2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Proposal of the Site-wise Abstract Image for the Web Image Resource Mining

Keisuke Shigemori*[1], Zoran Stejic*[3], Kaoru Hirota*[3], Toru Yamaguchi*[1]*[2], Yasufumi Takama*[1]*[2]

*[1] Tokyo Metropolitan Institute of Technology
*[2] Japan Science and Technology Corporation (JST)
*[3] Tokyo Institute of Technology

Address : 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
Email : ytakama@cc.tmit.ac.jp, shigemori@krectmt3.tmit.ac.jp

Abstract - As the web is vast and disorderly, it is difficult to find desired information on the web. In particular, finding image resources (knowing where and what kind of images can be found on the web) is very difficult but challenging. As the first step towards the web resource mining, this paper reports the preliminary results of collecting a number of images by a web robot as well as presenting those meta information.

1. Introduction

The recent development of information technology has realized a database of large capacity and low cost. As a result, the disorder of web is increasing. However, we must obtain useful information from the disorderly web. In particular, it is more difficult to search and collect images than text, though there are a large number of images on the web. This paper aims at discovering image resources on the web so that we can find a lot of images efficiently from the web.

2. Discovery of web image resource

We define web image resource as a group of image sites on the web (Fig.1). The sites are linked each other. Web image resource owns a lot of images. It is difficult and inefficient to obtain many interesting images one by one [2]. Therefore, discovery of web image resource makes such tasks easy and efficient. This paper suggests the presentation method of meta information about web image resource as the first step toward the discovery of web image resource.

The method for finding image resources on the web is as follows:

(1) Collection of images and meta information by a web robot.
(2) Presentation of meta information about collected image.

(3) Presentation of the site-wise abstract image based on LSP (Local Similarity Pattern).
(4) Presentation of the relation among sites based on a keyword map.

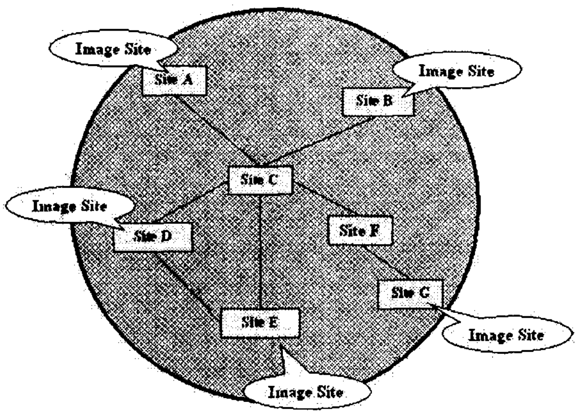

Fig.1 : Web image resource

3. Collecting image information by web robot

First we develop the web robot that collects images and those meta information from the web. A web robot is the program that traces a link on the web and collect information [3]. Depth-first search (Depth Level 5) is employed as searching strategy (Fig.2). Yahoo!'s category URL is used as a seed URL.

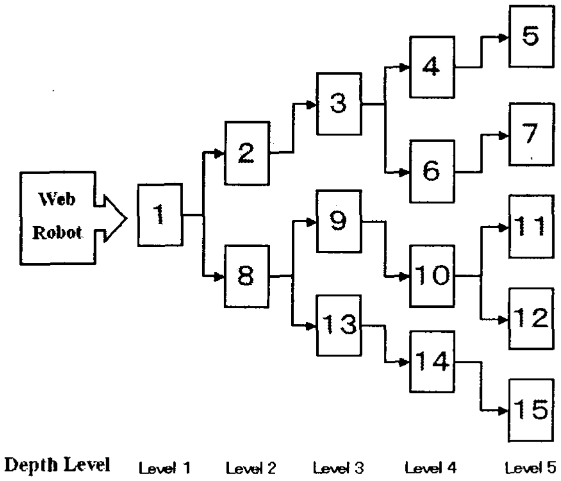

Fig.2 : Searching strategy

The following information is collected with an image as its meta information: Image itself, its filename in a local database, original name found on the web, the time when it was saved, URL of the page containing it, and URL of the image itself (Fig.3). The collected data is shown in Table.1.

