Seeing and Reading Red: Hue and Color-word Correlation in Images and Attendant Text on the WWW

Shawn Newsam
School of Engineering
University of California at Merced
Merced, CA 95340
snewsam@ucmerced.edu

ABSTRACT

This work represents an initial investigation into determining whether correlations actually exist between metadata and content descriptors in multimedia datasets. We provide a quantitative method for evaluating whether the hue of images on the WWW is correlated with the occurrence of color-words in metadata such as URLs, image names, and attendant text. It turns out that such a correlation does exist: the likelihood that a particular color appears in an image whose URL, name, and/or attendant text contains the corresponding color-word is generally at least twice the likelihood that the color appears in a randomly chosen image on the WWW. While this finding might not be significant in and of itself, it represents an initial step towards quantitatively establishing that other, perhaps more useful correlations exist. These correlations form the basis for exciting novel approaches that leverage semi-supervised datasets, such as the WWW, to overcome the semantic gap that has hampered progress in multimedia information retrieval for some time now.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: clustering; information filtering; retrieval models.

General Terms

Algorithms.

Keywords

Multimedia and metadata data mining, semi-supervised learning, image content descriptors.

1. INTRODUCTION

The impasse presented by the proverbial *semantic gap* has hampered progress in multimedia information retrieval over the past several years. The early successes that promised access to multimedia data based solely on its content without manual annotation have failed to develop into useable systems. Content descriptors that can be automatically extracted from images, such as color and texture, provide limited high-level information and usually only in highly constrained situations. Focus needs to be shifted away from developing new content descriptors to investigating novel ways in which metadata that is available without manual annotation can aid multimedia information retrieval. This work stipulates that one promising approach to addressing the semantic gap is to leverage the semi-supervised multimedia datasets that are appearing in our information society.

Early investigations into leveraging semi-supervised datasets are encouraging. Examples include:

- Searching for images on the world wide web (WWW) using image URLs, image names, image ALT tags and attendant text on the webpage containing the images. This includes commercial systems such as Google's Image Search [4] and Yahoo's Image Search [5].
- Searching for images on the WWW using a combination of textual information and content descriptors [10].
- Searching for images on the WWW using category structure, textual information, and content descriptors [8][1][9][2].
- Using annotated stock photography or art collections to learn statistical word-descriptor correlations to perform auto-annotation or auto-illustration [6][7].

These investigations suggest that using available metadata, such as URLs, annotations, etc., instead-of or in-addition-to content descriptors results in better performance than solely relying on content descriptors alone. This, in turn, suggests that the metadata and image content are correlated, but improved performance is only anecdotal evidence of such a correlation. The work presented in this paper represents initial investigations into whether such correlations actually exist. We focus on the particular question of whether the colors or hues of images on the WWW are correlated with the occurrence of color-words, such as red or blue, in the URLs, image names, and attendant text corresponding to the images. The existence of such a correlation might not have direct application, such as searching for images on the

MDM/KDD 2005 Chicago, August 21, Chicago, Illinois, USA

WWW, but serves to *provide initial insight into the nature of other correlations that might prove more useful*.

2. THE APPROACH

In summary, the objective is to estimate the likelihood that a particular color appears in an image on the WWW given that the corresponding color-word occurs in the metadata associated with the image. This metadata includes the image URL, the image name, and the text, if any, that appears in attendance with the image on the webpage. This section describes the steps taken towards this end. First, nearly 2 million images and their metadata are downloaded from the WWW. Next, hue histograms are computed for each image. Finally, the conditional hue probabilities are estimated by averaging the hue histograms over subsets of images corresponding to different hypotheses, such as the word red occurring in the attendant text. The conditional color likelihoods are then computed based on these conditional probabilities.

2.1 Acquiring the Images and Attendant Text

The first step is to acquire a sizable set of images and their metadata from the WWW. This data was generously provided by Till Quack from the Cortina content-based web retrieval dataset [2][9]. The Cortina dataset was collected using WWW Uniform Resource Locators (URLs) from the DMOZ open directory project [3]. The entire DMOZ directory can be downloaded as a Resource Description Framework file which can be parsed for URLs. Only URLs from the Shopping and Recreation categories were used in creating the Cortina dataset. These URLs were used to locate webpages containing images. The URL for each image was then stored in a database along with other key information extracted from the webpage. This includes up to 20 words above and 30 words below the image. These word sets constitute the attendant text utilized in this paper.

