

Multimedia Search Technologies

Marie- Claude Lavoie

Department of Education
Concordia University, CANADA
m_lavoi@education.concordia.ca

Steven Shaw

Chief Learning Officer
EEDO Knowledgeware Corporation, CANADA
steven.shaw@eedo.com

Abstract

One of the most ubiquitous activities related to learning in the digital age is “search”. In recent years, computers have rapidly evolved from numeric and text processing to include multimedia, specifically audio, video, and images. However, few methods exist for searching multimedia, apart from text-based strategies operating on keywords, metadata and filenames.

Creating text descriptions for multimedia is resource-consuming and problematic. For example, sounds are difficult to describe in words and as a result, audio collections have been largely inaccessible. This paper will review the different technologies, including open source tools, which have been developed to facilitate multimedia searches. Many of these are based on some form of pattern recognition, and non-textual user input.

These tools, though not widely utilized in the world of learning, have tremendous potential to facilitate access to, and re-use of, multimedia, offering benefits to learners and instructional developers.

Introduction

One of the most important operations performed by computers is searching. There are many tools that aid in the search and retrieval of information but usually they treat only one source. The most obvious of these are Internet search engines. There are so many queries being done from sites such as Google, MSN Search, or Yahoo!, that browsers are now able to incorporate textboxes linked to search engines on their interface, thereby bypassing the step to navigate to the link in the first place. Purpose-built tools for more specific types of search have also emerged, such as search technology that operates on email.

A more recent tool combines several different types of search tools into one so that it is easier to include these into users' everyday workflow. Desktop search tools are being produced by companies such as Google and MSN Search. These are installed on a user's computer and are able to search. However, these tools are still not complete – as we will discuss later, searching text is not enough. Nor are they adequate in other respects. Current challenges include the burgeoning proportion of web pages which are dynamic and therefore not captured with conventional searches based on approaches using metadata or stable linkages. A dissertation by Pederson (2001) concluded that as of October 2000 there were 2.5 billion static pages published on the web and some 550 billion dynamic pages. His analysis suggests the most powerful commercial search engine, Google, can only retrieve

.1% of these documents – a serious decrease from figures published in an earlier study (Lawrence & Giles, 1999) that concluded all search engines combined could retrieve upwards of 16% of web published documents. Users have experienced this deterioration and can describe it anecdotally in terms of growing frustration seeking answers to questions. In response, some advances in the last couple of years include technology that does not merely search keywords, but that also parses large quantities of information to address queries more intelligently. At the forefront of this work is the Acquaint project (advanced question answering for intelligence) currently underway with the CIA, National Security Agency and other US federal intelligence organizations. Less sophisticated, but sometimes effective, current commercial innovations include familiar services such as Ask Jeeves (Ask.com) and Answers.com (a free version of GuruNet). Other sites, such as Clusty.com and Vivisimo.com provide a layer of categorization of the results of each search, facilitating question answering to some degree.

In recent years, the situation regarding search and retrieval has become even more critical, as computers have rapidly evolved from numeric and text processing to include multimedia - specifically audio, video, and images. Few methods exist for searching multimedia. Audio and video collections have been searchable only through text descriptions. That is, each audio or video clip is described in words by a human cataloguer, who types the description into the computer. These descriptions are then searched by keyword to locate clips of interest.

This paper will give examples of different technologies to aid in multimedia searches. It does not list every technology available nor does it provide in-depth explanations of how they work; what it does do, however, is provide a broad overview of what is possible.

Text Searches and Beyond

Searching text can be done easily using string matching or pattern recognition algorithms. Hence, it is easy to search files containing text - but how do you search multimedia such as digital photographs or music?

One possibility is, of course, attaching textual metadata to the multimedia files. The Dublin Core Initiative (dublincore.org) describes metadata as "...descriptive information about an object or resource whether it be physical or electronic... Library card catalogs represent a well-established type of metadata that has served as collection management and resource discovery tools for decades."

