide range of retrieval results in order to achieve a high recall rate, we use a meta-search engine to combine several text-based image search engines.

3. IMAGE CONTENT BASED ORDERING

Let's first look at some sample results from the text-based image retrieval over the Internet. Figure 1 shows some relevant and irrelevant results of two image search engines for two text queries. We notice that, for a given text query term, the items in the *relevant* set have similar visual features (e.g. images in group a are similar in shape, images in group c are similar in color), while the images in the *irrelevant* set differ greatly from each other. A visual feature based clustering method should be able to group the *relevant* images together thus give a better retrieval performance.



(a)

(b)

(c)

(d)

Figure 1. Text term search results of some internet image search engines. (a) Relevant results from *PicSearch* with the term ***cup***; (b) Irrelevant results from *PicSearch* with the term ***cup***; (c) Relevant results from *Google* with the term ***cocacola***; (d) Irrelevant results from *Google* with the term ***cocacola***

The *relevant* and *irrelevant* mentioned above are based on the judgment of the user, who expects to retrieve images related to his query text and his anticipation. Our goal is to analyze the visual content automatically to provide a result that fits both the text description and the user's expectation.

In the following sections, we use two visual content-based processing methods on the initial text retrieval results to improve the retrieval performance. One is the unsupervised clustering method and the other is ranking with user's relevant feedback.

3.1 Unsupervised Clustering

Since the expected relevant images are likely to be similar to each other in content, we can perform an unsupervised clustering to group them together. To illustrate the process, we use a very simple but representative feature set and a simple clustering algorithm. A color histogram in the HSV color space described in the *ScalableColor Descriptor* in MPEG 7 [11] is used as the feature vector. We quantify the HSV color histogram with 16 bins in H, and 4 bins each in S and V (256 color bins in total). We choose k-means algorithm for the feature clustering. Experimental results show that most of the *relevant* images gather into one or two clusters because of their similarities, while *irrelevant* images assemble in other clusters.

3.2 Ranking with Relevant Feedback

Another reordering approach is to bring some "typical" images to the user and re-rank all the items based on the user's relevant feedback. The selection of "typical" images could be some representative items of each cluster. It could also be some top ranked items by the text-based retrieval. The user could select one or more items to express his anticipation. Thus the issue turns to a query-by-example problem on a relatively small image set. By analyzing the similarity or distance between the example image(s) and other items in the set, we are able to give a new ranking of the items.

Query-by-example and relevant feedback algorithms have been studied extensively in the CBIR research. Our experimental results show that, even with the simplest retrieval method, the new system provides a much more relevant and reasonable ranking than the purely text-based systems.

4. THE META-SEARCH ENGINE IMPLEMENTATION

We now can build an efficient image retrieval engine on the Internet based on the two-step text and visual content based approach. As discussed in section 2, there are already several text-based image search engines on the Internet. We implement the first step of our system (text content based query) as a meta-search engine, which sends queries to other text-based image search engines and summarizes their results to build an initial image set for later image content based processing. This brings us a more comprehensive initial data set with a higher recall than using a single search engine. Although the precision is comparatively low, it can be greatly improved in later steps.

Figure 2 shows the system architecture of the hybrid meta-search engine. We can summarize the image retrieval procedure as follows:

1. User inputs a few query text terms to the browser.
2. The Query Translator extracts the query terms from the HTTP request, then translates the query terms into the input format for each text-based image search engines, such as *Google Image Search, PicSearch, AltaVista Image Search*, and passes to the Page Crawler.
3. The Page Crawler sends the query to each search engine and collects the HTML files containing the URLs of images retrieved by the search engines, then parses the HTML files to obtain the individual URL.
4. The Result Collator merges the results and shows the first page of retrieved images and URLs to the user, and at the same time, send all the URLs to the Image Crawler.
5. Using the URLs, the Image Crawler retrieves the images from the Internet to construct the initial image set;
6. Feature Extractor computes the image content feature vectors for all images in the initial image set.
7. Upon the user's new request, cluster the image set using the feature vectors and k-means algorithm. Return the clustered images in a new HTML file for display in the user browser.
8. Based on the feedback image(s) selected by the user, compute the distance of feature vectors between the feedback image(s) and the images in the initial image set. Re-rank the images according to the distance and display the re-ranked images to the user.