The Cortina dataset also includes two color descriptors for each image. We decided, however, that these descriptors were not appropriate for our analysis and so we downloaded each image again using the URLs in the Cortina dataset to extract our own color descriptor termed *hue histograms*. Only 1,751,578 of the 2 million image URLs obtained from the Cortina dataset were still valid. Of these, 933,875 were located on webpages that contained attendant text.

2.2 Hue Histograms

Our analysis required a compact characterization of the color distribution of each image. Rather than use a traditional multi-dimensional histogram, we derived a one-dimensional histogram of just the hue values of the pixels in an image. The pixel values are first transformed from the red-green-blue (RGB) colorspace to a hue-lightness-saturation (HLS) colorspace. This hue channel is similar to an angle on the color-wheel in that its range from 0 to 360 corresponds to the colors red, orange, yellow, etc., through pink, and then back to red again. The hue histograms are computed by binning the hue channel into 360 one-degree intervals.

The HLS colorspace presented some expected problems. First, even very dark (low lightness value) or very light (high lightness value) colors have associated hue values, the colors corresponding to the hue values are not really perceivable. To deal with this we added two additional histogram bins, one for pixels with lightness values below an empirically chosen threshold and another for pixels with lightness values above an empirically chosen threshold. These bins correspond to the "colors" black and white, respectively. The second problem is that grey-ish colors also have an associated hue. This was dealt with by adding a third histogram bin for pixels with saturation values below another empirically chosen threshold.

color	central hue value	hue range
red	0	335-20
orange	30	10-50
yellow	60	40-90
green	120	85-185
blue	240	175-275
purple	300	265-320
pink	330	310-350

Table 1: The seven colors and their central hue values and hue ranges.

The complete hue histograms contain 363 bins. Three for black, white, and grey, and 360 for the hue values at one-degree intervals. Table 1 indicates the central hue values for the seven colors analyzed: red, orange, yellow, green, blue, purple, and pink. It also indicates the range of hues for each color. The ranges for adjacent colors overlap by 10 degrees to account for the ambiguous regions between colors.

Finally, the hue histograms are normalized so that they sum to one. This accounts for the different image sizes and makes the histograms interpretable as estimates of probability density functions (PDFs).

2.3 Conditional Hue PDFs

The conditional hue PDFs are estimated by averaging the hue histograms corresponding to a particular hypothesis. For example, if $hist_i(h)$ is the value of the hue histogram for image i at hue h then the conditional hue PDF for the hypothesis of "all images whose attendant text contains the word red," is estimated as:

$$p(h|red \in text) = \frac{1}{\#(i : red \in text)} \sum_{i:red \in text} hist_i(h) \ . \quad (1)$$

2.4 Conditional Color Likelihoods

Finally, the conditional hue PDFs, such as $p(h|red \in text)$, are used to compute the conditional likelihoods that colors occur in sets of images satisfying hypotheses. For example, the likelihood that the color blue occurs in images whose attendant text contains the color-word blue is computed as:

$$p(blue \in image|blue \in text) = \sum_{h=175}^{275} p(h|blue \in text) \ . \quad (2)$$

The bounds on the summation for each of the seven color considered are listed in Table 1. These conditional color likelihoods can be used to compare different hypotheses, such as color-words appearing in attendant text, URLs, image names, etc., with each other as well as with baseline hypotheses such as the color blue occurring in any image on

the WWW which can be computed as:

$$p(blue \in image | all\ images) = \sum_{h=175}^{275} p(h | all\ images)\ . \quad (3)$$

3. EXPERIMENTS AND RESULTS

The conditional hue PDFs and conditional color likelihoods are computed for a variety of hypotheses as well as the baseline hypothesis. This section describes these hypotheses and their results.