An important point to keep in mind is that a tool must be *appropriate* for the type of query to be completed. Creating text descriptions for audio and video is not only a burden costing time and money, but the value of such descriptions is limited. Sounds are difficult to describe in words (is it a bang, a crash, a thud?) and as a result, audio collections have been largely inaccessible. A more effective, albeit complex, method lies in abstract pattern recognition

Audio Search Technologies

One of the standard ways of searching for audio files is using ID3 tags that are, essentially, metadata header tags for files. The metadata information (such as title, date of recording, subject, or person) is not sufficient for the accurate and rapid retrieval of specifically requested data.

The open-source program, Find Duplicate Music Files (Rosenfeld, n.d.), or FDMF, is designed to identify music files that contain the same music, even if the files are differently named and contain different, or no, meta-information.

The program works by calculating the energy in four frequency bands for each one-

second chunk and determining the sum of the four bands as well as the ratio between each consecutive chunk so it can analyze the power spectrum. It then fits the power spectra to a fixed set of frequency points and then compares the two files' frequency charts.

Shazam Entertainment, based in the United Kingdom, offer a service built around a proprietary pattern recognition technology that can identify recorded audio even under noisy conditions (Shazam Entertainment, n.d.). It compares a thirty second clip sent by wireless device, such as a cellphone, to a huge database of songs from recording companies and sends back the information through SMS messages. This technology is being licensed in North America by AT&T and Microcell.

The New Zealand Digital Library's Web-based MELody inDEX (McNab et al., 1997) is designed to retrieve melodies from a database on the basis of a few notes sung into a microphone. MELDEX accepts acoustic input from the user, transcribes it into the ordinary music notation, and then searches a database for tunes that contain the pattern, or patterns similar to it. Retrieval is ranked according to how well the items match. This differs from Shazam's system because it does not require the query to be exactly the same as the file it is searching. Therefore, you are much more likely to get approximate matches. It is also much easier to browse in the MELDEX system.

FDMF is useful for detecting superfluous files and Shazam's and MELDEX's systems are great for identifying songs that keep getting stuck in you head. When considering audio files, there are several parameters we can choose to look at to increase the complexity of, and refine, the search, besides metadata, including voice, speech, musical instruments, and overall mood of a musical piece.

Biometrics research has lead to the ability to identify a specific voice based on a recorded sound clip. Usually, this is useful for forensic investigations but it can be employed to search recorded personal conversations or meetings. In theory, we could keep a voice clip of everyone in our address book which would allow us to search for a specific recorded meeting in which any given person was present.

One such project is underway with IBM's Conversational Biometrics Group, a part of the Human Language Technologies department at the Thomas J. Watson Research Center. They have created the Conversational Biometric Authentication in Real Time System (IBM Conversational Biometrics Group, n.d.), C-BART for short, that demonstrates how conversational biometrics may be used for voice-based personal authentication. Speaker recognition encompasses all the activities involving the identification of a speaker, based on his or her voice and the clustering of speakers based on similarities of their voices. This application is able to detect whenever a speaker changes, regroup speech segments spoken by the same person, and cluster speakers who speak similarly (e.g. same accent).

A harder parameter to search for is emotion in someone's voice. James Noble has decided to tackle this issue in his thesis presented to the Department of Computer Science at the University of Melbourne. Although not perfect by any means, the author suggests an algorithm that could differentiate acoustically between thirteen different emotions at an average success rate of over 60% when using a dependent speaker (Noble, 2003). Also, an accuracy of 23.6% was obtained when analyzing emotion independent of the speaker; this figure is only 11.4% less than that obtained by human judges. As Noble (2003) states, "this is an excellent result considering that humans have learnt from a lifetime of examples."

