COMPAQ

Searching for Multimedia on the World Wide Web

Michael J. Swain

Cambridge Research Laboratory

Cambridge Research Laboratory
Technical Report Series

CRL 99/1
March 1999

Cambridge Research Laboratory

The Cambridge Research Laboratory was founded in 1987 to advance the state of the art in both core computing and human-computer interaction, and to use the knowledge so gained to support the Company's corporate objectives. We believe this is best accomplished through interconnected pursuits in technology creation, advanced systems engineering, and business development. We are actively investigating scalable computing; mobile computing; vision-based human and scene sensing; speech interaction; computer-animated synthetic persona; intelligent information appliances; and the capture, coding, storage, indexing, retrieval, decoding, and rendering of multimedia data. We recognize and embrace a technology creation model which is characterized by three major phases:

Freedom: The life blood of the Laboratory comes from the observations and imaginations of our research staff. It is here that challenging research problems are uncovered (through discussions with customers, through interactions with others in the Corporation, through other professional interactions, through reading, and the like) or that new ideas are born. For any such problem or idea, this phase culminates in the nucleation of a project team around a well articulated central research question and the outlining of a research plan.

Focus: Once a team is formed, we aggressively pursue the creation of new technology based on the plan. This may involve direct collaboration with other technical professionals inside and outside the Corporation. This phase culminates in the demonstrable creation of new technology which may take any of a number of forms - a journal article, a technical talk, a working prototype, a patent application, or some combination of these. The research team is typically augmented with other resident professionals—engineering and business development—who work as integral members of the core team to prepare preliminary plans for how best to leverage this new knowledge, either through internal transfer of technology or through other means.

Follow-through: We actively pursue taking the best technologies to the marketplace. For those opportunities which are not immediately transferred internally and where the team has identified a significant opportunity, the business development and engineering staff will lead early-stage commercial development, often in conjunction with members of the research staff. While the value to the Corporation of taking these new ideas to the market is clear, it also has a significant positive impact on our future research work by providing the means to understand intimately the problems and opportunities in the market and to more fully exercise our ideas and concepts in real-world settings.

Throughout this process, communicating our understanding is a critical part of what we do, and participating in the larger technical community—through the publication of refereed journal articles and the presentation of our ideas at conferences–is essential. Our technical report series supports and facilitates broad and early dissemination of our work. We welcome your feedback on its effectiveness.

Robert A. Iannucci, Ph.D.
Director

Searching for Multimedia on the World Wide Web

Michael J. Swain

March 1999

Abstract

The proliferation of multimedia on the World Wide Web has led to the introduction of Web search engines for images, video, and audio. On the Web, multimedia is typically embedded within documents that provide a wealth of indexing information. Harsh computational constraints imposed by the economics of advertising-supported searches restrict the complexity of analysis that can be performed at query time. And users may be unwilling to do much more than type a keyword or two to input a query. Therefore, the primary sources of information for indexing multimedia documents are text cues extracted from HTML pages and multimedia document headers. Off-line analysis of the content of multimedia documents can be successfully employed in Web search engines when combined with these other information sources. Content analysis can be used to categorize and summarize multimedia, in addition to providing cues for finding similar documents.

CRL Technical reports are available on the CRL's web page at
http://www.crl.research.digital.com.

Compaq Computer Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700
Cambridge, Massachusetts 02139 USA

1 Introduction

The World Wide Web is full of images, video, and audio, as well as text. Search engines are starting to appear that can allow users to find such multimedia, the quantity of which is growing even faster than text on the Web. As 56 kbps (V.90) modems have become standardized and widely used, and as broadband cable modem and ADSL services gain following in the United States and Europe, multimedia on the Web is becoming freed of its major impediment: low-bandwidth consumer Internet connectivity.

Traditional search engines on the World Wide Web index HTML pages, treating each one as a document, and indexing the text on the page to allow users to find them. Multimedia search engines index the images, video, or audio documents that are linked to HTML pages, or embedded within them, serving the needs of those who are interested in the multimedia itself. Most of these multimedia documents are images, although a growing number of video and audio documents are appearing on the web as well. If a number of different multimedia documents appear on one page, or are linked to a page, each one can be given a separate entry in the index, and users are able to search for each multimedia document separately.

Creating a multimedia search engine introduces challenging issues in a number of different areas. One, naturally, is how to represent and index the documents for efficient retrieval. Intertwined with this question is what interface to present the user, including how to query the index, and how to summarize the results. And there are other issues, such as respecting the rights of the copyright holders, and obtaining the cooperation of the sites that are being indexed.

