Using HTML Metadata to Find Relevant Images on the World Wide Web*

Yelena Tsymbalenko
Ethan V. Munson
Dept. of EECS
University of Wisconsin-Milwaukee
Milwaukee, WI 53201 USA
{yelena,munson}@cs.uwm.edu

January 9, 2001

1 Introduction

The World Wide Web has become one of the largest information repositories the world has ever seen. While much of that information is textual, a substantial amount is drawn from other media, especially static images.

An obvious way to use the Web is to treat it as a sort of library that can be indexed, catalogued, and queried. Web search engines and directory-style portals do just that by indexing or cataloguing the Web's textual content to provide convenient access services to end-users.

Providing search services for the Web's image content has been more difficult. A number of researchers have developed Web image search tools, but these systems are limited by the use of visually-based queries and databases that represent a small subset of the Web's content. Companies like Alta Vista and Lycos Multimedia are also beginning to provide image search services. Some features of Alta Vista's image search system suggest that their technology is similar to that reported in this paper.

This paper describes a new approach to finding images on the Web. Instead of analyzing the images themselves using image processing techniques, our software examines the HTML source code that refers to the image and using only this textual information, decides whether or not an image is relevant to a query.

In the next section, we provide some background on prior research into image search and multimedia research using similar strategies. In Section 3, we describe the software testbed we built, while in Section 4 we describe the results of experiments that we ran using this testbed. Section 5 closes with conclusions and suggestions for future research.

*This research was sponsored by the U. S. Department of Defense. Ethan V. Munson was also supported by NSF CAREER award CCR-9734102.

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. Examples of such research are easily found in recent conference proceedings [2, 1] and journals [8].

A good example of this course of research is the IBM Almaden Center's Query-By-Image-Content (QBIC) system [6]. QBIC allows users to make image queries based on image features such as shape, color, texture, and object layout. Users define queries either by providing a sample image or by using a graphical tool to make a sketch or diagram. QBIC has a well-developed visual query language and an interesting GUI. Its use of image features for indexing and querying is both an advantage and a disadvantage. When users are seeking images with a particular appearance (e.g. mix of colors, object with a particular shape), it is very helpful. When users are looking for pictures of particular content, it is less helpful because the low-level image characteristics give only limited insight into the real semantics of images. For example, a person's facial appearance may be fairly constant, but their clothing may not be. Many objects (e.g. pencils, fish, or motorcycles) look radically different depending on camera viewpoint. QBIC's approach also appears ill-suited to the scale of the Web. QBIC constructs its indices by pre-analyzing each image in its database. This is computationally demanding and it is difficult to see how it can be done for the Web as a whole.

WebSeek [10] is a more direct attempt to create a directory and database of images from the Web. WebSeek uses a mix of automated and manual techniques to create a database of images downloaded from the Web. It automatically inspects HTML documents, extracting keywords from the image file names that are used to create a histogram of file names. This histogram is used to manually construct a subject hierarchy for the downloaded images. In another manual step, the downloaded images are mapped into the subject hierarchy. Once this is done, WebSeek users can browse the categories in the subject hierarchy, search the categories by keyword, and search the database using image features, especially color histogram information.

WebSeek has a large database of Web images and supports both text-based and image-based queries. Text-based queries have more semantic content than image-based queries. However, it seems unlikely that WebSeek's database can approach the scale of the entire Web, since manual categorization of images is a slow and labor-intensive process.

WebSeer [7] is the system most closely related to this research. The principal investigator, Swain, now works for Alta Vista. Alta Vista has a new image search tool whose qualities appear to derive from Swain's research on WebSeer.

The goal of research on WebSeer was 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. Since the WebSeer research emphasized image categorization, this use of metadata is not discussed in detail in any of the WebSeer papers. We assume that the metadata was helpful, but a detailed analysis was not provided.

