Are Web Pages Characterized by Color?

Norifumi Murayama
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuda-cho,
Midori-ku, Yokohama, JAPAN
murayama@lr.pi.titech.ac.jp

Suguru Saito
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda-cho,
Midori-ku, Yokohama, JAPAN
suguru@pi.titech.ac.jp

Manabu Okumura
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda-cho,
Midori-ku, Yokohama, JAPAN
oku@pi.titech.ac.jp

ABSTRACT

When human guess the content of a web page, not only the text on the page but also its appearance is an important factor. However, there have been few studies on the relationship between the content and visual appearance of a web page. We investigating the tendency between them, especially web content and color use, we found a tendency to use color for some kinds of content pages. We think this result opens the way to estimating web content using color information.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence; I.2.0 [**Artificial Intelligence**]: General; H.1.m [**Information Systems**]: Miscellaneous

General Terms

Measurement, Experimentation

Keywords

color, contents of web page

1. INTRODUCTION

When we guess the content of a web page, the page's visual appearance gives us a clue. However, most studies on automatically estimating web page content used only textual information and link structures with a few exception.

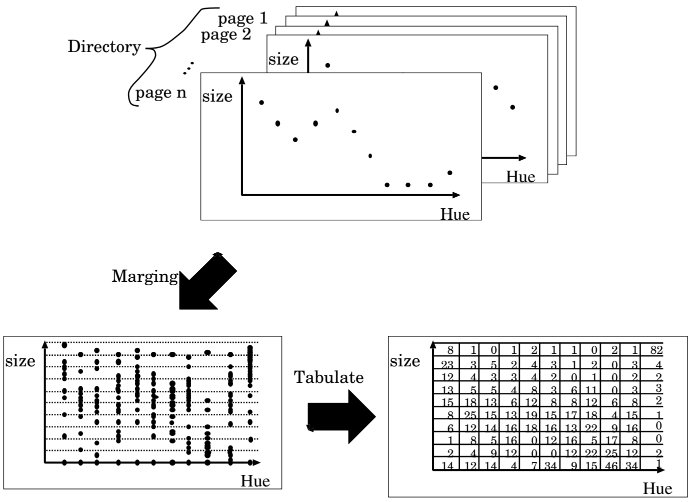
For example, Jianying and Amit[2] proposed a method of categorizing images in web pages. James and Jia[3] filtered out adult sites by using the skin color of photographic images contained in websites.

Those methods focused only on images in web pages. If there are an general method of dealing with visual information on the whole web page, then it would improves the accuracy of guessing web page content.

In this paper, we focus on color use as one measure of page appearance. We explore the relationship between webpage content and color use.

In this research, we used an existing database constructed as a directory structure based on page content categories. For those categories, we researched the color tendencies.

WWW2004, May 17–22, 2004, New York, NY USA.
ACM xxx.xxx.



8	1	0	1	2	1	1	0	2	1	82
23	3	5	2	4	3	1	2	0	3	4
12	4	3	3	4	2	0	1	0	2	2
13	5	5	4	8	3	6	11	0	3	3
15	18	13	6	12	8	8	12	6	8	2
8	25	15	13	19	15	17	18	4	15	1
6	12	14	16	18	16	13	22	9	16	0
1	8	5	16	0	12	16	5	17	8	0
2	4	9	12	0	0	12	22	25	12	2
14	12	14	4	7	34	9	15	46	34	1

Figure 1: computation of color information

2. WEB PAGE SAMPLING

2.1 Selection of target web pages

The database that we selected is a collection of web pages by Open Directory Project (ODP) [1]. It has a standard directory structure for representing a hierarchy of categories and has links to connect subdirectories of similar meaning. In this study, we used only the directory structure to simplify our process.

The number of web pages registered in ODP is about three million. Accessing and capturing every pages waste a lot of time. Therefore, we selected samples from the database by a rule that subdirectories in a certain lower level are ignored randomly.

Consequently, the number of web pages, that we actually access to and capturinged is decrease to 116,403.

2.2 Capturing and cleansing the data

The number of target pages is still large, so we capture rendered web pages automatically. However, the captured pages includes obviously unsuitable ones. For example, an image where the whole area is the default background color of the browser. This happens when a response from a server is too late and nothing is rendered.

Therefore, we apply post-processing to remove these inappropriate captured images.

As the result, we had 62,052 pages as captured images.

Table 1: Hue histogram for the whole dataset

Hue	G	↔	B	↔	P	↔	R	↔	Y	↩	AC
large	1	0	2	3	0	0	1	3	2	0	80
	0	0	3	4	0	0	1	2	2	0	6
↑	1	1	5	8	0	1	5	5	3	0	5
	1	1	7	13	1	4	11	10	5	1	4
size	2	2	8	16	3	11	17	14	7	2	1
	3	4	9	14	4	13	17	14	9	3	0
	5	5	9	9	6	14	12	11	10	5	0
↓	5	6	8	6	8	11	7	8	9	6	0
	6	7	6	3	8	8	4	5	7	6	0
small	71	69	38	17	64	32	20	24	41	72	0

3. RESEARCH METHOD

3.1 Color information

Before analysis, we translate captured pages into histograms. The process flow is shown in Figure 1. First, we make histograms of hue, chroma, and brightness from one captured image. In the histogram, the vertical axis is a log scale. Three types of histograms are used separately in the following steps. Second, we merge histograms of one type for one subdirectory, and make one two dimensional histogram. This histogram shows the relationship between area size and one of hue, chroma or brightness in one content category.

