To Search for Images on the Web, Look at the Text, Then Look at the Images*

Ethan V. Munson
Department of EECS
University of Wisconsin-Milwaukee
Milwaukee, WI 53211 USA
munson@cs.uwm.edu

Yelena Tsymbalenko
GE Medical Systems
Waukesha, WI USA
Yelena.Tsymbalenko@med.ge.com

1 Introduction

The World Wide Web can be viewed many different ways. It is simultaneously a giant library, a world-encompassing read-mostly file system, a novel means of expression, and an incredible business opportunity. In addition, the Web is probably the largest public repository of images ever created. Images are particularly important to the Web because it was their introduction into HTML by the developers of NCSA Mosaic that created the explosion of interest in the Web in 1993 and 1994. While the popular saying that "a picture is worth a thousand words" is not uniformly correct, it is definitely true that people find certain pictures far more compelling and informative than any text addressing the same topic. So, it is natural that we should want to use the Web as a sort of "image library" that we can search for images relevant to topics that interest us.

As is described briefly in the next section, considerable research has investigated techniques for indexing and analyzing images using image processing techniques, particularly in closed image and video databases. In general in these systems, users construct visually-based queries using sketches, sample images, or specifications of image parameters. The system compares the image parameters of the query to those of the images in the database in order to find good matches. This approach can be quite effective and these techniques have been applied to the Web with some success [7].

In this paper, we want to argue that image processing is the wrong starting point for most Web image searches. We believe that most users would prefer to create queries using words. Furthermore, images on the Web are almost always accompanied by text that gives useful clues to the images' content. Finally, there are compelling performance reasons to avoid downloading and analyzing images until textual clues suggest that the image might be relevant to a user's query.

The next section gives a brief survey of related research. Section 3 describes our experiments with finding images using HTML metadata. Section 4 discusses the implications of this research and suggests future research directions.

2 Background

There is a large body of research on multimedia indexing and retrieval. Most of this research has been performed using closed databases whose content was under the direct control of the researchers. Systems like QBIC [4] have users provide sample images or sketches and then find images in the database with similar image features, such as shape, color, texture, or object layout. This approach can work well when users seek images with a particular appearance, but it fails when users want to find images with semantics that do not map neatly to image features.

WebSeek [7] is an attempt to create a directory and database of images from the Web. Images in WebSeek's database are categorized into a hierarchy of topics derived from an analysis of image file names. The individual images are assigned to a topic by a human judge. Users can browse or search the topics in the database and can also search using

*This material is based upon work supported by the U. S. Department of Defense and by the National Science Foundation under Grant No. 9734102. Additional support was provided by Sun Microsystems.

image features.

Research on WebSeer [5] investigated how to classify images into categories such as photographs, portraits and computer-generated drawings. To do this, WebSeer supplemented information from image content analysis with information from HTML metadata. WebSeer used several kinds of HTML metadata including the file names of images, the text of the ALT attribute of the IMG tag, and the text of hyperlinks to images to help identify relevant images.[1]

On the Web, the connection between text and image information is strong. HTML documents always contain some text, because HTML is by definition, a textual language. When HTML documents contain images, it is very likely that some other text in the document conveys some of the images' semantics. So, it is natural for Web image search systems to use text as part of the process. This approach is similar to that of Brown et al. [2, 3], who used close-captioned text and speech analysis to index video data.

3 Using HTML Metadata to Find Web Images

We studied the effectiveness of HTML metadata (textual content and structure) for finding images on the Web [8]. We built a prototype image search system that accepted one word textual queries and returned a set of images. This system used the Alta Vista search engine to find HTML documents matching the textual query. It downloaded those documents and analyzed them, using a set of eight clues to decide whether the images in those documents matched the query.

We tested the effectiveness of this system using twelve one-word queries. The system was configured to download up to thirty documents (and all images in those documents) for each query. Decorative images (buttons and advertising banners) were filtered out via a simple heuristic rule that rejected any image with a horizontal or vertical size smaller than 65 pixels. This procedure could have produced 360 Web pages, but only 276 pages containing a total of 1578 non-decorative images were accessible. Each image was examined by one of the experimenters to determine its relevance to the query.