Search engines are typically free to the user, financed only by banner ads or product tie-ins. So the cost per search has to be kept low. On the other hand, the number of multimedia documents, like the number of HTML pages, is extremely large. AltaVista$^{\mathrm{TM}}$ counts 94 million pages with embedded images or links to images, and 1.2 million linked to audio and video files. [1] Therefore, as for text search engines, satisfying a multimedia search query must be an extremely efficient operation, scalable to millions of documents.

The user interface to an image search engine will be used by millions of different people, most of whom want quick results and will have not read any instructions on how to use the system. So the user interface should be simple and, as far as possible, familiar to users of text search engines such as AltaVista.

2 Indexing Images: WebSeer

The main cue for indexing HTML pages is the text that appears on the page, ignoring HTML tags. What should be used to index, say, an image? An image by itself defies

[1] According to [1], AltaVista indexes somewhat under one-half of the static pages on the World Wide Web. Each page that is linked to multimedia may be linked to a number of multimedia documents. On the other hand, a number of HTML pages can be linked to the same document, or to copies of the document. So these numbers give only rough estimates of the number of multimedia documents on the Web. The number of multimedia documents is probably significantly higher than the numbers given, perhaps by a multiple of 2-4 times.

the standard techniques used for indexing text documents. The three main techniques that have been used in multimedia search engines to date are:

1. Text found on the HTML page that references the image determined to be relevant to the content of the image.

2. Category selection based on attributes extracted from the header of the image, from analysis of its contents, and from the relevant text.

3. Similarity to another image in the collection.

WebSeer [5], an image search engine developed at the University of Chicago, used the first two techniques. It used text from the filename, alternate text, text determined to be a caption, text in hyperlinks to the image, and text from the HTML page title. Text was determined to be a caption by being within the same `<center>` tag, or within the same table cell. Weights were assigned to text from these different location, based on the developers' prior judgements about the likelihood of the text to be relevant to the image.

Users could specify categories based on the image dimensions, the file size, whether the image was color or black and white, whether it was a photograph or a non-photograph (typically computer-generated graphic), and, when looking for people, specify the number of faces in the scene, and the size of the largest face (close-up, bust, half-body, full-body). The photograph/graphic computation used the type of image (GIF, JPEG), dimensions of the image, its color distribution, the distribution of pixel-neighbor differences, and other tests. The results of these tests were combined using multiple decision trees trained on hand-labeled images.

Rowley *et al*'s face finder [9] was used to locate the faces in the images. The interface allowed users to select the number of people in the image, and the size of the largest face, indicating a close-up, bust shot, half-body shot, or full-body shot. While the system suffered false-negatives, usually due to rotated, small, or partially-occluded faces, there were few false positives.

In WebSeer, thumbnails of the images were presented to the user; clicking on the thumbnail allowed the user to view the actual image being indexed, and another icon next to the thumbnail allowed the user to go to the page that contained or referenced the image.

WebSeer incorporated a form of relevance feedback, that modified the boolean query to incorporate terms that were contained in the relevant text associated with the majority of the positive examples, and to negate terms that were relevant to the majority of the negative examples. Terms that occurred in the majority of both positive and negative examples were ignored.

Related research to WebSeer in Web image retrieval includes Columbia University's WebSEEk [12] and work by Neil Rowe [8].

3 The AltaVista Photo Finder

The AltaVista™ Photo Finder shares some features with WebSeer. Users enter text queries, and can select to narrow the search to photos/non-photos, and to color/black

Figure 1: Top: WebSeer query page. Bottom: WebSeer response page.

and white. The text queries match relevant text extracted from the Web pages, and thumbnails summarize the results of the query. Unlike WebSeer, the indexing engine for the Photo Finder is the same indexing engine as powers the main AltaVista site, and so the system can efficiently deliver responses to millions of users per day, querying a database of tens of millions of images.

A feature of the Photo Finder is a link to visually similar images to an image on the results page. This link leads to a pre-computed set of images that are the most visually similar within the subject category to which the image has been assigned. Visual similarity is determined by computing a multidimensional feature vector composed of subvectors computed from color and texture distributions, and what $\mathrm{Virage}^{\mathrm{TM}}$ calls composition, apparently related to the spatial layout of color, and structure, related to the edge orientations and positions in the image. Nearest neighbors of the resulting multi-dimensional vectors are computed to determine the similar images. The similarity is restricted to a subject category subset of the images that more or less evenly divide the photos on the Web; a few of the categories are architecture, art, animals, nature, space, science, sports, entertainment, politics, computer, science fiction, supermodels, trains, and military. Doing so has two advantages: The similar images returned tend to be related in subject matter as well as visually similar, and the cost of computing visual similarity is reduced because of the much smaller number of potential similars. The cost of computing all the nearest neighbors is proportional to the square of the number of images in the subject category, so the operation is performed off-line. AltaVista's logs indicate that 5-10% of the searches performed on the Photo Finder are similarity queries.

