Accuracy Enhancement of Function-Oriented Web Image Classification

Koji Nakahira
Department of Frontier Informatics,
The University of Tokyo
5-1-5 Kashiwano-ha, Kashiwa-shi,
Chiba, 277-8561, Japan
nakahira@hal.k.u-tokyo.ac.jp

Toshihiko Yamasaki
Department of Frontier Informatics,
The University of Tokyo
5-1-5 Kashiwano-ha, Kashiwa-shi,
Chiba, 277-8561, Japan
yamasaki@hal.k.u-tokyo.ac.jp

Kiyoharu Aizawa
Department of Frontier Informatics,
The University of Tokyo
5-1-5 Kashiwano-ha, Kashiwa-shi,
Chiba, 277-8561, Japan
aizawa@hal.k.u-tokyo.ac.jp

ABSTRACT
We propose a function-oriented classification of web images and show new applications using this categorization. We defined nine categories of images taking into account of their functions used in web pages, and classified web images by using Support Vector Machine (SVM) in tree structured way. In order to achieve high accuracy of classification, we employed two kinds of features, image-based features and text-based features, and the two kinds can be used together or separately for the stages of the classification. We also utilized DCT coefficients to classify photo images and illustrations. As a result, accurate classification has been achieved. Finally, we show the page summarization as a new application that is made feasible for the first time by our new categories of WWW images.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *abstracting methods, indexing methods.*

General Terms
Theory

Keywords
Web images, Classification, Support Vector Machine.

1. INTRODUCTION
The number of images utilized in web pages has been increasing drastically. The images on web pages do not only convey the content of the images themselves but also provide important visual structure of the web pages.

In most of the web services, however, images are discarded and only text data are analyzed. Although image features were employed in some researches to improve image search accuracy [1], main parts of their algorithms were still based on text data analysis.

On the other hand, we have proposed web image classification [2, 3] based on the function of the images using SVM and demonstrated some new applications such as summarization of web pages and categorization of web pages. In our previous work, the discrimination accuracy of photos and illustrations was guide poor. Therefore, the purpose of this paper is to enhance the accuracy of the classification by introducing frequency analysis using discrete cosine transform (DCT) [4]. As a result, much better performance has been demonstrated.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. CATEGORIZATION OF WEB IMAGES

2.1 Categories Based on the Functions
We have defined nine function-based categories by analyzing a number of web sites. Namely, main-titles, section-titles, photos, illustrations, icons, list of pages, logos, advertisements, and segmented-Images. Examples of some categories are shown in Figure 1. It has been verified that the images in other web sites can be classified to the categories..

2.2 Classification Algorithm
For the classification, we employ SVM [5] and we introduced two kinds of features of the images: image-based features and text-based features. Image features consist of image size, the number of pixels, aspect ratio, the number of colors, color histograms of Cr and Cb, and file type. On the other hand, location of the image in the page, length of comments added to the image, existence of a link, existence of the same image in the same page, and so on, were used as text-based features.

Web images are classified according to the classification tree shown in Figure 2. One of the two features or both were selected at each stage of the classification tree to achieve better performance by empirical study.

However, in our previous work, it has been revealed that photos and illustrations cannot be separated well by these two features. Therefore, we introduced DCT coefficients as new features. We divided the images into 8x8 blocks and calculated the mean of the absolute DCT coefficients. Then, average DCT coefficients matrices for photos and illustrations were obtained and regularized. The differences of the regularized coefficients are illustrated in Figure 3. (a). The darker the brightness is, the larger the difference is. For efficient computation, we selected the most significant 20 coefficient indices as shown in Figure 3. (b).

2.3 Experimental Results of Classification
2,500 images randomly selected from 130 company web sites were utilized as a training data set. Then, 1,500 images that were not contained in the training set were evaluated. We classified the images into the nine categories as listed section 2.1 using the SVM with an RBF kernel. The experimental results are shown in Table 1. In order to evaluate the total performance of the classification, the F-measure is also introduced. The F-measure is calculated by the following equation:

$$F-measure = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$$

In our previous work [2], the average of F-measure of all categories was 0.339. By employing two kinds of features, new parameters, and regularization of the parameters of features, the average of F-measure came to 0.712. Particularly, the DCT contributed very well to discriminating between photos and illustrations. By utilizing the DCT, the F-measure of photos, and illustrations went up from 0.684 and 0.358 to 0.889 and 0.450.

3. SUMMARIZATION OF WEB PAGES

New applications have been made possible by using the image classification results based on their roles and functions. In this paper, the web summarization is shown as an example.