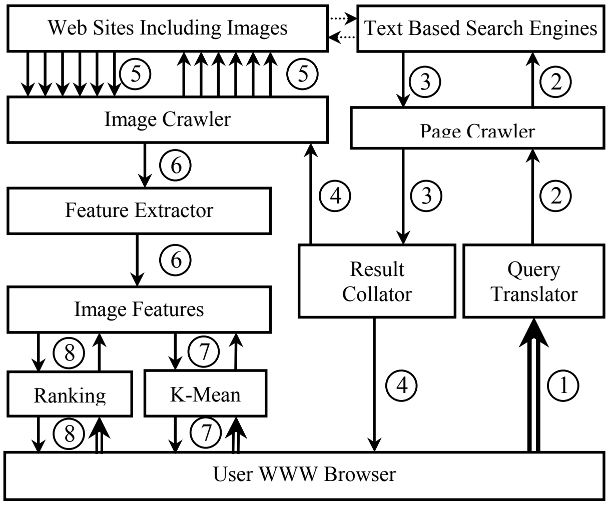


Figure 2 Architecture of the hybrid image retrieval system

5. EXPERIMENTS

A set of experiments is conducted to test the performance of the system. We use the recall and precision measures for retrieval performance evaluation [12].

- Recall is the fraction of the relevant images (*R*) which has been retrieved (*Ra*), i.e.,

$$Recall = \frac{|Ra|}{|R|} \quad (1)$$

- Precision is the fraction of the retrieved images (*A*) which is relevant (*Ra*), i.e.,

$$Precision = \frac{|Ra|}{|A|} \quad (2)$$

5.1 Text-based Image Retrieval

Five terms, "sun", "forest", "ocean", "CocaCola", and "desert", are used as the query terms. For each query, the first 60 images returned from text-based image metasearch engine are chosen as the initial image set *(I)*. A relevant set *(R)* from the initial image set is defined as the set of images containing the subject of the query term. (For example, the relevant set for "forest" is the set of images containing forest.) The number of relevant images,

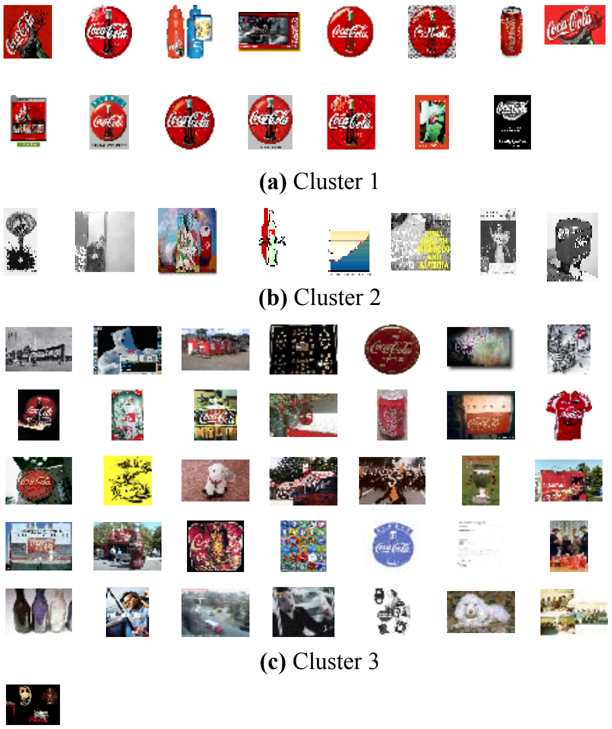
|R|, in the initial image set for each query is shown in Table1. The results show that the text-based retrieval can provide many relevant images in the first step, but the precision is low.