Besides using *how* someone talks, we can analyze *what* they are actually saying. Phonemes are the smallest unit of human speech and all utterances made in the entire world can be catalogued within a 400 phoneme range. By being able to identify these, researchers at Nexidia were able to devise a method which uses an open-vocabulary retrieval system using these phonemes instead of entire words (Clements et al., n.d.). Their Phonetic Search Engine is one of the most advanced speech recognition technologies as it reduces the time, and increases the accuracy of searches, against large collections of recorded speech. Searches can be conducted 100,000 times faster than real-time playback of recordings. This is one of the first technologies that analyzes sound instead of converting the speech to text and then searching that.

As for the general mood a piece of music can have, German developers at PW Soft have created a program called MoodMixer (PW Soft, n.d.) which can be plugged-in to music players such as WinAmp and automatically produce playlists that satisfy any mood. It uses different parameters, such as tempo, to determine songs suitable for any situation.

These were a few examples of technologies that have pushed the limits of searching audio. Another popular form of multimedia, however, is graphics and the boundaries in this realm are being pushed even harder.

Graphic Search Technologies

Graphic searches need to take into account several parameters to be effective, including text recognition, image composition, and facial and emotive recognition. They also need to differentiate between two-dimensional or three-dimensional images.

The Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst, has developed a system that retrieves actual handwritten pages given text queries. The system was based on the George Washington collection on catalogue the Library of Congress (Lavrenko et al., 2004).

Their goal was to allow easy searches of historical documents without having the high costs of transcribing their meaning first. In order to address the poor quality of some of the page images, the group chose a holistic word recognition approach that does not require character segmentation (Rath et al., n.d.). The reason for this is that most documents are severely degraded and segmentation of words into characters will produce very poor results.

The primary users of facial recognition software have been law enforcement agencies, which use the system to capture random faces in crowds. Visionics, a company based in New Jersey, is one of many developers of facial recognition technology. The twist to its particular software, FaceIt (Identix, n.d.), is that it can pick someone's face out of a crowd, extract that face from the rest of the scene and compare it to a database full of stored images.

In order for this software to work, it has to know what a basic face looks like. Facial recognition software is based on the ability to first recognize faces and then measure the various features of each face. Once a face is detected, the system determines the head's position, size and pose. A face needs to be turned at least 35 degrees toward the camera for the system to register it.

The heart of the FaceIt facial recognition system is the Local Feature Analysis (LFA) algorithm. This is the mathematical technique the system uses to encode faces. The system maps the face and creates a faceprint, a unique numerical code for that face. Once the system has stored a faceprint, it can compare it to the thousands or millions of faceprints stored in a database.

Besides recognizing faces, software can also recognize emotions. The affective social computing group at the University of Central Florida has created a program with the ability for facial expression recognition. They believe that there are six basic universal facial expressions: these are identified as fear, anger, surprise, disgust, happiness, and sadness (University of Central Florida, n.d.). For their research, they collected numerous facial images of many different individuals of different gender, ethnicity, and age groups, exhibiting different expressions of emotions including the six universal ones. Once a pattern among facial expressions is found, the next natural step was to ask computer programs to identify them in subjects (Nasoz et al., 2003).

ImgSeek (Niederberger, n.d.) is an open-source photo collection viewer and manager with many features. The most interesting of these is the ability to submit queries expressed as a rough sketch painted by the user. Figure 1 shows the user interface and one such query. The searching algorithm makes use of multi-resolution wavelet decomposition of the query and database images based on an idea by a research team at the University of

Washington (Jacobs et al., 1995). Another interesting feature is the ability to submit a specific picture for which the software then is able to find similar pictures. For example, if the user submitted a picture of sunsets, the software would return all other sunsets in the database.

Other types of images that can be searched are three-dimensional (3D) images. The Princeton Shape Retrieval and Analysis Group have created a tool for shape-based queries (Princeton Shape Retrieval and Analysis Group, n.d.). It allows the user to draw one or more 2D shapes with a pixel paint program and then have the system match the resulting image(s) to 2D projections of 3D objects. For example, in the figures below, the user has drawn outline contours specifying a shape, and the system has returned a set of matching objects.