Another feature of the Photo Finder is a Family Filter, that screens out objectionable images from the Photo Finder results. AltaVista automatically categorizes entire HTML pages using a text-based classifier to separate the objectionable pages from others. In addition, it employs the results of a commercial rating service, and trust-worthy editors and AltaVista users.

One way to improve the performance of pre-computed image similarity is to use relevant textual information as part of the vector representation of the image content, not only to categorize the image. La Costia and Sclaroff at Boston University present one approach for doing this in their Image Rover search engine [10, 3], and give evidence that the textual information is of considerable value for user satisfaction in finding images they consider "relevant" to the original image, in the information retrieval sense. But some of the more promising ways to improve the effectiveness of such a system that have been explored may be impractical in a system that has to serve an index of tens of millions of images to millions of users per day, at a cost of a few cents per page view.

One example of a feature that has appeared in research systems but not commercial image search engines is relevance feedback. Similarity can mean many different things to different users, or the same user at different times; relevance feedback could help to give enough information to the system to determine the type of similarity desired by the user. But relevance feedback that computes nearest neighbors in a high-dimensional vector space is a computationally expensive feature at runtime, because if feedback is permitted, nearest neighbors cannot be computed off-line. An example of a research Web search engine that incorporates relevance feedback including both text and image

Figure 2: AltaVista Photo Finder response page.

information is the ImageRover system.

Another attractive but computationally expensive feature would be to allow the user to introduce images to the system, for similarity computation. To do so would require computing the features for the user's image at runtime, and finding nearest neighbors in an on-line computation. Finally, users might wish to select a subset of an image to be the target for image similarity. Unless the image is pre-segmented into a small number of regions, and the similarity pre-computed for these, finding similarity to image subsets could also be prohibitively expensive.

Another area where image processing is of value is filtering adult images, which are prevalent on the World Wide Web. For this task there are a number of different sources of information that can be fruitfully combined. Rehg and Jones have shown that adding skin-color detection to an automated text-analysis system significantly increases the detection rate of adult images [7]. The skin-color detector is a learning-based system, trained on large amounts of labeled skin data sampled from the World Wide Web (over 1 billion pixels). As Jones and Rehg mention, this system is expected to be improved by texture analysis, and face recognition, which can determine if a large patch of skin viewed in the image is explained by being part of a face or not.

3.1 Other Issues

An image search engine will bring users to the sites that it indexes, so these sites are usually predisposed to be cooperative in order to maximize their user traffic. But site operators can become concerned if the users visit their sites without viewing the banner ads from which they receive their revenue. For this reason, the Photo Finder links only to the HTML pages linking to or embedding the images, not the images themselves – even though each image is addressable by its own URL. If sites do not wish to be indexed they can easily block the crawler visiting the site. In fact, crawlers, including the AltaVista crawler respect a standard protocol known as the robots exclusion standard for protecting content from being indexed.

Some concern has been expressed by copyright holders over reduced-resolution thumbnail summaries of images as being "whole works", and therefore possibly copyright violations in themselves. AltaVista's contention is that their use of thumbnails falls under fair use, as do the summaries supplied by text search engines. Copyright holders can become particularly concerned about indexes of illegal copies of their images, encouraging users of the search engine to view their images at sites other than the sites they sanction and receive royalties from. AltaVista, in particular, a highly-trafficked and publicly-visible site, has to be sensitive to these issues.

4 Indexing Web Video and Audio

Video and audio documents are not as common on the Web as images, but their numbers are growing as rapidly, if not faster, than the Web as a whole. A crawl of 623 thousand HTML pages starting at http://www.yahoo.com, performed in July 1998, found 49 times as many links to images (6.6 per page) as to all types of video and audio (0.13 per page), and 8.2 times as many links to GIF images than JPEG images. Of course, many of the links to GIF images are links to structural elements of Web pages such as bullets, buttons, and separators. A large crawl of the World Wide Web performed in November 1998 found 1.2 million links to multimedia documents. Figure 3 shows the relative proportion of different formats of audio and video, with audio formats (AIFF, AU, MIDI, MP3, RealAudio, WAV) displayed on the left, and video formats (ASF, AVI, MPEG, Quicktime, RealVideo) displayed on the right. Audio files in various formats make up 72%, with 28% being video. RealNetworks' RealAudio and RealVideo formats make up 46% percent of the files between them.