The research most similar in spirit to that reported in this paper was conducted by Brown et al. [3, 4]. In their first study [3], they used textual "closed captions" transmitted with broadcast news to index stored video. In the second study [4], they used speech recognition techniques to analyze the audio components of video mail. Then, the textual content of the recognized speech was used for indexing the video content. In both studies, Brown et al. took advantage of the fact that data in two media were traveling together and exploited data in one medium to better understand the content in the other.

3 Image Search Architecture

We applied the cross-media indexing strategy of Brown et al. to the Web image search problem. We started with the observation that images on the Web are almost always accessed through HTML documents and that the bulk of the content of HTML documents is textual. In addition, the HTML source includes text that defines a hierarchical information structure. We consider both the textual content and the structure of HTML documents to be "metadata" describing images and use this metadata to determine which images may be relevant to a query.

The second aspect of our strategy was to exploit existing Web search engines in order to search the entire Web, rather than a closed database of previously downloaded images. By using existing search engines, we saved considerable engineering effort and were able to exploit the search engine designers' considerable expertise in computing the relevance of Web documents to textual queries.

We constructed a Web image search application composed of four modules: text search, document download and cleaning, document analysis, and search results interface.

The text search module accepted a one-word query and sent it to the Alta Vista search engine. Alta Vista returned an HTML document with links to ten Web pages that best matched the query. In addition, the bottom of this document had links to as many as nineteen other pages of search results. In effect, Alta Vista returned links to 200 pages having some relevance to our one word queries. The text search module extracted the URLs of these pages from the search results documents and sent a subset of these URLs to the document download and cleaning module.

The download and cleaning module first used the low-level HTTP interfaces to download the Web pages for each URL. In addition, this module downloaded every image referenced by each document, in order to facilitate later analysis. At this point, we confronted the problem that many HTML documents on the Web are ill-formed and thus are difficult to analyze. We solved this problem by using the "Tidy" application [9] developed by Raggett for the Web Consortium. Tidy uses heuristic rules to translate HTML (well-formed or ill-formed) to well-formed XHTML (an analog of HTML that conforms to the XML specification [5]).

The document analysis module parsed the well-formed XHTML documents into an internal tree representation and then searched for "clues" that might indicate that an image in the document matched the query. The analysis module considered an image to match the query if the query appeared in any of the following eight places:

1. An image's file name;
2. The textual content of the document's TITLE element;
3. The value of the ALT attribute of the IMG element;
4. The textual content of an anchor (A) element whose target was the image's file;
5. The value of the TITLE attribute of an anchor (A) element;
6. The textual content of the paragraph that was the parent of the IMG element;
7. The textual content of any paragraph located within the same CENTER element as the IMG element; and
8. The textual content of heading elements that precede the image.

Finally, the search results interface module took the list of matching images generated by the document analysis module and created a Web page interface with links to the matching images and the pages that they came from. This final interface was not designed for end-users, who would certainly prefer an interface based on thumbnail images, but it was suitable for our image search experiments.

At this point, some comments on the design of the testbed are appropriate.

- By using a commercial search engine as the first step in image search, we saved a tremendous amount of engineering effort. However, it clearly makes the set of images returned by the system depend on the behavior of the search engine. At this time, we have no idea what effect the choice of search engine had on our research.
- The eight "clues" used to find matching images were derived from the work on WebSeer and from our own study of the HTML specification and of Web document design practice.
- About 1% of the HTML documents we downloaded were so ill-formed that the Tidy program could not produce an XHTML version.
- We determined that images smaller than 65 pixels in either the horizontal or vertical dimension could be ignored. We found through informal experimentation that such images were essentially always "decorative" elements like borders, bullets, or banner advertisements.

4 Image Search Experiment

Using the testbed described in the previous section, we conducted an image search experiment in the fall of 1999 to assess the effectiveness of our strategy. Our goal was to answer two research questions:

- Which HTML features reveal the most information about images in a document?

- Do image search results depend on the type of query made?