PicKosmo.jpg
PicKosmo.jpg
Fri Dec 20 02:49:59 JST 2002
http://www2r.biglobe.ne.jp/%7Enmom/guestbook2.html
http://www2.justnet.ne.jp/~rmom/PicKosmo.jpg

Fig.3 : Example of image and its meta information

Table.1 : Collected data

		Data1	Data2
Category		Photographer	Illustrator
Page		5450 page	2947 page
Site		228 site	124 site
Total Number		23811 pieces	12445 pieces
Total Size		366,189,855	111,571,679
Site	Max Number	2185	2286
	Average Number	104.4	100.4
	Average Capacity	1,606,096	889,772
Individual	Size Average	15,379	8,965
	JPEG Average	27,386	22,137
	GIF Average	4,676	6,395
Percentage	10kByte Over	7579 (31.8%)	2728 (21.9%)
	Directly Link	437 (1.8%)	29 (0.2%)
	JPEG & GIF Percentage (Number)	J : 11222 (47.1%) G : 12589 (52.9%)	J : 2032 (16.3%) G : 10413 (83.7%)
	JPEG & GIF Percentage (Size)	J : 307,324,225 (83.9%) G : 58,865,630 (16.1%)	J : 44,982,352 (40.3%) G : 66,589,327 (59.7%)

※ A unit of size is Byte

SEED Page
Data1:(http://dir.yahoo.co.jp/Arts/Visual_Arts/Photography/Photographers/)
Data2:(http://dir.yahoo.co.jp/Arts/Visual_Arts/Illustration/Illustrators/)

4. The site-wise abstract image

4.1 Concept of the site-wise abstract image

In these days the study of a content-based image retrieval has been considerably advanced [4][5]. However, it is not suitable for searching an image site, because it aims to search an individual image. This paper proposes for making the site-wise abstract image based on LSP (Local Similarity Pattern). The site-wise abstract image is useful for a user to grasp the contents of a site.

The purpose of site-wise abstract image is to present users what are common features of images in a site. Suppose that two different sites contain the same image of flower, as shown in the top in Fig.4. When a site collects blue images as shown in the left-middle in Fig.4, the common feature of the site is color, and other features like text and shape are not common. On the other hand, when another site collects the images containing an object at center, as shown in the right-middle in Fig.4, its common features are shape and text. Here, we define the site-wise abstract image as the image in which the common features are highlighted, while other features are inconspicuous. In Fig.4, the flower image of left-middle site are presented as blue, with low resolution (left-bottom in Fig.4). On the other hand, that of right -middle site are presented as monochrome, with high resolution (right-bottom in Fig.4).

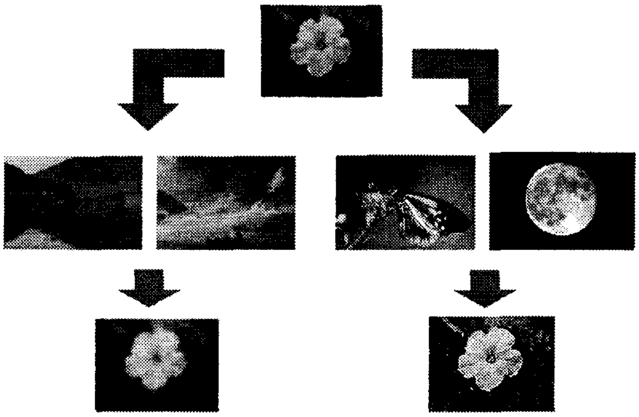

Fig.4 : Feature of images in a site

4.2 The site-wise abstract image based on LSP

In this paper, site-wise abstract image is generated with LSP. LSP is a method for computing image similarity while considering a searcher's viewpoint [6]. It divides an image into blocks, in each of which appropriate features such as color, shape and texture are applied. LSP-based similarity computation is interactively performed based on genetic algorithm.

Extracted LSP is shown in Fig.5. Fig.6 is the image which is processed by using the LSP as a filter. A processing method of an image by LSP is as follows: a block without C is displayed as monochrome because color is uncommon in the block, and a block without S or T is displayed with low resolution because shape or texture is uncommon in the block. Fig.6 shows the site-wise abstract image generated with the LSP of Fig.5. In this case an image of this site has a color feature centrally and a shape feature on the circumference.