Engineers at Purdue University are developing a similar system (Venere, 2004.). However, instead of allowing for a portal-like search of 3D objects they enable engineers to search for parts in industry databases by sketching the part from memory, penciling in modifications to an existing part or selecting a part that has a similar shape.

Video Search Technologies

Although video can be thought of as graphic images in subsequence with audio, searching video can identify motion and scenes.

A joint research project between Columbia University and two commercial stock footage companies, Hot Shots Cool Cuts, Inc., and Action Sports Adventure, Inc., has been undertaken to evaluate new content-based video search tools (Image and Advanced Television Lab at Columbia University, n.d.). VideoQ expands the traditional search methods (e.g. keywords and subject navigation) and allows users to search video based on visual features and spatio-temporal relationships (Chang et al., 1997.). For example, VideoQ can use spatio-temporal description of a video object as the basis of a query. The user assigns the duration,

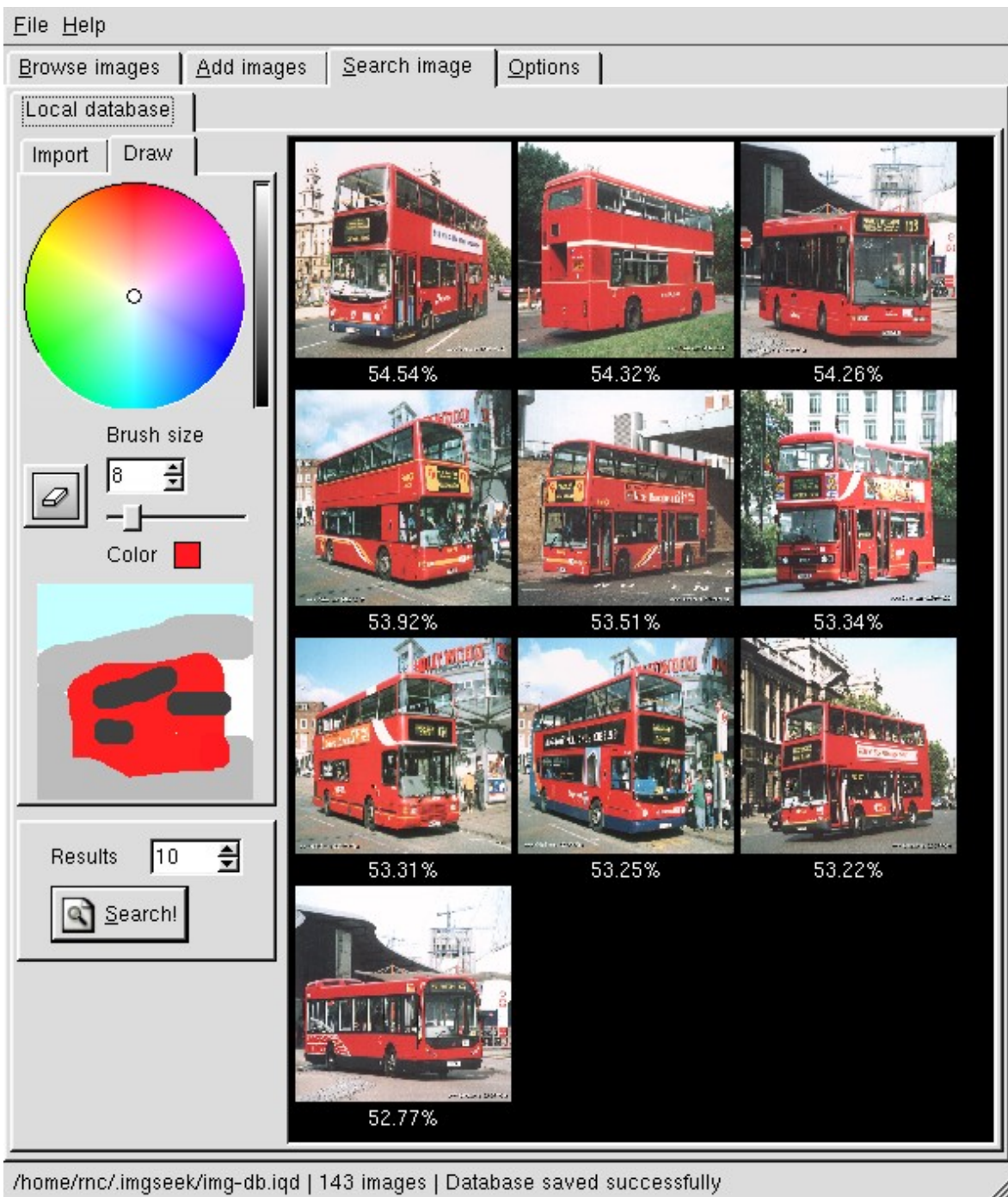


Figure 1. A query expressed as a rough sketch painted by the user.
(Source: <http://imgseek.sourceforge.net/sshot/>)

and an arbitrary trajectory to a video object and VideoQ will look for an object moving in the specific direction. Users can also specify colors and textures as well as shapes to the objects they are describing.

The KIA Project (Sundaram & Chang, 2000) looks at different aspects of video files based on production techniques. It uses a video skim to create a shortened version of the video. Using various pattern recognition algorithms, the shots and the corresponding audio are selected. The algorithm can detect shots with faces, scene boundaries, and audio structures, such as dialogues. Overall, they determine relationships between descriptive complexity of shots and decoding the time for that shot.

Conclusion