3.1 Hypotheses

The conditional hue PDFs and conditional color likelihoods are computed for a total of eight hypotheses:

all This is the baseline of all 1,751,578 images.

all with text The 933,875 images with attendant text (many images appear alone on webpages).

color-word ∈ text A color-word occurs in the attendant text for an image.

color-word ∈ URL A color-word occurs in the URL for an image.

color-word ∈ image name A color-word occurs in the image name. The image name is considered as the substring in the URL after the rightmost backslash.

color-word ∈ text,URL Conjunction of two hypotheses above.

color-word ∈ text,image name Conjunction of two hypotheses above.

only color-word ∈ text A color-word occurs in the attendant text for an image without any other color-words.

The color-words considered are red, orange, yellow, green, blue, purple, and pink. The meaning of occurrence varies by hypothesis above. In the case of attendant text, the color-word must occur as a separate word; that is, it must be preceded and followed by a non-alphabetic character. This constraint is not enforced for occurrence in a URL or image name. Thus, the color-word red is considered to occur in the URL http://threddies.com/images/tinysageside.jpg and the image name redruff_computer.jpg. This likely introduces noise but considerably more complex string matching algorithms would be required to identify only those URLs and image names that contain the color-word red as a "separate word" as in the attendant text.

3.2 Conditional Hue PDF Results

The conditional hue PDFs are shown for several hypothesis in Figure 1. Note that the PDFs are plotted only for hue values corresponding to the 360 color-bins in the hue histograms. The three bins corresponding to black, white and grey have been left out since their magnitudes are generally much larger and their values are not informative for our analysis. Thus, the PDFs as plotted do not necessarily sum to one. Note, also, that the PDFs have been smoothed by averaging the values over five-degree intervals.

Figures 1(a) through 1(f) show the PDF for the proposed hypothesis with a solid line, and the PDF for the baseline hypothesis (all images) with a dashed line for comparison.

The first thing to note from Figure 1 is that the baseline PDF is not uniform. There appear to be distinct peaks around the red-orange and green-blue regions. Second is that there is approximately only a 30% probability that a pixel is not black, white, or grey. The images corresponding to the baseline hypothesis are approximately 16% black, 30% white, and 25% grey.

The PDFs in Figure 1 all correspond to hypotheses relating to the color-word red. Thus, it is significant that they all exhibit the following noteworthy characteristics. First, they are all greater than the baseline PDF for hue values corresponding to red. Second, they are generally all less than the baseline PDF for other hue values. This indicates that there is a correlation between the color-word red occurring in the attendant text, URL and/or image name, and the color red appearing in the image.

The conditional hue probabilities for the other color are similar to those for red but space restrictions prevent them from being included.

3.3 Conditional Color Likelihood Results

The conditional color likelihoods allow a more quantitative comparison of the hypotheses. Table 2 shows the conditional color likelihoods for a number of hypotheses. The rows represent the hypotheses, such as the color-word red occurring in both the attendant text and URL (red ∈ text,URL) for an image, and the columns represent the color likelihood conditioned on the hypothesis. The maximum for each column is shown in bold. There is also a column indicating the number of images that satisfy the hypothesis.

The results in Table 2 again demonstrate there is a correlation between a color-word occurring in attendant text, URL and/or image name, and the color appearing in an image. The correlations are strongest for the color-word occurring in the image name but the correlations for the other hypotheses are not significantly weaker. In general, the likelihood that a color appears in an image given the corresponding color-word occurs in the attendant text, URL and/or image name is at least twice the baseline likelihood.

The following specific observations can be made from Table 2:

- The hypotheses for the color-word appearing in the image name maximizes the conditional color likelihoods of all hypothesis for orange, yellow, green, blue, and purple, and there is the correct correspondence between the color-word and color; i.e., p(orange|orange ∈ image name) is maximal.
- The hypothesis for the color-word red appearing in both the text and image name maximizes the likelihood that red appears in the image.
- The hypothesis for the color-word pink appearing in both the text and URL maximizes the likelihood that

pink appears in the image.