Compaq's Cambridge Research Laboratory has developed a research prototype for indexing video and audio documents on the Web [4]. As for images, relevant text from the HTML pages referencing the multimedia documents is used to build the index, and the multimedia documents themselves are downloaded and converted to common formats for analysis. The header information for some types of multimedia documents can be quite rich; for example RealAudio and RealVideo files, widely-used streaming multimedia formats, typically contain title, author, and copyright information that is

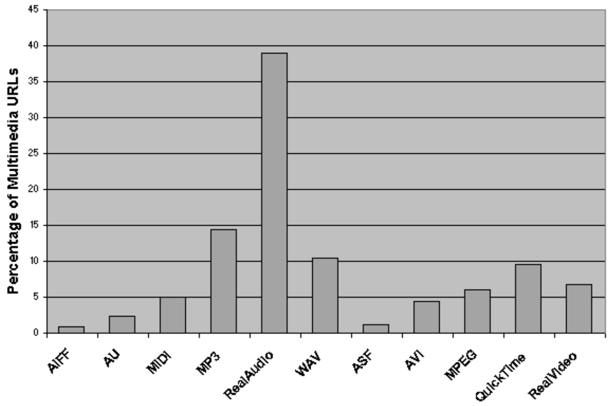

Figure 3: Multimedia formats on the Web.

displayed by the player, in addition to other useful information such as duration and bandwidth.

Synchronized multimedia files, that stream multiple sources of information in parallel, or coordinated in sequences, can contain even more useful information for indexing. An example is the World Wide Web Consortium's Synchronized Multimedia Integration Language (SMIL) format, which is supported by the RealSystem G2 player. SMIL documents can contain indexable streamed, formatted text that is displayed to the user, in addition to title, author, and copyright information. A special category of text, extremely useful for indexing, is displayed only if the user selects to display optional captions for hearing impaired viewers, in analogy to closed captions for TV shows.

The system analyzes the content of video documents to extract a representative keyframe, to be displayed as thumbnails are for the images. To do so, it segments the video into shots, locates the best shot, and then selects a representative frame from the selected shot. The best shot has significant motion and spatial activity, and is likely to include people. For efficiency, motion is measured by frame differences, spatial activity by the distribution of grayscale values, and the presence of humans by skin color detection. Furthermore, long shots are given a preference as they are often a sign of the intent of the content producer to emphasize a portion of the video.

The representative frame from within the shot contains less motion (i.e. is not blurred), has large spatial activity, and, preferably, contains people. To save computation and bandwidth, only the first 30 seconds of video are analyzed to select the representative keyframe.

5 What's Next?

There is a tremendous amount of research being done in multimedia indexing that could at some point be of use to Web multimedia search engines. Here I will describe some near-term directions that are likely to provide some significant benefit to the typical user of a Web multimedia search engine.

The most important research direction in this area is likely to be the use of speech recognition. Speech recognition will open much of the video and audio on the Web to be indexed like text documents can be indexed today, since about 55% of the audio we've encountered on the Web contains speech. To obtain these numbers, random samples of audio and video URL's were made, from broad crawls of the World Wide Web starting at http://www.yahoo.com. The 12 thousand files were hand-labeled. The distribution found was 59% speech, 40% percent music, and 1% percent "other", that did not fall into either of these two categories. About 90% of the speech in audio documents obtained by randomly sampling the 623 thousand document crawl is in English, with almost all of the rest in other European languages.

Applying automatic speech recognition engines to English speech found on the Web results in widely varying error rates, depending on the audio recording quality, presence of background sounds, vocabulary, amount of compression, accents, and so on. On uncompressed broadcast news sources, error rates of 20-30% are obtainable, for systems that run 10 times slower than real-time on a single Pentium II processor. If such a stream is RealAudio-compressed down to 8 kilobits per second, a typical bandwidth for streamed audio documents on the Web, the error rates increase by about 5%, provided appropriate audio models are used for the distorted speech.

Commercial systems exist for a number of European and Asian languages besides English; error rates are comparable to English.