A section-title image is a symbol image of a section and shows the section content. Therefore, web pages can be summarized by using section-title images. Figure 4 shows experimental results. When a page has section-title images, showing only them and texts around these images will produce a summarization of the page. The most important point here is that our approach is language-independent, because it is based on images.

4. CONCLUSION

In this paper, we proposed web image classification based on the roles and functions of the images. By employing both image-based features and text-based features, pretty accurate performance have been achieved. In addition, we introduced DCT as new features. As a result, much better performance in classification of photo images and illustrations has been achieved as compared to our previous work. Based on the classification results, we have demonstrated language-independent summarization of web pages.

5. REFERENCES

[1] John R. Smith and Shin-Fu Chang, "Visually searching the Web for content," IEEE Multimedia, 4(3):12-20, 1997.

[2] K.Nakahira, S.Ueno, and K.Aizawa, "Automatic Categorization WWW Images with Applications for Retrieval Navigation," in Proceeding of 5th Pacific-Rim Conference on Multimedia, pp. 174-181, Springer, 2004.

[3] K.Nakahira, T.Yamasaki, and K.Aizawa, "Classification of WWW Images based on Their Functions," submitted to 2005 International Conference on Multimedia & Expo.

[4] K.R.Rao, and P.Yip, "Discrete cosine transform: algorithms, advantages, applications," Academic Press Professional, Inc., San Diego, CA, 1990.

[5] C.Cortes and V.Vapnik, "Support Vector Networks," Machine Learning, Vol. 20, pp. 273–297, 1995.

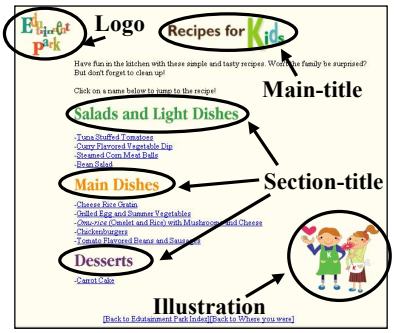


Figure 1. Examples of WWW image definition.

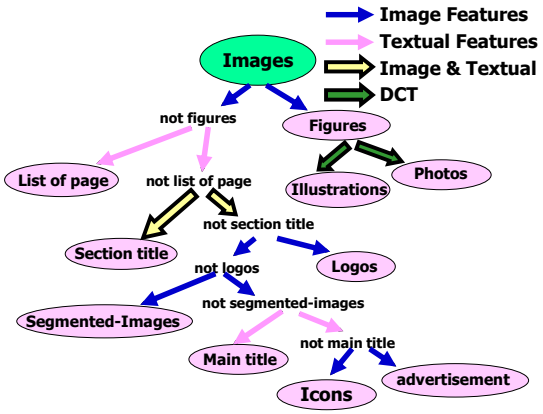
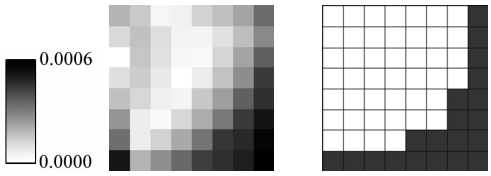


Figure 2. Classification tree. Best feature for each stage of classification is also shown.



(a) (b)

Figure 3. DCT coefficient analysis of photo and illustration: (a) difference between matrix elements of photo and illustration; (b) most significant 20 indices used as features.

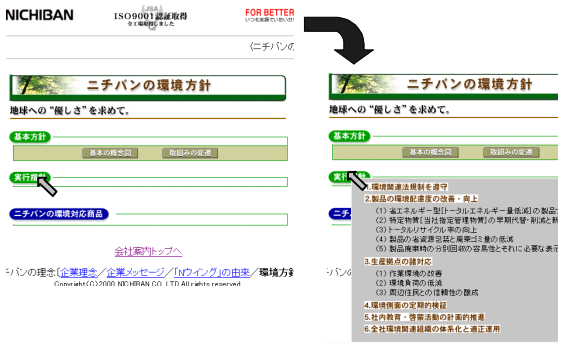


Figure 4. Result of page summarization.

Table 1. Experimental results of web image classification.

	Precision	Recall	F-measure
Figure	0.853	0.935	0.892
Figure (photo)	0.893	0.884	0.889
Figure (illustration)	0.439	0.462	0.450
List of Page	0.728	0.776	0.751
Section title	0.500	0.615	0.552
Segmented	0.697	0.528	0.601
Main title	0.351	0.191	0.248
Logo	0.114	0.116	0.115
Icon	0.292	0.111	0.161
Advertisement	0.000	0.000	0.000