We used our testbed to search for images using twelve one-word queries drawn from five categories. The queries, listed by category, were:

Famous People: "Gorbachev," "Yeltsin," and "Streisand"

Non-famous People: "Yelena" and "Ekaterina"

Famous Places: "Paris" and "London"

Less-famous Places: "Bremen" and "Spokane"

Phenomena: "Explosion," "Sunset," and "Hurricane"

We modified the testbed so that, for each query, it downloaded 30 of the 200 pages returned by Alta Vista and all of the images on those pages. The 30 pages were taken from the first, eleventh, and twentieth search results pages.[1] This procedure could have produced 360 Web pages, but only 276 pages containing a total of 1578 non-decorative images were accessible. For each image, we recorded which of the eight clues would have caused that image to be retrieved by our software. In addition, one of us (Tsymbalenko) looked at each image and classified it as either "relevant" or "not relevant" to the query word.

4.1 Results

We used the human relevance ratings and the data about which images would have been retrieved to compute the standard information retrieval measures of precision and recall. Precision is the proportion of images that a clue caused to be retrieved that are actually relevant to the query. It is computed by the formula

$$\text{Precision} = \frac{\text{Retrieved images that are relevant}}{\text{Total retrieved images}}$$

Recall is the proportion of relevant images (out of the "complete" collection) that are retrieved and computed by the formula

$$\text{Recall} = \frac{\text{Relevant images that were retrieved}}{\text{Total relevant images in collection}}$$

It is important to give a cautionary note about our recall statistics. Recall is normally computed using some standard body of material (e.g. one year's issues of a major newspaper), called a corpus, which is used as the entire "collection" over which searches are performed. Our recall statistics were computed using the 276 HTML documents returned by Alta Vista as our corpus. This is clearly *not* a valid approach, since the set of documents returned by Alta Vista were chosen precisely because they were

[1] We originally chose this approach in the mistaken belief that there would be interesting differences between the first search results (high relevance) and the last search results (low relevance). In retrospect, this was a pointless exercise, because Alta Vista was finding tens of thousands of Web pages that matched our queries, but only providing the best 200 of these. All of these 200 pages were highly relevant to our queries.

Query	Clue							
	1	2	3	4	5	6	7	8
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 percent	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.

highly relevant to our query and this could easily bias our results. Unfortunately, we know of no standard Web corpus on which to perform our tests. Thus, we use our recall statistics only to give initial results on the relative merits of our metadata clues, not to make comparisons to other search approaches.

Our recall results are shown in Table 1. Only clues 1 (image file name) and 2 (content of document's TITLE element) show high levels of recall with medians of 39.9% and 79.5%, respectively. Clue 3 (value of the ALT attribute of the IMG element) shows a modest level of recall with a median of only 2.5%, but individual recall percentages as high as 33%. Only one other clue, number 6 (textual content of a paragraph that contains an IMG element), showed any recall at all.

Precision results are shown in Table 2. For clue 1 (image file name), precision ranges from 36% to 100%

4.2 Discussion

Examining the results of the image search experiment closely, several key results emerge.

The three clues (1, 2, and 3) that show significant levels of recall are relatively simple. Image file name (clue 1) presumably works because Web site designers prefer mnemomic names for image files. The TITLE element (clue 2) is designed to provide a high-level description of a document's content and is widely used because it get listed in search engine results and the browser's title bar. The ALT attribute of the IMG element (clue 3) is explicitly designed to be a textual alternative to the image itself. The remaining five clues generally emphasize HTML's underlying structural model, and based on our results, do not seem to be widely used idioms among HTML authors and showed essentially no recall.

Looking at the types of queries, image file name (clue 1) had poor recall for the

Query	Clue							
	1	2	3	4	5	6	7	8
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 percent	83.0	55.0	87.5	0.0	0.0	0.0	0.0	0.0

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.

names of people, but excellent recall for place name queries, particularly the less famous cities. An informal look at the details of this phenomenom suggests that Web designers often use nicknames for the image file names people (e.g. “Gorby” for “Gorbachev”), but usually use full names for places.