Retrieval has been found to be remarkably robust in the presence of errors in speech recognition [6], although extrapolating results of multi-word queries and relatively small databases to Web search engines should be done with some caution [11]. One source of robustness to error is that words important to the subject of a document tend to be repeated a number of times, raising the chance of detection by the speech recognition engine. Words less important to indexing such as prepositions are at least as likely as e.g. proper nouns to be mis-recognized, provided the proper nouns are in the language model. And false positives are usually broadly distributed over the words in the language model, which contains about 50 thousand words for speech recognition systems applied in such domains.

Many other solvable problems are of potential value to multimedia search. Audio/video classification could subdivide the search space and provide useful annotations on the results page. An important one is family-friendly screening. Skin detection and word-spotting will be of use in these domains.

Many of the searches done by users of image search engines are for pictures of people. People/portrait detection was implemented with some success in WebSeer, and could be in commercial search engines as well.

Finding near-duplicate images and multimedia would be of value to those trying to find illegal copies of their copyrighted works on the Web. A key challenge will be to make these duplicate-detection techniques scalable, so that duplicates can be found within the hundreds of millions of other images and multimedia documents on the Web. Scalable text document-clustering techniques such as described in [2] may possibly be extended to solve these problems.

6 Conclusions

Along with the proliferation of multimedia information on the Internet has come tools for users to find the images, and more recently, the video and audio documents that interest them on the Web. Content analysis of image and multimedia has a role in multimedia search engines on the World Wide Web. Yet, because of the wealth of textual information associated with multimedia on the web and stringent performance requirements, content analysis has played a complementary role to other sources of information, and will probably continue to do so for the near future. Content analysis has been found to be useful for categorizing multimedia, such as speech versus music for audio, and photo versus non-photo for images. It also has a role in summarizing the multimedia, such as obtaining a representative keyframe of a video. With other cues, such as relevant text extracted from the linking HTML page and from the multimedia header, content analysis can also be used for finding similar images, video, or audio.

7 Acknowledgements

The work described in this review paper is that of a large number of dedicated individuals. WebSeer was designed and implemented in cooperation with Charles Frankel, Vassilis Athitsos, Bryan Sivak, and others at the University of Chicago. The AltaVista Photo Finder was designed and implemented by Nick Whyte, Charles Frankel, Venkat Raman, and others at AltaVista, in partnership with Virage. The video and audio indexing research prototype was the result of a team effort by Frank Bomba, Frederic Dufaux, Brian Eberman, Timothy Haven, Bob Iannucci, R. Paul Johnson, Chris Joerg, Leonidas Konthothanassis, Gabe Mahoney, Jim Oliver, Gene Preble, L.J. Ruell, David Rodal, Lou Romm, Mike Schexnaydre, Chris Weikart, Bill Wilder, and the author at Compaq's Cambridge Research Laboratory. Dave Goddeau, Pedro Moreno, and Jean-Manuel Van Thong from the CRL Speech group also contributed to this paper.

References

[1] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the 7th International World Wide Web Conference*, pages 379–388, 1998.

[2] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proceedings of the Sixth International World Wide Web Conference*, pages 391–404, 1997.

[3] M. L. Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 1998.

[4] F. Dufaux, B. Eberman, L. Kontothanassis, P. Moreno, M. Swain, and C. Weikart. A system for indexing web multimedia. Technical Report 99-3, Cambridge Research Laboratory, Compaq Computer Corporation, 1999.

[5] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. Technical Report 96-14, Department of Computer Science, University of Chicago, 1996.

[6] K. S. Jones, G. J. F. Jones, J. T. Foote, and S. J. Young. Experiments in spoken document retrieval. In K. S. Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 493–502. Morgan Kaufmann, 1997.

[7] M. J. Jones and J. M. Rehg. Statistical color models with applications to skin detection. Technical Report 98-11, Cambridge Research Laboratory, Compaq Computer Corporation, 1998.

[8] N. C. Rowe and B. Frew. Finding photograph captions multimodally on the world wide web. In *AAAI Spring Symposium: Intelligent Integration & Use of Text – Image, Video, and Audio Corpora*, pages 45–51, 1997.

[9] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–208, 1996.

[10] S. Sclaroff, L. Taycher, and M. L. Cascia. Imagerover: A content-based image browser for the world wide web. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

[11] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. Technical Report 1998-014, Systems Research Center, Compaq Computer Corporation, 1998.

[12] J. R. Smith and S.-F. Chang. An image and video search engine for the world-wide web. In *Symposium on Electronic Imaging: Science and Technology - Storage & Retrieval for Image and Video Databases V*, 1997.

COMPAQ

Searching for Multimedia on the World Wide Web

Michael J. Swain

CRL 99/1

March 1999