Table 1. Number of relevant images for five queries

Term	sun	forest	ocean	CocaCola	desert
|R|	14	30	35	30	28

5.2 Content-based Image Clustering

The 60 images in the initial set are clustered into 4 folders based on the color feature. Figure 3 shows the clustering result for the query term "Cocacola". Most of the images relevant to "Cocacola" are grouped into the first cluster which has a very high precision.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

Figure 3. Clustering result for the query term "CocaCola"

Reordering the clusters by precision, the recall (r) and precision (p) for each cluster are shown in Table 2. In Table 2, C1 is the most relevant cluster with the highest precision, and C4 is the most irrelevant cluster. Refering to the definition of recall and precision in (1) and (2), here, *R* contains all the relevant images within the 60 images retrieved by the text-based method, *A* is the cluster set, and *Ra* is the relevant image set in each cluster. So the recall computed here is different from the true recall rate of the system, which should be the ratio between the number of relevant retrieved images and the number of all relevant images on the Internet. Since it is impossible to computer the true recall rate, we use the recall in Table 2 to illustrate the performance of the clustering technique. Figure 4 shows the average precision and recall for each cluster based on the five queries. The average precision and recall for C1 are 92% and 51%. That means the user can easily find most of the relevant images in this cluster. Since a user is usually looking for "some relevant results" rather than "all the relevant stuff" on the Internet, C1 cluster is more meaningful to the user than pages of images which may or may not be relevant.

Table 2. Recall and precision for each cluster

		C1	C2	C3	C4
"sun"	p (%)	100	29	15	11
	r (%)	43	14	21	21
"forest"	p (%)	73	33	19	17
	r (%)	73	3	13	3
"ocean"	p (%)	86	83	22	17
	r (%)	71	14	11	3
"CocaCola"	p (%)	100	52	33	0
	r (%)	50	43	7	0
"desert"	p (%)	100	89	36	0
	r (%)	18	32	53	0
average	p (%)	92	57	25	9
	r (%)	51	21	21	7

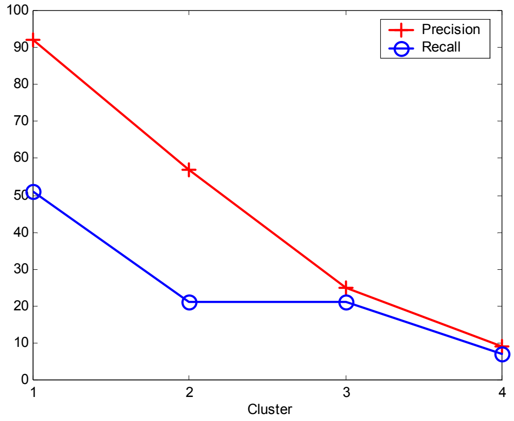
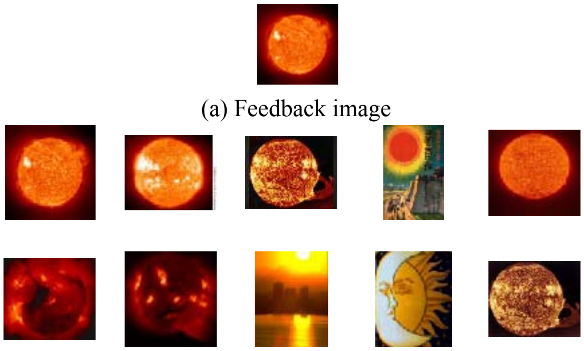


Figure 4. Average precision and recall for the 4 clusters. The horizontal axis represents the cluster number, and the vertical axis represents the precision and recall rate.

5.3 Ranking with Relevant Feedback

The 60 images are re-ranked according to the user's feedback. Figure 5 shows a feedback image and the top 10 re-ranked images for the query term "sun".

To illustrate the improvement to the retrieval performance using the feedback mechanism, we calculate the curve of precision versus recall by averaging the previous five queries. Assuming that the user examines the images by rank from high to low, *R* is relevant image set for the initial 60 images, *A* is the set of images the user has examined and *Ra* is the relevant image set from *A*. The more images are examined, the larger recall can be achieved. The two curves for the re-ranked images set and the initial image set are plotted in Figure 6. The curve for the re-ranked image set is high above that of the initial image set. This means in order to get the same number of relevant images the user only needs to examine a much smaller number of images after feedback re-rank.