We computed recall and precision statistics for each of the eight HTML metadata clues (shown in Figures 1 and 2, respectively). Only three clues had significant recall[2]: the file name of the image (median = 39.9%), the textual content of the TITLE element of the HTML document (median = 79.5%), and the value of the ALT attribute of the IMG element (median = 2.5%). The remaining five clues, all of which were structurally oriented and depended on HTML authors making good use of HTML's structural features, were essentially useless. Precision for the three successful clues was high, with medians of 83.0% for the image file name, 55% for the TITLE element and 87.5% for the value of the ALT attribute.

The most surprising result is the high level of precision for the ALT attribute of the IMG element. Previous research by Antonacopoulos et al. [1] and by Lopresti and Zhou [6] has shown that more than half of the text values of ALT attributes are empty or wrong. This would suggest that the ALT attribute would have little utility in identifying Web image content.

In fact, our results do not conflict with those of Antonacopoulos or Lopresti. Their research addresses the question of whether, given an IMG element, the ALT attribute value is both present and relevant to the image's content and they both found that it frequently was not. Our poor recall results for the ALT attribute are in accord with this result. However, our study's precision results look at a different question: given an IMG element whose ALT attribute contains the query word, is the referenced image relevant to the word? In this relatively infrequent case, we found that the image was very likely to be relevant to the query word.

Our results should be viewed with caution because this was a small study and it has some flaws. Our use of the Alta Vista search engine, which has a proprietary document relevance rating system, probably affected the results. Our relevance ratings were performed by one person when they should really be based on the judgement of multiple relevance raters. Our use of one-word queries is unrealistic, because one-word queries are probably not specific enough to yield useful results for most users. Our use of the Tidy program to clean the

[1] Swain, the principal investigator for WebSeer, now works for Alta Vista. The Alta Vista image search tool has several features that closely resemble WebSeer.

[2] Because of the lack of a standard corpus for Web image search, our recall statistics were computed by treating the full set of 276 documents retrieved from Alta Vista as the corpus. Thus, our recall numbers can be used to compare our clues to each other, but not to the results of other studies.

HTML documents may have removed some clues due to structural assumptions made by Tidy. In order to address these issues, we have begun a second study to replicate and expand on our initial results.

Finally, it is worth noting that our technique cannot find an image based on text that appears in that image or in other images in the document. Given that image-based text is widely used to control formatting in Web pages, this limits the recall of our technique.

4 Discussion

Our results suggest that textual information in HTML documents can be very useful for finding images on the Web. This is not to say that image features have no place in the image search task, but rather that text may be the best starting point. Once candidate images are identified by clues in the HTML, image processing techniques can be used to refine the search. These are some of the reasons for our belief:

Composing Queries We believe that users primarily seek images that match an idea. In general, users will better express that idea with words than by drawing a sketch or providing a sample image. For example, how do you construct a sketch that will match a motorcycle as seen from many different viewpoints?

Image Features The image features commonly used by image retrieval systems are very low-level and sometimes correspond poorly to human perceptual concepts. Users rarely know what color histogram they seek or how to specify textures.

Performance Image data is relatively large and requires substantial download time. Furthermore, image processing can be computationally intensive. Thus, a system that can make a good estimate of image content from the text of the HTML document can avoid downloading and analyzing images that are probably irrelevant to the query.

Software Engineering Effort By using an existing Web search engine, we can leverage the considerable expertise and development effort in the search engine toward the construction of an image search tool.

We plan to continue to investigate the use of HTML and XML metadata for Web image search. We hope to improve our prototype software, making it faster and connecting it to a variety of search engines. We will replicate our initial results and experiment with other models for identifying relevant images. We also want to explore other hypotheses, such as whether image search techniques should be varied for different kinds of queries (e.g. people, places, and events).

