IMAGE COLLECTOR: AN IMAGE-GATHERING SYSTEM FROM THE WORLD-WIDE WEB EMPLOYING KEYWORD-BASED SEARCH ENGINES

Keiji Yanai

Department of Computer Science,
The University of Electro-Communications, JAPAN

ABSTRACT

Due to the recent explosive progress of WWW (World-Wide Web), we can easily access a large number of images from WWW. There are, however, no established methods to make use of WWW as a large image database. In this paper, we describe an automatic image-gathering system from WWW employing keywords and image features, which is called the Image Collector. By exploiting some existing keyword-based search engines and selecting images by their image features, our system obtains, with high accuracy, images that are strongly related to query keywords. We have implemented the system that gathers more than one hundred images from WWW in about five minutes.

1. INTRODUCTION

Due to the recent explosive progress of WWW (World-Wide Web), we can easily access a large number of images from WWW. Therefore, we can consider WWW as a huge image database. However, most of those images on WWW are not categorized in terms of their contents and are not labeled related keywords. We can use commercial search engines in order to search WWW for HTML documents by giving them keywords. In the similar way, some search engines can also search for images related to keywords. However, most of image search engines search for images based on only keywords in HTML documents including images without analyzing contents of images. As a result, they tend to return images that are not appropriate images to the given keywords.

As a method of image search, content-based image retrieval (CBIR) has been researched [1, 2, 3]. Conventional keyword-based image search requires that all images in a database are attached keywords to by hand in advance, while CBIR doesn't require attaching keywords. In CBIR, the similarity between images is computed using image features extracted from images, and we can search similar images to query images.

To achieve image search for WWW based on only keywords but also contents of images, in this paper, we propose an automatic image-gathering system from WWW that is constructed by integrating a keyword-based search method and a CBIR method, which is called the Image Collector. In our system, a user gives query keywords to the system in the beginning, and obtains images associated with the keywords finally. First, using the existing commercial WWW search engines for HTML documents, the system gathers images embedded to HTML documents related to query keywords. Next, the system selects output images from collected images based on image features extracted from images themselves. We describe a method of image-gathering, implementation of a prototype system and experimental results.

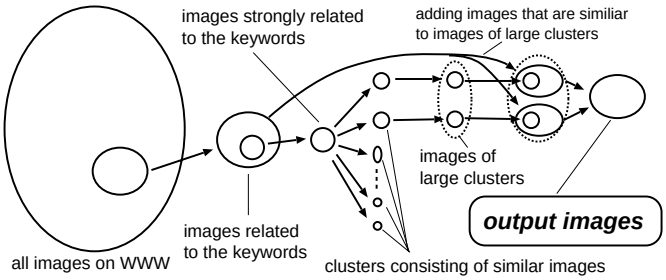

Figure 1: Processing flow of the image-gathering from WWW.

2. A METHOD OF IMAGE-GATHERING

The final goal of our image-gathering system is gathering images on WWW related to the query keywords given by a user. Note that our system is not called an image "search" system but an image "gathering" system, since our system has the following properties: (1) it does not search for images over the whole WWW directly, (2) it does not make a database in advance, and (3) it makes use of search results of commercial keyword-based search engines for query keywords..

Figure 1 shows the processing flow. Since an image on WWW is usually embedded in an HTML document that explains it, the system exploits some existing commercial keyword-based WWW search engines, and it gathers URLs (Universal Resource Locator) of HTML documents related to query keywords. In the next step, using those gathered URLs, the system fetches

HTML documents from WWW, analyzes them, and evaluates the extent of relation between the keywords and images embedded in HTML documents. If it is judged that images are related to keywords, the image files are fetched from WWW. According to the extent of relation to the keywords, we divide fetched images into two groups: images in group A having stronger relation to the keywords, and others in group B. For all gathered images, image features are computed.

In CBIR, a user has to provide query images to the system, because it searches for images based on the similarity of image features between query images and images in an image database. In our system, instead of providing query images, a user only needs to provide query keywords to the system. Then, we select images strongly related to the keywords as group A images, remove noise images from them, and regard them as query images only by examining keywords. Removing noise images is carried out by eliminating images which belong to relatively small clusters in the result of image-feature-based clustering for group A images. Images which are not eliminated are regarded as appropriate images to the query keywords, and we store them as output images. Our preference of larger clusters to smaller ones is based on the following heuristic observation: an image that has many similar images is usually more suitable to an image represented by keywords than one that has only a few similar images. Next, we select images that are similar to the query images from group B in the same way as CBIR, and add them to output images.