C	T	C	T
S	C	C	T
ST	C	CS	S
T	S	ST	ST

C: COLOR

S: SHAPE

T: TEXTURE

Fig.5 : Image features computed by LSP

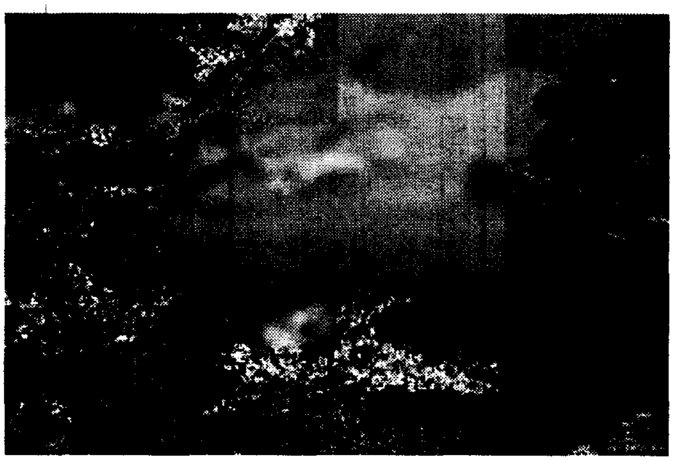

Fig.6 : The site-wise abstract image

4.3 Selection of positive images

Twenty images are selected as positive images in a site in order to use LSP for generating site-wise abstract image. We selected 20 images in the following viewpoints.

- Size-based selection: the 20 largest images in file size are chosen.
- Random-sampling: the 20 images are chosen randomly from a site.

The generated images by Size-based selection is shown in Fig.7. Those by Random-sampling is shown in Fig.8. In both figures the left images are original images, and the right images are generated site-wise abstract image.

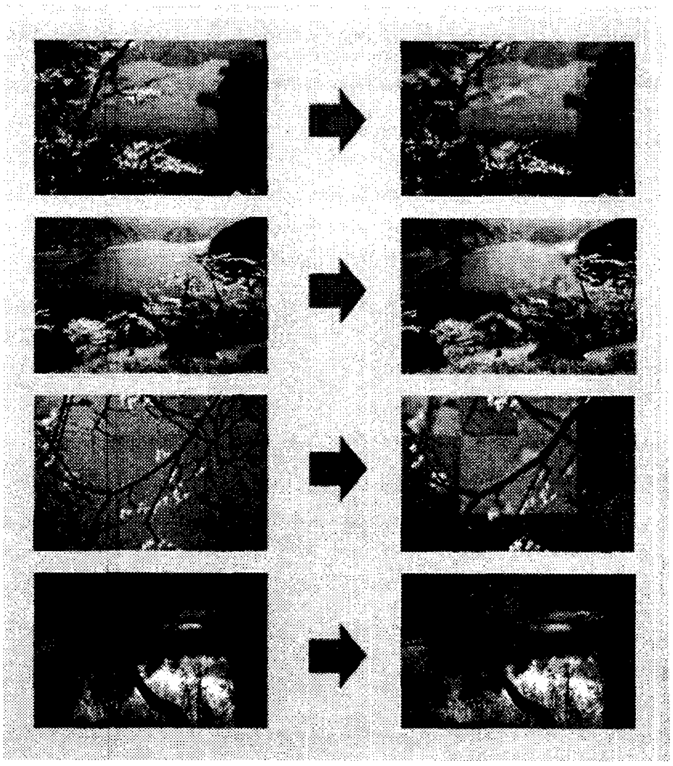

Fig.7 : Size-based selection

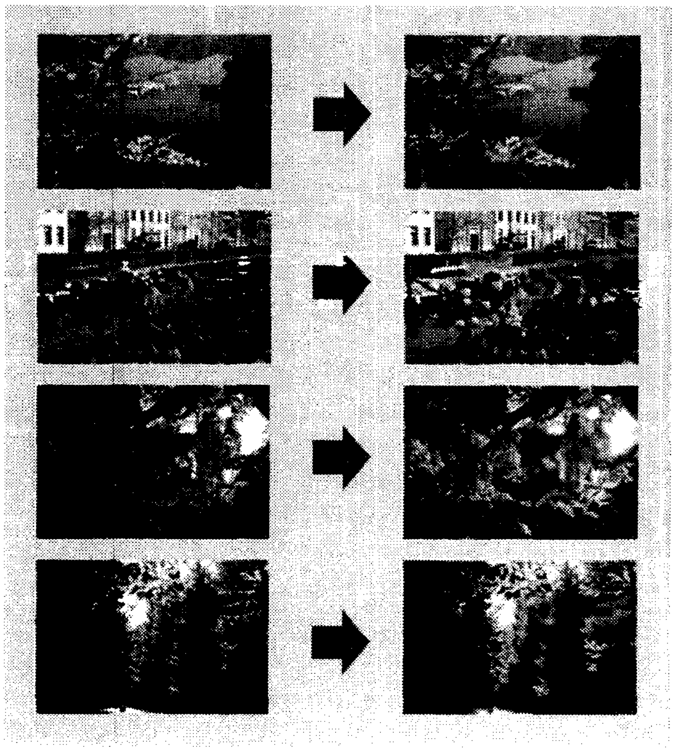

Fig.8 : Random-sampling

Preliminary experimental result shows that size-based selection is more suitable than random-sampling.