The precision results are more striking. The three queries that had some recall all showed good precision. The image file name had precision ranging from 36% to 100% with a median of 83%. The content of the TITLE element had precision ranging from 35.7% to 100% with a median of 55%. These precision results are strong by the normal standards of textual information retrieval. The precision results for the value of the ALT attribute of the IMG entity are quite impressive. In general, this clue had 100% precision.

Now, it is possible to examine the original research questions that we posed.

First, we asked which HTML features revealed the most information about the images in a document. It is clear that only three of the HTML features that we tested showed any real utility for identifying the content of images. Image file name, the content of the TITLE element, and the value of the ALT attribute appear useful in image search, while the other clues we tested do not appear useful.

Second, we asked whether the type of query affected our image search results. The type of query does seem to affect our recall results, where the names of people show less recall than the names of places. No consistent effect can be seen in the precision results.

4.3 Cautionary Notes

These results should be viewed cautiously. This was a small study and it has some flaws.

- Our results are affected by our use of the Alta Vista search engine. It is not clear what effect this had, but the use of a different search engine might produce different results.
- Our relevance ratings for the images were performed by one person. They should really be based on the judgement of multiple relevance raters. Also, image relevance is harder to judge than text relevance and may require somewhat different rating methods.
- The distinction between famous and non-famous people is confounded with another effect. All of the famous names used were family names. Both of the non-famous names used were personal names. Also, one of the non-famous names is actually the personal name of moderately famous person (the skater Ekaterina Gordeeva).
- The one word queries we used do not allow the construction of very precise queries, especially for the names of people.
- Our use of the Tidy program may have removed some clues. We believe that an author writing HTML like the following, probably views the image as part of the first paragraph (that is, a child of the paragraph). Tidy makes the image a sibling.

```
<P>Some text. <IMG href="img.gif">
<P>More text.
```

5 Conclusion and Suggestions for Future Research

This paper has described new software for finding images on the Web based on simple text queries and an experiment testing the techniques used by that software. The software used a text search engine to find documents containing text matching the query and then analyzes the content of the document to determine whether the images in that document may be relevant to the query. The experiment demonstrated that some of the techniques used to identify relevant images were effective and that others were not. Its results also suggested that the type of query made alters the effectiveness of the search technique.

Why is this software interesting? Image search is an inherently interesting problem and is being studied widely. This software is interesting because it is able to find relevant images without actually downloading or analyzing those images. Instead, it examines only the text that surrounds the reference to the image in the HTML document. It is widely known that image download requires substantial amounts of time when traversing the Web. Any system that can find images without downloading them

has an inherent performance advantage over systems that must download images. Furthermore, we believe that the text in an HTML document gives more precise semantic information about the content of images than any existing image processing technique. Image processing can be used to determine that an image shows the face of a person, but the file name of the image may say exactly what person is shown.

Considerably more research is called for. Our basic results showing the success of our technique need to be replicated using a larger study, more complex queries, and a more robust relevance rating system. While our techniques appear strong and they have an efficiency advantage over image processing approaches, there is no reason that textual metadata cannot be combined with image processing in order to produce even better results. Finally, we continue to believe the type of query will interact with search heuristics in interesting ways.

References

[1] *Proceedings, ACM Multimedia '00*. ACM Press, November 2000.

[2] *Proceedings, ACM Multimedia '99*. ACM Press, November 1999.

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

[4] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck 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, November 1996.

[5] World Wide Web Consortium. Extensible markup language (xml) 1.0. Available on the Web at http://www.w3.org/TR/2000/REC-xml-20001006, October 2000. Second edition.

[6] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: the QBIC system. *Computer*, 28(9):23–32, September 1995.

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

[8] *Multimedia Systems*. Published jointly by Springer Verlag and ACM Press. Quarterly academic journal.

[9] Dave Raggett. Clean up your web pages with HTML TIDY. Available at www.w3.org, August 2000. Version of August 4, 2000.

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