The reason we use hue, chroma, and brightness instead of RGB value is that those values are more appropriate to human color perception. They are calculated from CIELAB, which is converted from sRGB values. Note that in the hue histogram we include an achromatic field.

Furthermore, the histogram is represented by a table where the that sum of each column is normalized. We performed this calculation for all subdirectories and whole the dataset.

The hue/size histogram for the whole dataset is shown in Table1[1]. It shows the tendency of hue and size in all web pages.

3.2 Tendency analysis

To analyze the color tendency of a subdirectory, we subtract values in the table for the whole dataset from corresponding values in the subdirectory table. In the subtracted table, the color of a cell where the absolute value is large is a characteristic for the subdirectory.

Because the sum of each column is Table 1 is normalized, values can be regarded as the probability in the column. By using the probability equation of a multinomial distribution with Table 1, we calculate P_D, which is the appearance probability of each column in a subdirectory table.

From the value of $1/P_D$, we measure the strength of the tendency of color used in a subdirectory.

4. RESULTS

One table obtained for a directory "*Home/Gardens*" is shown in Table 2. Since the values of $1/P_D$ for yellow and yellow-green are large, we can recognize a characteristic in the hue for this category. Furthermore, by looking at the table carefully, we can determine that the area size for such a hue grows from the general tendency.

Table 3 is the hue/size table of "*Adult/Image_Galleries*", which are observed to have the strongest tendency about hue. From this, we can see that red, which includes skin color, is used more than for other subdirectories.

Table 2: Result of hue of *Home/Gardens*

Hue	G	↔	B	↔	P	↔	R	↔	Y	↩	AC
$1/P_D$	30	40	35	48	51	56	30	39	109	189	70
large	0	0	-2	-2	0	0	1	1	**3**	0	**-3**
	0	0	-2	**-4**	0	0	-1	**3**	**6**	0	0
↑	-1	-1	**-4**	**-5**	0	1	-1	**3**	**9**	1	1
	1	-1	**-3**	**-5**	0	2	-2	**4**	**13**	**3**	2
size	2	**3**	-2	1	0	-1	-1	-2	**5**	**11**	0
	2	-2	0	0	2	0	**-3**	**-4**	**6**	**7**	0
	3	-2	0	**3**	0	1	**4**	-2	**-5**	**3**	0
↓	-2	**-4**	0	**3**	1	**-3**	1	**3**	**-4**	**3**	0
	-2	-2	1	1	1	0	1	1	-2	1	0
small	-2	**9**	**12**	**12**	-2	2	2	**-8**	**-29**	**-30**	0

Table 3: Result of hue of *Adult/Image_Galleries*

Hue	G	↔	B	↔	P	↔	R	↔	Y	↩	AC
$1/P_D$	40	27	72	54	80	103	217	70	39	38	83
large	-1	0	**5**	**5**	**3**	0	**3**	0	0	0	**-18**
	0	0	-1	-1	0	0	1	-1	-1	0	**14**
↑	-1	-1	-2	**-4**	0	2	**14**	2	-2	0	**5**
	0	0	**-5**	**-5**	**4**	**7**	**17**	**5**	-2	-1	0
size	-1	0	-2	-2	1	**8**	2	**12**	-2	-1	0
	0	0	-2	0	**3**	**7**	**-10**	0	-1	-2	0
	-2	-2	2	0	2	1	**-6**	**-3**	1	-1	0
↓	-1	-1	**-3**	1	1	**-4**	**-5**	-2	0	-2	0
	-3	-1	1	**3**	0	**-5**	**-3**	-1	**3**	-2	0
small	**11**	**6**	**8**	**6**	**-14**	**-16**	**-12**	**-10**	**5**	**11**	0

Table 4: Result of chroma of *Kids and Teens*

Chroma	small	←						→	large	
$1/P_D$	66	35	30	45	69	62	109	77	64	27
large	**-15**	0	2	0	1	1	2	0	0	0
	5	0	0	**3**	2	1	2	0	0	0
↑	**3**	-1	1	1	2	**3**	**4**	2	0	0
	4	**-3**	-2	1	**5**	**3**	**6**	**4**	0	1
size	2	-1	1	2	**3**	**3**	2	**4**	2	1
	0	**4**	1	0	2	1	2	**3**	2	-1
	0	1	-1	-1	-2	0	1	2	2	1
↓	0	2	1	-1	**-3**	-1	-2	1	**4**	2
	0	-1	-1	-1	**-3**	**-3**	-1	0	2	0
small	0	-1	-1	**-4**	**-7**	**-10**	**-15**	**-16**	**-14**	**-5**

In the same way, Table 4 indicates the tendency to use high chromatic color in the "*Kids and Teens*" category.

We examined 386 subdirectories having more than 100 web pages. In total, we found color tendencies in 139 categories, with the threshold value of $1/P_D$, 50.

However, in the case of abstract categories like "*Business*", we could not find any significant tendencies.

5. CONCLUSION

These results show that there is a tendency of color use in some categories of web pages. This suggests the possibility of using color information for estimating web page.

6. REFERENCES

[1] Open directory project. In *http://dmoz.org/*. Netscape, 1998-2003.

[2] J. Hu and A. Bagga. Functionality-based web image categorization. WWW2003, 2003.

[3] J. Z. Wang and J. Li. Classifying objectionable websites based on image content. International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services, 1998.

[1] G:Green, B:Blue, P:Purple, R:Red, Y:Yellow and AC:Achromatic Color