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

Layout Analysis for Calculation of Web Page Similarity as Image

Noriaki Mitsuhashi*1 Toru Yamaguchi*1*2 Yasufumi Takama*1*2

*1 Tokyo Metropolitan Institute of Technology
*2 PRESTO, Japan Science and Technology Corporation (JST)

Address:6-6 Asahigaoka, Hino, Tokyo 191-0065

E-mail: noriaki@krectmt3.tmit.ac.jp, ytakama@cc.tmit.ac.jp

Abstract - When we search information on the Web using search engines, they only analyze the text information collected from the source files of Web pages. However, there is a limit to analyze the layout of a Web page only from its source file, although Web page design is the most important factor for a user to estimate a page. In particular, it often happens on the Web that the pages of similar design offer similar information. We propose a method to analyze layout for comparing the design of pages by treating the displayed page as image.

1 Introduction

These days anyone can easily collect information on the Web. Because of the amount of information on the Web, it is important that necessary information is found by using information retrieval system. When we search information on the Web using search engine like Google[1], it only analyzes the text information (including link information) collected from source files of Web pages. However, analyzing source files has a limit of their capability, as the important factors for human to estimate a page are not only its contents, but also its layout. In particular, it often happens on the Web that the pages of similar design offer similar information. There have been studies that analyze the layout of Web pages [3, 4], but they have only analyzed the source file, and not considered displayed images at all. We propose a method of region segmentation and layout analysis for comparing the design of Web pages, based on the visual similarity of the displayed images.

2 Comparison of Web Page

First a Web page is transformed into the image of RGB color space. RGB color space is transformed into Y image of luminance component and Cb · Cr images of chrominance components. Because the human eyes are sensitive to luminance component, detecting edge is done on Y image. Y image is divided into $n \times n$ pixel blocks, and the block average over the pixel values in each block is calculated. Edge detection uses a value that is calculated by dividing each pixel value by the block average value. A binary image is extracted from Y image using edge detection. After region segmentation is applied to the binary image, each region is classified into text or image region. Finally Web page is compared with others by using the obtained image.

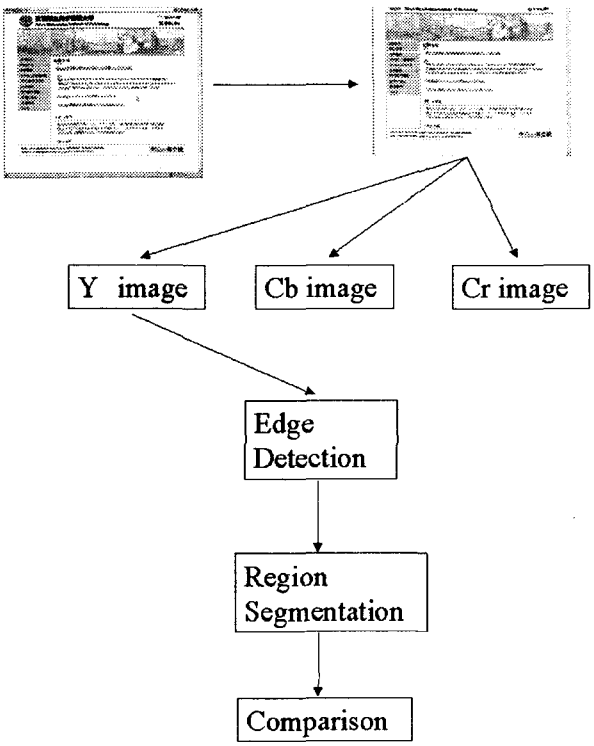

Figure 1: Comparison of Web page

3 Layout Analysis

This section describes layout analysis before a comparison of Web pages. Text and images are segmented into different regions.

3.1 Edge Detection

First, RGB color space is transformed into Y image of luminance component and Cb · Cr images of chrominance components. Image processing is applied to Y image. Y image is divided into 2×2 pixel blocks, and the block average over the pixel values in each block is calculated. Edge detection uses α_{ij} that is calculated by dividing each pixel value by the block average value.