- In all cases but two, the conditional color likelihood for a hypothesis is maximum for the correct color. I.e., p(red|red ∈ text) is greater than p(red|color-word ∈ text) for any other color-word. The exceptions are p(red|color-word ∈ URL) and p(red|color-word ∈ image name).

- The conditional likelihoods "bleed" into adjacent colors. Not only are the conditional likelihoods corresponding to the correct color-words large but so are the conditional likelihoods for adjacent colors on the hue axis.

4. DISCUSSION

This work represents an initial investigation into determining whether correlations actually exist between metadata and content descriptors in multimedia datasets. We provide a quantitative method for evaluating whether the hue of images on the WWW is correlated with the occurrence of color-words in metadata such as URLs, image names, and attendant text. It turns out that such a correlation does exist: the likelihood that a particular color appears in an image whose URL, name, or attendant text contains the corresponding color-word is generally at least twice the likelihood of the color appears in a randomly chosen image on the WWW. As pointed out in the Introduction, this finding might not be significant in and of itself, but represents an initial step towards quantitatively establishing that other, perhaps more useful correlations exist. These correlations form the basis for exciting novel approaches that leverage semi-supervised datasets, such as the WWW, to overcome the semantic gap that has hampered progress in multimedia information retrieval for some time now.

As this is only an initial investigation, there are plenty of directions for this work to proceed in. Establishing quantitative ways to evaluate correlations between higher-level textual concepts and image content would be very useful for designing the learning algorithms for tasks such as retrieval, classification, and auto-annotation.

5. ACKNOWLEDGMENTS

This work was performed in part under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

6. REFERENCES

[1] Categorical image search demo. http://vision.ece.ucsb.edu/multimedia/search.html.

[2] Cortina: Large-scale, content-based image retrieval on the www. http://vision.ece.ucsb.edu/multimedia/cortina.html.

[3] The DMOZ open directory project. http://www.dmoz.org.

[4] Google image search. http://www.google.com.

[5] Yahoo image search. http://www.yahoo.com.

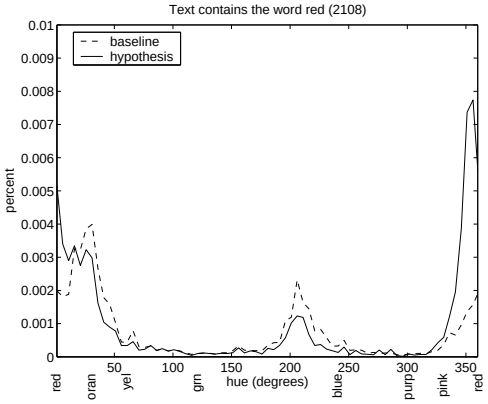
[6] K. Barnard and D. Forsyth. Clustering art. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 434–439, 2001.

[7] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–415, 2001.

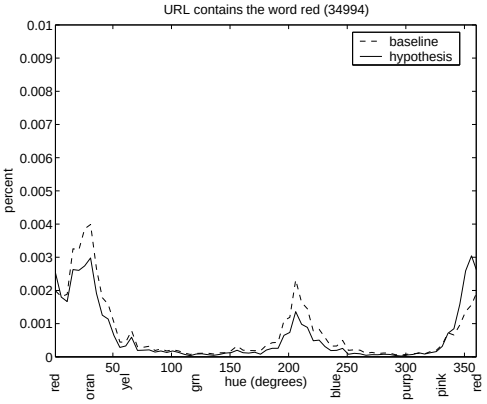
[8] S. Newsam, B. Sumengen, and B. S. Manjunath. Category-based image retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 596–599, 2001.

[9] T. Quack. Cortina: A system for large-scale, content-base web image retrieval and the semantics within. Master's thesis, Swiss Federal Institute of Technology Zurich, April 2004.