Some WWW image search systems such as WebSeer[4], WebSEEk[5] and Image Rover[6] have been reported so far, which can be regarded as an integration of keyword-based search and content-based image retrieval. These systems search for images based on the query keywords, and then a user selects query images from search results. After this selection by the user, the systems search for images that are similar to the query images based on image features. These three systems carry out their search in an interactive manner. Our system is different from those in that our system only needs one-time input of query keywords. Our system is able to gather a large number of various images related to the keywords, since it is unnecessary for a user to indicate query images during the processing, and the whole processing is executed automatically. The three systems quoted above require gathering images over WWW in advance and making big indices of images on WWW. In contrast to those systems, due to exploiting existing keyword-based search engines, our system does not require making a large index in advance.

3. COLLECTION AND SELECTION

The processing of the Image Collector consists of a collection stage and a selection stage.

3.1. Collection Stage

In the collection stage, the system obtains URLs by means of some commercial keyword-based WWW search engines, and by using those URLs, it gathers images from WWW. The algorithm as follows:

1. A user supplies the system with query keywords.
2. The system sends queries to commercial keyword-based search engines, and obtains URLs of HTML documents related to the keywords.
3. The system fetches HTML documents indicated by the URLs from WWW.
4. The system analyzes HTML documents, and extracts URLs of images embedded in the HTML documents with image-embedding-tags ("`IMG SRC`" and "`A HREF`"). For each of those images, the system calculates a score which represents the intensity of relation between the image and the query keywords. The score is calculated by checking the following conditions:

   Condition 1: Every time one of the following conditions is satisfied, 3 points are added to the score.

   - In case the image is embedded by "`SRC IMG`" tag, "`ALT`" field of "`SRC IMG`" includes the keywords.
   - In case the image is linked by "`A HREF`" tag directly, words between "`A HREF`" and "`/A`" include the keywords.
   - The name of the image file includes the keywords.

   Condition 2: Every time one of the following conditions is satisfied, 1 point is added to the score.

   - "`TITLE`" tag includes the keywords.
   - "`H1, ..,H6`" tags include the keywords, if these tags are located just before the image-embedding-tag.
   - "`TD`" tag including the image-embedding-tag includes the keywords.
   - Ten words just before the image-embedding-tag or ten words after it include the keywords.

   If the final score of an image is higher than 3, the image is classified into group A. If it is higher than 1, the image is classified into group B. The system fetches only image-files whose image belongs to either group A or B. If the size of a fetched image-file is larger than a certain predetermined size, the image is handed to the selection stage.
5. In case the HTML document does not include image-embedding-tags at all, the system fetches and analyzes other HTML documents linked from it in the same manner described above, if it includes a link tag ("`A HREF`") which indicates URL of HTML documents on the same web site.

3.2. Selection Stage

In the selection stage, the system selects appropriate images for query keywords out of images collected in the collection stage. The selection is based on image features as described below.

1. For all collected images, first, the system makes a color histogram as image features [7]. We make a color histogram not from the RGB color space directly, but from the Lu^*v^* color space that converted from the RGB color space, because the Lu^*v^* color space is known to represent human color sense better than the RGB color space. We quantize the Lu^*v^* color space into 216 (6 for each axis) bins, and make a color distribution histogram for each collected image.
2. For images in group A, the distance which represents the degree of dissimilarity between two images is calculated based on image features. In the calculation of the distance, we adopt not the Euclid distance but the distance which considers the proximity in the color space [7].
3. Based on the distance between images, images in group A are clustered by the cluster analysis method. We adopt the farthest neighbor method (FN): we define the distance between clusters as the largest distance between two images belonging to mutually different clusters. In the beginning, each cluster has only one image. For each pair of clusters, if the distance between them is smaller than a certain threshold, they are merged into the same cluster. The system repeats merging clusters, until all distances between clusters are more than the threshold.
4. The system throws away small clusters which have fewer images than a certain threshold value. It stores all images in the remaining clusters as output images.
5. The system selects images from group B if they have a small distance from images in the remaining clusters of group A, and adds them to output images.

4. EXPERIMENTS

We have implemented an experimental system in C and Perl on Linux-based PC (CPU:Athlon 750Mhz, memory:384MB). In the experiments, we limited WWW sites to access to only Japanese (.jp) domain.

We show experimental results for eight keywords in Table 1, which describes the number of URLs of HTML documents obtained from search engines, the number of images collected from WWW, and the number of selected images.

In the collection stage, we used five major Japanese search engines, Goo, Infoseek Japan, Lycos Japan, Oc-n Navi and Excite Japan to obtain URLs related to query keywords, and merged the search results of five engines by omitting duplications. In each experiment, we obtained about 2000 URLs in about ten seconds. We fetched and analyzed HTML documents of about 2000 URLs, and we got several hundreds of images from WWW in about three or four minutes. Fetched images were divided into two groups, A and B, by analyzing HTML documents as shown in Table 1.