References

[1] A. Antonacopoulos, D. Karatzas, and J. O. Lopez. Accessing textual information embedded in internet images. In *Proceedings of Electronic Imaging 2001: Internet Imaging II, San Jose, California USA*. SPIE, Jan. 2001.

[2] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young. Automatic content-based retrieval of broadcast news. In *Proceedings ACM Multimedia '95*, pages 35–44. ACM Press, Nov. 1995.

[3] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings ACM Multimedia '96*, pages 307–316. ACM Press, Nov. 1996.

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the QBIC system. *Computer*, 28(9):23–32, Sept. 1995.

[5] C. Frankel, M. Swain, and V. Athitsos. WebSeer: An image search engine for the World Wide Web. Technical Report 96-14, University of Chicago, Department of Computer Science, July 1996.

[6] D. P. Lopresti and J. Zhou. Locating and recognizing text in WWW images. *Information Retrieval*, 2(2/3):177–206, May 2000.

[7] J. Smith and S. F. Chang. Visually searching the Web for content. *IEEE Multimedia*, 4(3):12–20, July-September 1997.

[8] Y. Tsymbalenko and E. V. Munson. Using HTML metadata to find relevant images on the Web. In *Proceedings of Internet Computing 2001, Volume II, Las Vegas*, pages 842–848. CSREA Press, June 2001.

Query Word	Clue							
	Image Filename	Document Title	Image ALT text	Anchor text	Anchor Title	Paragraph text	Centered text & image	Nearby heading
Gorbachev	26.0	84.0	5.0	0.0	0.0	5.0	0.0	0.0
Yeltsin	60.0	60.0	0.0	0.0	0.0	0.0	0.0	0.0
Streisand	23.0	84.6	0.0	0.0	0.0	0.0	0.0	0.0
Yelena	11.1	85.0	5.0	0.0	0.0	0.0	0.0	0.0
Ekaterina	0.0	60.0	14.3	0.0	0.0	0.0	0.0	0.0
Paris	62.0	62.0	0.0	0.0	0.0	0.0	0.0	0.0
London	12.5	95.8	12.5	0.0	0.0	0.0	0.0	0.0
Bremen	80.0	90.0	33.0	0.0	0.0	0.0	0.0	0.0
Spokane	90.0	100.0	30.0	0.0	0.0	0.0	0.0	0.0
Explosion	25.0	75.0	0.0	0.0	0.0	0.0	0.0	0.0
Sunset	88.8	11.1	0.0	0.0	0.0	0.0	0.0	0.0
Hurricane	53.8	38.5	0.0	0.0	0.0	0.0	0.0	0.0
Median %	39.9	79.5	2.5	0.0	0.0	0.0	0.0	0.0

Table 1. Recall percentages for each clue and each query.

Query Word	Clue							
	Image Filename	Document Title	Image ALT text	Anchor text	Anchor Title	Paragraph text	Centered text & image	Nearby heading
Gorbachev	83.0	46.0	100.0	—	—	100.0	—	—
Yeltsin	100.0	60.0	—	—	—	—	—	—
Streisand	100.0	47.8	—	—	—	—	—	—
Yelena	66.7	89.5	100.0	—	—	—	—	—
Ekaterina	—	100.0	100.0	—	—	—	—	—
Paris	84.0	70.0	—	—	—	—	—	—
London	60.0	46.9	75.0	—	—	—	—	—
Bremen	66.7	69.2	100.0	—	—	—	—	—
Spokane	75.0	71.4	100.0	—	—	—	—	—
Explosion	100.0	50.0	—	—	—	—	—	—
Sunset	36.0	50.0	—	—	—	—	—	—
Hurricane	100.0	35.7	—	—	—	—	—	—
Median %	83.0	55.0	87.5	—	—	—	—	—

Table 2. Precision percentages for each clue and each query. Dashes are used for clue-query combinations that had zero recall, since precision cannot be computed when there is no recall. Median precision percentages are computed based only on those queries that had some recall.