5. Visualization of site-to-site relation

It is also important to know links between sites in order to discover web image resources. Thus this paper employs a method to display links between image sites by a keyword map. We have already developed keyword map tool TMIT [7]. The distance between objects is defined based on the number of mutual links (Eq.(1)).

$$D = \log(a + b + 1) \qquad (1)$$

D : Distance between objects
a : Number of link to site B from site A
b : Number of link to site A from site B

Fig.9 is the generated keyword map from data2 in Table.1. Fig.9 displays only the sites with 100 or more images. The node shows the number of image in the site and its host name. A frame color of a site with 100-200 images is green, and that of a site with 200 or more images is red.

In Fig.9, the site "www.idea.gr.jp" is arranged on the center. This site has a search engine in addition to an illustration library. Therefore this site has many links with other sites and is supposed to be the center of this web image resource. In addition, there is a site "www.adobe.co.jp" at the lower left of the center. This is a site of a maker of image software. Because this web image resource has been collected from an illustrator category of Yahoo!, it is supposed that an object of common interest is appeared.

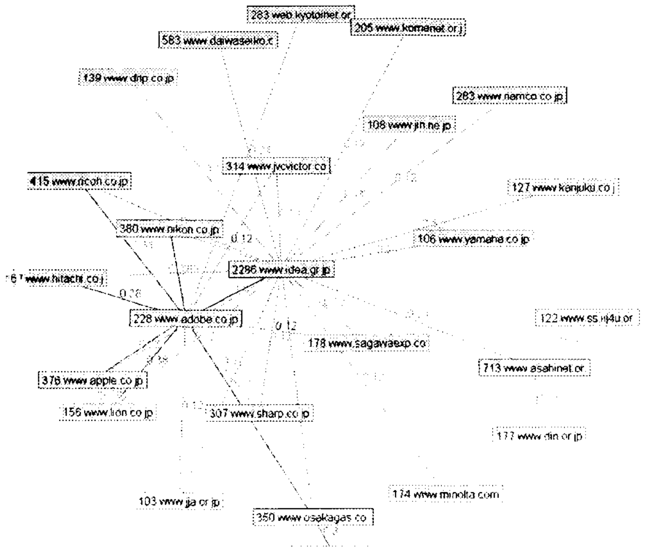

Fig.9 : Keyword map of links between image sites

6. Conclusion

This paper proposes for finding image resources on the web, in order to locate a set of interesting images efficiently from the web. The concept of site-wise abstract image is also proposed, which is useful for a user to grasp the contents of a site. As for the future study, a web image database system will be implemented based on the proposed fundamental techniques.

7. Bibliography

[1] Keisuke Shigemori, Zoran Stejic, Kaoru Hirota, Toru Yamaguchi, Yasufumi Takama, Towards the Web Image Resource Mining, The 17th Annual Conference of the Japanese Society for Artificial Intelligence, 1F1-07, 2003

[2] Keiji Yanai, An Image-gathering System from WWW Employing Keywords and Image Features, IPSJ-TOD, Vol.42, No.SIG10 (TOD11), pp.79-91, 2001

[3] Seiji Yamada, Tuyosi Murata, Yasuhiko Kitamura, Intelligent Web Information System, JSAI Journal, Vol.4, pp.495-502, 2001.

[4] Teruyoshi Washizawa, Toru Yada, Yasuhiko Yasuda, A Fast Algorithm of k-Nearest Neighbor Search in the Similarity Retrieval of Image Databases, NII Journal, NO.2, pp.27-37, 2001.

[5] Asanobu Kitamoto, Mikio Takagi, Fast Matching of Hierarchical Attributed Relational Graphs for an Application to Similarity-Based Image Retrieval, Meeting on Image Recognition and Understanding 1996 (MIRU'96), Vol.2, pp.331-336, 1996.

[6] Zoran Stejic, Yasufumi Takama, Kaoru Hirota, Genetic algorithm-based relevance feedback for image retrieval using local similarity patterns, Information Processing & Mnagement, Vol.39, pp.1-23, 2003.

[7] Yasufumi Takama, Tetsuya Hori, Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization, CIRA2003, pp.770-775, 2003.