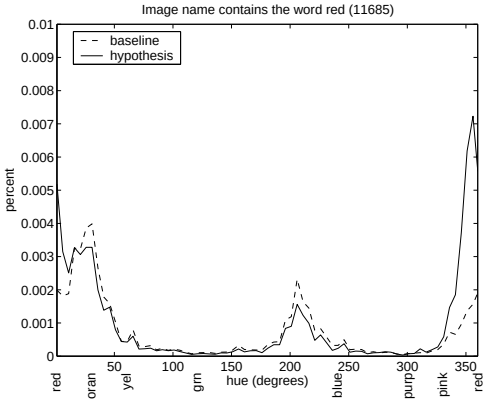
[10] S. Sclaroff, M. L. Cascia, S. Sethi, and L. Taycher. Unifying textual and visual cues for content-based image retrieval on the world wide web. *Computer Vision and Image Understanding (CVIU)*, 75(1):86–98, 1999.



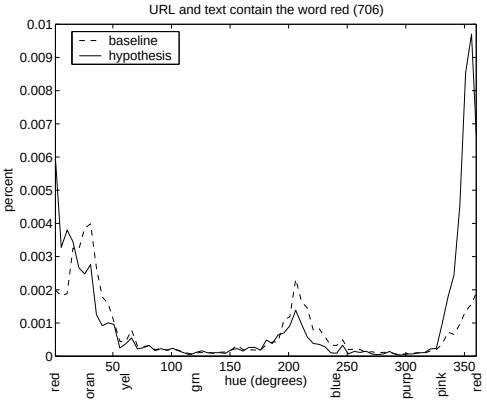
(a) Text contains the word red.



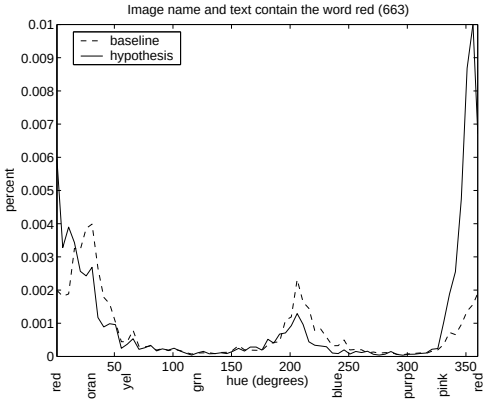
(b) URL contains the word red.



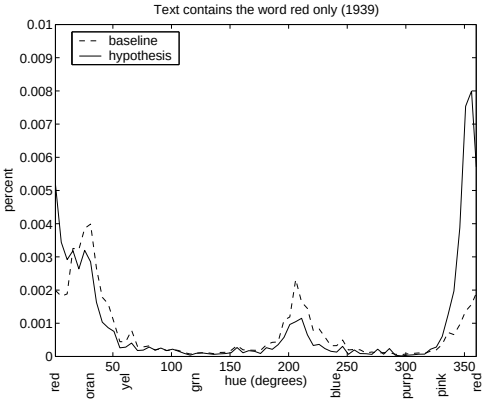
(c) Image name contains the word red.



(d) Both URL and text contain the word red.



(e) Both image name and text contain the word red.



(f) Text contains the word red but no other color-words.

Figure 1: Conditional hue PDFs for hypotheses corresponding to (a) the word red appearing in text, (b) the word red appearing in the URL, (c) the word red appearing in the image name, (d) the word red appearing in both the URL and text, (e) the word red appearing in both the image name and text, and (e) the word red appearing alone in the text. The solid line shows the PDF for the hypothesis. The dashed line shows the baseline PDF computed over all images. The number of images satisfying the hypothesis is indicated in the title in parentheses.