(a) Feedback image

(b) Top 10 re-ranked images

Figure 5. A feedback image and the top 10 re-ranked images for query with the term "sun".

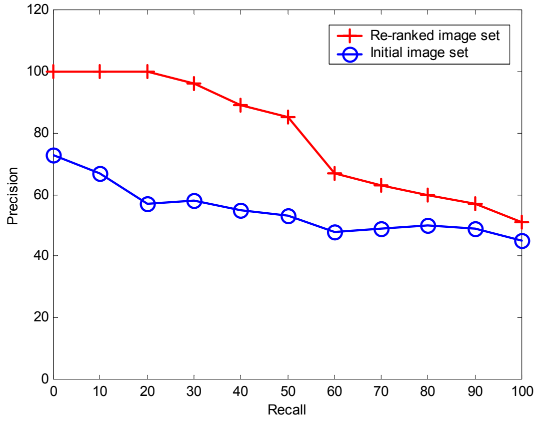


Figure 6. Average recall versus precision figure for the initial image set and the re-ranked image set.

6. CONCLUSION

The main purpose of this paper is to provide an efficient and truly realizable approach for WWW based image retrieval. We first perform a text-based meta-search to obtain an initial image set with relatively high recall rate and low precision rate. Then the image content based processing is employed to produce a much more relevant output.

There are three ways to combine the text-based method and the visual content-based method: use the text-based method first; use the visual content-based method first; use the two methods at the same time. The key to the success of our system is using the first approach. By using the high-recall, low-precision, and low-cost text-based method first, we can easily collect as many relevant images as possible over the Internet at a very low cost. Then the high-precision and high-cost visual based method is used to improve the relevance precision on a significantly smaller image set.

Experimental results show that, even with the simplest image feature and clustering algorithm the system achieves promising results. More extensive experiments are needed to test the system. We expect improved performance using more elaborate visual features to describe the image content. Better clustering and ranking algorithms will also help. Because of the first step text-based method greatly limited the number of images that need to be processed, we can afford to use more complicated visual features in the second step.

7. REFERENCES

[1] Hideyuki Tamura and Naokazu Yokoya. Image database systems: A survey. Pattern Recognition, 17(1), 1984.

[2] Shi-Kuo Chang and Arding Hsu. Image information systems: Where do we go from here? IEEE Trans. On Knowledge and Data Engineering, 4(5), Oct. 1992.

[3] W. Niblack, R. Barber, and et al. The QBIC project: Querying images by content using color, texture and shape. In Proc. SPIE Storage and Retrieval for Image and Video Database, 1996.

[4] Amarnath Gupta and Ramesh Jain. Visual information retrieval. Communication of the ACM, 40(5), 1997.

[5] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image database, International Journal of Computer Vision, 1996.

[6] Remco C. Veltkamp, Mirela Tanase. Content-Based Image Retrieval Systems: A Survey, Technical Report UU-CS-2000-34, October 2000.

[7] J. R. Smith and S.-F. Chang. Visually searching the web for content. IEEE Multimedia Magazine, 4(3): 12-20, 1997.

[8] I. Kompatsiaris, E. Triantafylluo, and M. G. Strintzis. A World Wide Web Region-Based Image Search Engine. International Conference on Image Analysis and Processing, 2001.

[9] Michael J. Swain, Charles Frankel, Vassilis Athitsos. WebSeer: An Image Search Engine for the World Wide Web. Technical Report TR96 -14, University of Chicago, July 1996.

[10] Sclaroff S., Taycher L. and Cascia M. L. ImageRover: a content-based image browser for the World Wide Web. Proc of IEEE Workshop on Content-based Access of Image and Video Libraries, 1997.

[11] Leszek Cieplinski, Munchurl Kim, Jens-Rainer Ohm, Mark Pickering, Akio Yamada. Text of ISO/IEC 15938-3/FCD Information Technology – Multimedia Content Description Interface – Part 3 Visual, March 2001.

[12] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval, ACM Press, 1993.