In the selection stage, we selected images from group A by image-feature-based clustering and removing small clusters, and selected images from group B by CBIR. This processing took about one minutes. Total processing time was about five minutes.

We judged selected images either as OK or NG by the subjective evaluation. OK means that the image exactly corresponds to the keywords, and NG means that it does not. In Table 1 we describe the precision, which is defined to be $N_{OK}/(N_{OK} + N_{NG})$, and the recall, which is defined to be $N_{OK_{sel}}/N_{OK_{col}}$, where N_{OK}, N_{NG}, $N_{OK_{sel}}$, and $N_{OK_{col}}$ are the number of OK images, the number of NG images, the number of OK images in selected images, and the number of OK images in collected images, respectively. Both the precision and the recall of images selected from group A are over 90% except three keywords. This shows that most of high-scored images at the keyword-based evaluation are correct. The precision of images selected from group B is between 36% and 83%. It is superior to the precision of images collected as group B in all experiments. It was, however, less than the precision of images selected from group A, since images fetched in group B included many inappropriate ones, and the system selected some of them by mistake. Figure 2 and Figure 3 show "lion" images selected from group A and B, respectively.

As the final output of each experiment, we obtained output images the number of which was about half of the number of collected images. Both the precision and the recall of output images are about 70%, except the precision of "tiger" and the recall of "Mt.Fuji", which implies that our method is effective for image-gathering from WWW.

Since Mt.Fuji is the most popular mountain in Japan, there are many images of Mt.Fuji in Japanese web sites. There are relatively fewer images of "Nomo" than images related to other keywords, since "Nomo" is a person's name.

5. CONCLUSIONS

In this paper, we described a method, implementation and experiments of an automatic image-gathering system from WWW. We have achieved the high precision that are about 70% without any knowledge about target images by means of both the keyword-based selection and the image-feature-based selection.

Table 1: Experimental results. This table describes the number of URLs obtained from search engines, the number of collected images from WWW, the number of selected images out of them. Numerical values in () represent the precision and the recall of the collected or selected images.

query keywords	num. of URLs	images in group A		images in group B		total (A+B)	
		collected	**selected**	collected	**selected**	collected	**selected**
lion	1979	72 (84)	**62 (93,95)**	216 (26)	**66 (42,49)**	288 (41)	**128 (67,73)**
apple	2054	97 (86)	**76 (95,87)**	237 (50)	**99 (72,60)**	334 (61)	**175 (82,71)**
baby	2031	85 (48)	**73 (53,95)**	528 (74)	**272 (83,58)**	613 (70)	**345 (77,62)**
desk	2112	76 (90)	**72 (92,97)**	212 (50)	**84 (71,56)**	288 (61)	**156 (81,72)**
keyboard†	2194	39 (95)	**38 (95,97)**	167 (60)	**58 (73,43)**	206 (66)	**96 (82,57)**
tiger	2006	57 (71)	**51 (75,95)**	178 (33)	**71 (42,50)**	235 (42)	**122 (56,69)**
Nomo‡	1778	38 (95)	**34 (97,92)**	28 (25)	**14 (36,72)**	66 (65)	**48 (79,88)**
Mt.Fuji	1981	541 (71)	**317 (91,75)**	837 (42)	**158 (66,30)**	1378 (53)	**475 (82,53)**

†. personal computer's keyboard ‡. name of a Japanese major leaguer

Figure 2: "Lion" images selected from group A.

Figure 3: "Lion" images selected from group B.

In the current implementation, we use only a simple keyword-based search method and simple image features for image-clustering. For future works, we plan to exploit more sophisticated keyword-based search methods and image features.

In addition, since the collection stage and the selection stage are separated in the present implementation, the processing time becomes quite long. We plan to implement a parallel system by integrating two stages on PC cluster system in order to achieve speed-up of the processing time.

Acknowledgments

A part of this work was supported by grants from the Telecommunications Advancement Foundation and the Okawa Foundation for Information and Telecommunications.

6. REFERENCES

[1] V.N. Gudivada and V.V. Raghavan, "Content-based image retrieval-systems," *IEEE Computer*, vol. 28, no. 9, pp. 18–22, 1995.

[2] A. D. Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, 1999.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and J. Ramesh, "Content-based image retreival at the end of early years," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[4] C. Framkel, M.J. Swain, and V. Athitsos, "Webseer: An image search engine for the world wide web," Tech. Rep. TR-96-14, University of Chicago, 1996.

[5] J. Smith and S.F. Chang, "Visually searching the web for content," *IEEE Multimedia*, vol. 4, no. 3, pp. 12–20, 1997.

[6] S. Sclaroff, M. LaCascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Computer Vision and Image Understanding*, vol. 75, no. 1/2, pp. 86–98, 1999.

[7] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.