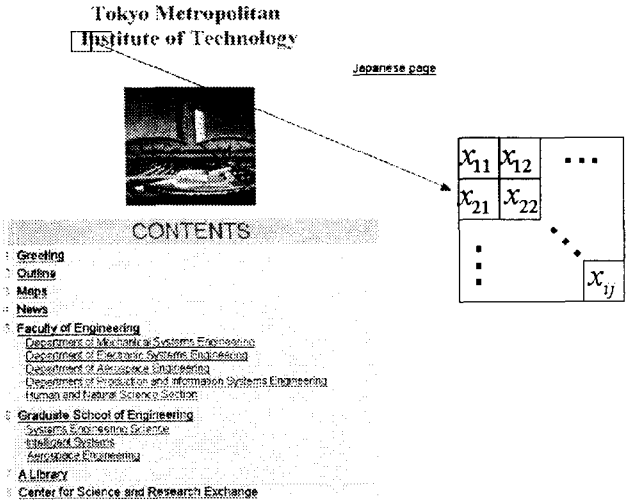

Figure 2: Edge detection

$$X = \frac{\sum_{i=1}^{n}\sum_{i=1}^{n} x_{ij}}{n \times n} \quad (1)$$

$$\alpha_{ij} = \frac{x_{ij}}{X} \quad (2)$$

where x_{ij} is pixel value in a block, and X is block average. As the value of α_{ij} tends to distribute around α_{ij}=1.0 because of the high correlation between adjacent pixels, α_{ij} represents the degree of steepness for block average value.
When $|\alpha_{ij} - 1.0|$ is above threshold, the pixel is judged as edge[2]. A simple way of edge detection is to find the part of an image, where density drastically changes. Figure3 shows the histogram of coefficient α_{ij}. The degree of steepness is distinguished by using this figure. So edge detection is used by a model such as (3).

$$\beta \leq \alpha_{ij} \quad or \quad \gamma \geq \alpha_{ij} \quad (3)$$

Where β and γ are threshold.

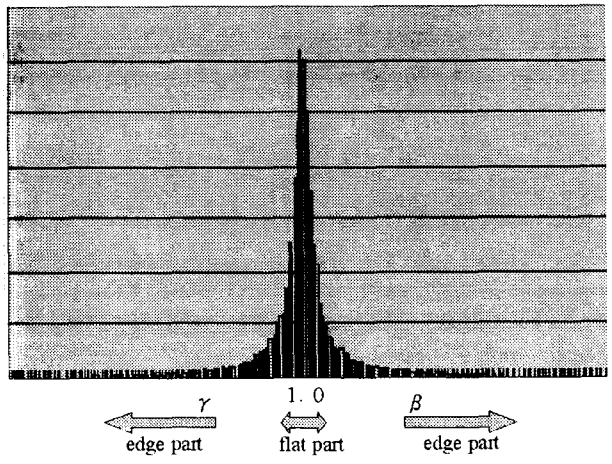

Figure 3: Histogram of coefficient α_{ij}

3.2 Region Segmentation for Web Page

Region of text and image is segmented using the binary image that is extracted by edge detection. There are text and images on a Web page. The region segmentation is performed based on the assumption that the gap between edges within a region is narrow, but that between different regions is wide. The region segmentation uses d, which is edge to edge spacing (Figure 4). When d is above threshold a, the region is segmented as shown in Figure 5. First the obtained image is searched for a raster direction, and the position where first edge part is found is recorded. Other edges are searched around the edge of that position until $d > a$. Maximum value and minimum value of location data are recorded, and a region is extracted by processing recursively. A region may be larger size than target's region because of the fake edges appeared by the change of a background color. In order to solve this problem, a region above a certain size is segmented by using a small value of threshold a in the region.