event	# images	$p(red\|e)$	$p(orange\|e)$	$p(yellow\|e)$	$p(green\|e)$	$p(blue\|e)$	$p(purple\|e)$	$p(pink\|e)$
all	1751578	0.0724	0.1309	0.0481	0.0181	0.0703	0.0059	0.0149
all with text	933875	0.0710	0.1292	0.0476	0.0182	0.0691	0.0056	0.0147
red ∈ text	2108	0.1881	0.1159	0.0334	0.0161	0.0434	0.0052	0.0379
orange ∈ text	365	0.1257	0.2153	0.0472	0.0134	0.0417	0.0039	0.0195
yellow ∈ text	808	0.0447	0.1774	0.1298	0.0136	0.0392	0.0040	0.0097
green ∈ text	1333	0.0517	0.1199	0.0858	0.0708	0.0569	0.0035	0.0091
blue ∈ text	2568	0.0503	0.0882	0.0323	0.0155	0.1588	0.0056	0.0121
purple ∈ text	404	0.0555	0.0754	0.0283	0.0155	0.1117	0.0694	0.0391
pink ∈ text	801	0.1383	0.0915	0.0281	0.0113	0.0368	0.0188	0.0902
red ∈ URL	34994	0.0861	0.1003	0.0336	0.0125	0.0416	0.0039	0.0179
orange ∈ URL	1422	0.1676	0.2742	0.0494	0.0152	0.0508	0.0074	0.0141
yellow ∈ URL	2658	0.0545	0.2018	0.1675	0.0161	0.0593	0.0035	0.0100
green ∈ URL	7734	0.0545	0.1151	0.0752	0.0678	0.0557	0.0043	0.0112
blue ∈ URL	12771	0.0462	0.0915	0.0366	0.0183	0.1756	0.0060	0.0106
purple ∈ URL	2569	0.0516	0.0767	0.0256	0.0122	0.0939	0.0597	0.0304
pink ∈ URL	3381	0.1322	0.1039	0.0292	0.0114	0.0417	0.0181	0.0967
red ∈ iname	11685	0.1679	0.1222	0.0401	0.0148	0.0501	0.0059	0.0375
orange ∈ iname	1161	0.1794	**0.2980**	0.0523	0.0136	0.0399	0.0039	0.0115
yellow ∈ iname	1953	0.0453	0.2169	**0.2036**	0.0157	0.0518	0.0030	0.0089
green ∈ iname	4200	0.0418	0.0953	0.0826	**0.0994**	0.0542	0.0038	0.0088
blue ∈ iname	8758	0.0392	0.0786	0.0302	0.0182	**0.2141**	0.0062	0.0099
purple ∈ iname	1499	0.0453	0.0732	0.0293	0.0126	0.1099	**0.0920**	0.0405
pink ∈ iname	2813	0.1366	0.0924	0.0261	0.0111	0.0365	0.0200	0.1118
red ∈ text,URL	706	0.2149	0.1076	0.0352	0.0196	0.0456	0.0074	0.0489
orange ∈ text,URL	104	0.1556	0.2740	0.0606	0.0103	0.0545	0.0093	0.0150
yellow ∈ text,URL	191	0.0337	0.1859	0.1887	0.0164	0.0454	0.0039	0.0097
green ∈ text,URL	437	0.0545	0.1151	0.0752	0.0678	0.0557	0.0043	0.0112
blue ∈ text,URL	804	0.0487	0.0960	0.0340	0.0198	0.1876	0.0075	0.0150
purple ∈ text,URL	146	0.0381	0.0624	0.0298	0.0188	0.1471	0.0839	0.0374
pink ∈ text,URL	250	0.1570	0.0784	0.0197	0.0137	0.0322	0.0219	**0.1338**
red ∈ text,iname	663	**0.2187**	0.1061	0.0351	0.0200	0.0445	0.0076	0.0509
orange ∈ text,iname	99	0.1583	0.2842	0.0635	0.0092	0.0512	0.0098	0.0154
yellow ∈ text,iname	184	0.0343	0.1871	0.1896	0.0167	0.0453	0.0041	0.0099
green ∈ text,iname	389	0.0428	0.0914	0.0786	0.0864	0.0741	0.0080	0.0074
blue ∈ text,iname	722	0.0476	0.0944	0.0346	0.0172	0.1940	0.0070	0.0147
purple ∈ text,iname	138	0.0394	0.0650	0.0298	0.0186	0.1543	0.0882	0.0393
pink ∈ text,iname	249	0.1569	0.0787	0.0198	0.0137	0.0323	0.0220	0.1333

Table 2: Conditional color likelihoods for different hypotheses. The rows represent the hypotheses, and the columns represent the color likelihood conditioned on a hypothesis. There is also a column indicating the number of images that satisfy the hypothesis.