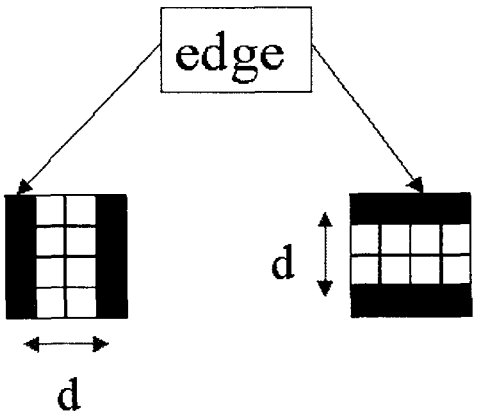

Figure 4: Edge to edge spacing

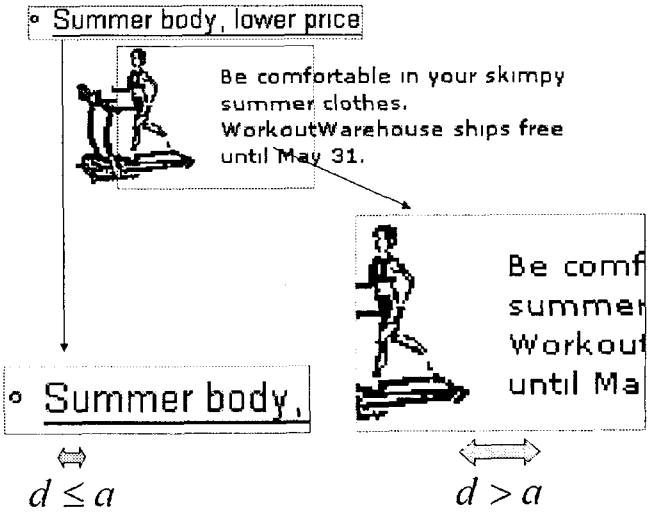

Figure 5: Region segmentation

3.3 Text and Image Distinction

In section 3.2 whether a region that is extracted by region segmentation is text region or image region is not determined. When Web pages are to be compared, distinction of text and image is important. Therefore an obtained region must be classified into text or image. We focus on the fact that text is almost horizontal writing on the Web. First each region is distinguished whether text is included in the region or not. It is observed that there is the pixel line in a region that consists of a number of edges. In Figure 6, there is a pixel line including edge in the region that contains text line. Text is distinguished in a range from the first pixel line to the last pixel line of an existing edge. Figure 7 shows that there exists edges at a uniform distance. When there exist edges at a uniform distance, the corresponding region is decided as text area. Image area contains edges at random. Then each region is decided whether it includes text or image based on the existence of such areas.

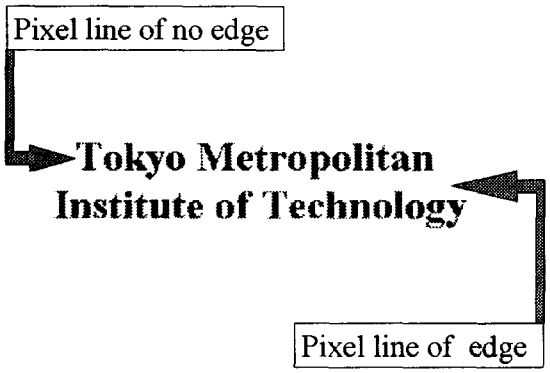

Figure 6: Distinction of text line

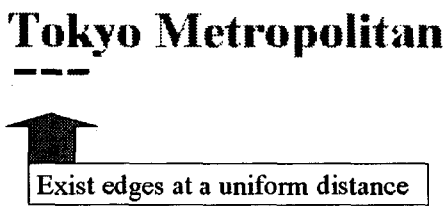

Figure 7: Distinction of character

4 Performed Based on The Web Page Comparison

Web page comparison is performed on the layout of pages extracted by region segmentation. The obtained binary image (i.e. edge image) is segmented into 3 regions; text region, image region, and those mixture. Each region is weighted with positive integer; text region is 1, image region is 2, those mixture is 3, otherwise is 0. These weights reflect the importance of regions. The similarity between two Web pages is measured using S_w, which describes the cosine value between page layout vectors w_i and w_j. Each dimension of a vector corresponds to a pixel in an image. A pixel is assigned the weight of the corresponding region. Figure 9 shows the layout of Web page, which is generated from Fig.8.

$$S_w = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\| \mathbf{w_i} \| \cdot \| \mathbf{w_j} \|} \quad (\in [0,1]) \tag{4}$$

Where $w_i \cdot w_j$ is inner product, and $\| \mathbf{w_i} \|$ is magnitude of vector.

Figure 8: Original image

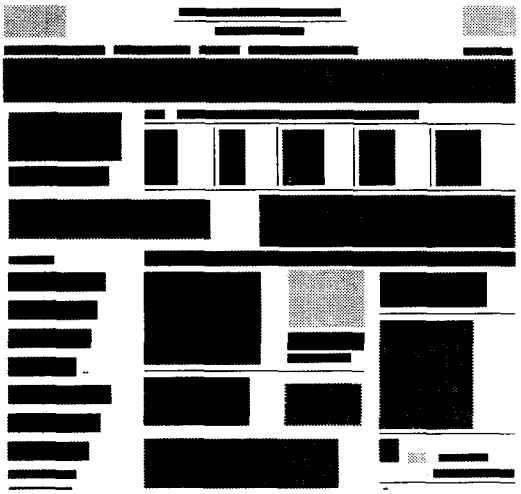

Figure 9: Image of Web page layout

5 Experiment Results

Experiments are performed to show the effect of region segmentation and Web page comparison. Figure 8 is the original image of a Web page. Figure 10 and Figure 11 are Processed image. It is said from Fig.8 and Fig.11 that region segmentation is almost correct. Table 1 shows the similarities between the page in Fig.8 and the top pages of popular sites. Table 2 shows the results of questionnaires, in which 7 subjects are asked to judge the visual similarities. Fig.8 shows the top page

of Website goo(http://www.goo.co.jp) and it is assumed that it looks like top pages of other search engines, such as yahoo! and infoseek. Both the results of the proposed method and the judgment by subjects show the assumption is valid.

Figure 10: Edge detection

Figure 11: Processed image

Table 1: Similarity by proposed method

page name	similarity
yahoo!	0.522
infoseek	0.503
onkyo	0.300
daihatsu	0.299

Table 2: Visual judgment

page name	visual judgment		
	looks similar	otherwise	looks different
yahoo!	6/7[subjects]	1/7	0/7
infoseek	7/7	0/7	0/7
onkyo	0/7	3/7	4/7
daihatsu	0/7	2/7	5/7

6 Conclusion

In this paper, we proposed a method for analyzing layout of Web pages in order to calculate their visual similarities. The proposed method is able to segment a Web page into text or image region and compare the layout of pages by treating the displayed pages as image. The proposed method is applied to actual Web pages, such as the top pages of search engines, and it can judge the visual similarity among pages like human subjects. The proposed method is promising as a basis for various applications, such as finding top pages and evaluating Web accessibility.

References

[1] Sergey Brin, Lawrence Page "The Anatomy of a Large-Scale Hypertextual Web Search Engine" Computer Networks and ISDN Systems, vol.30, no.1-7, pp.107-117, 1998.

[2] Noriaki Mitsuhashi, Toru Yamaguchi, Yasufumi Takama "Finding Web Page Based on Web Page Similarity as Image" The 17th Annual Conference of the Japanese Society for Artificial Intelligence, 1C4-01, 2003

[3] Vojtech Svatek, Jiri Braza, Vilem Sklenak "Towards Trple-Based Information Extraction from Visually-Structured HTML Pages" The Twelfth International World Wide Web Conference, 2003

[4] Shipeng Yu, Deng Cai, Ji-Rong Wen, Wei-Ying Ma "Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation" The Twelfth International World Wide Web Conference,2003