The future of search will not be limited to either centralized repositories, as in the present paradigm for enterprise content management, or to text-based categorization and search via metadata or string-matching. Instead, searches will encompass multiple repositories, desktops and other sources such as email and recorded audio-conferences. A variety of emerging tool, as described in this paper, will expand our ability to search multimedia components, independently of any manual or automated approach to tagging these elements with text (metadata). This will have an enormous impact on the power of individuals to search for content, and on fields that are, or can be, particularly dependent on the study, analysis or manipulation of images or sound. The list would include, but is not limited to cultural studies, architecture, history, anthropology, and art history. In recent years there has been a tremendous growth in the use of media and media manipulation across a variety of fields. We have provided a sample of technologies available to search audio, graphic, or video. It is not, by all means, an exhaustive list, but it does give an overall view of the possibilities in multimedia search beyond metadata tagging and text searches. The future of search will be shaped very much by these tools, as well as by advanced technologies for categorization that allow for the creation of user-centred views, and the expansion of ontological representations, such as Topic Maps (ISO, 1999).

References

- Center for Telecommunications Research at Columbia University, n.d. The KIA Project [online]. Found at: <http://www.ctr.columbia.edu/video-summary/index.htm> [April 7th, 2005].
- Chang, S., et al., 1997. VideoQ: An Automatic Content-Based Video Search System Using Visual Cues. In: Association for Computing Machinery (ACM) Multimedia Conference, November 9-13th 1997 Seattle.
- Clements, M., et al. n.d. Phonetic Searching of Digital Audio. Fast-Talk Communications, Inc. Found at: <http://nexidia.com/whitepaper.asp> [April 7th, 2005].
- Dublin Core Metadata Initiative, n.d. What is Metadata? [online]. Found at: <http://dublincore.org/resources/faq/#whatismetadata> [April 7th, 2005].
- Eronen, A., and Klapuri, A., 2000. Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 9th 2000 Istanbul, pp. 753-756.
- IBM Conversational Biometrics Group, n.d. Conversational Biometric Authentication in Real Time (C-BART) [online]. Found at:

- http://www.research.ibm.com/CBG/CBART_Demo.html [Aprilth, 2005].
- Identix, Inc, n.d. FaceIt ARGUS [online]. Found at:
http://www.identix.com/products/pro_security_bnp_argus.html [Aprilth, 2005].
- Image and Advanced Television Lab at Columbia University, n.d. VideoQ [online]. Found at:
<http://persia.ee.columbia.edu/VideoQ/.index.html> [Aprilth, 2005].
- Jacobs, C.E., Finkelstein, A., and Salesin, D.H., 1995. Fast multiresolution image querying. In *Proceedings of the 22nd Annual Conference on Computer Graphics and interactive Techniques* S. G. Mair and R. Cook, Eds. SIGGRAPH '95. ACM Press, New York, NY, 277-286. Found at: <http://portal.acm.org/citation.cfm?id=218454#> [April 7th, 2005].
- Niederberger, J., et al., n.d. ImgSeek [online]. Found at: <http://imgseek.python-hosting.com/>
- ISO, 1999. ISO/IEC 13250 Topic Maps. Found at:
www.y12.doe.gov/sgml/sc34/document/0129.pdf.
- Lavrenko, V., et al., 2004. Holistic Word Recognition for Handwritten Historical Documents. In: Document Image Analysis for Libraries (DIAL), January 23-24th 2004 Palo Alto, pp. 278-287.
- Lawrence, S & Giles, L., n.d. Accessibility of information on the web. *Nature*, 400, July, 107-109.
- Lou, K., et al., 2004. Content-based Three-Dimensional Engineering Shape Search. In: 20th International Conference on Data Engineering (ICDE), March 30 - April 2 2004 Boston, pp. 754-765.
- McNab, R., et al., 1997. The New Zealand Digital Library: MELody inDEX. *D-Lib Magazine*, May 1997. Found at: <http://www.dlib.org/dlib/may97/meldex/05witten.html> [Aprilth, 2005].
- Nasoz, F., et al., 2003. Emotion Recognition from Physiological Signals for Presence Technologies. *International Journal of Cognition, Technology, and Work - Special Issue on Presence*, Vol. 6(1).
- Niederberger, R., n.d. imgSeek - Intelligent Image Database [online]. Found at:
<http://sourceforge.net/projects/imgseek/> [Aprilth, 2005].
- Noble, J., 2003. Spoken Emotion Recognition with Support Vector Machines. Honours thesis (PhD). University of Melbourne. Found at:
<http://eprints.unimelb.edu.au/archive/00000497/> [Aprilth, 2005].
- Pederson, A., 2001. Semi-automated web discovery and analysis: an approach based on interactive machine learning principles. Dissertation. Agder University, Norway.
- Princeton Shape Retrieval and Analysis Group, n.d. Princeton 3D Model Search Engine [online]. Found at: <http://www.cs.princeton.edu/gfx/proj/shape/> [Aprilth, 2005].
- PW Soft, n.d. Moodmixer [online]. Found at:
<http://www.moodmixer.com/Englisch/Hauptseite.htm> [Aprilth, 2005].
- Rath, T., et al., n.d. Handwriting Retrieval Demonstrations. University of Massachusetts Amherst Center for Intelligent Information Retrieval. Found at:
http://ciir.cs.umass.edu/~trath/prj/hw_retr/demo_intro.html [Aprilth, 2005].
- Rosenfeld, K., n.d. Find Duplicate Music Files [online]. Freshmeat. Available from:
<http://freshmeat.net/projects/fdmf/> [April 7th, 2005].
- Shazam Entertainment, n.d. The Shazam Technology [online]. Available from:
<http://www.shazamentertainment.com/technology.shtml> [Aprilth, 2005].
- Sundaram, H. and Chang, S., 2000. Determining Computable Scenes in Films and their Structures using Audio Visual Memory Models, *ACM Multimedia 2000*, Oct 30 - Nov 3,

Los Angeles, CA.

University of Central Florida, n.d. Affective Social Computing Laboratory at the Facial Expression Recognition. Found at:

<http://www.cs.ucf.edu/~lisetti/research/emotionrecognition.html#fer> [Aprilth, 2005].

Venere, A., 2004. Purdue Engineers Design 'Shape- search' for Industry Databases. Purdue News March 30th 2004. Found at:

<http://news.uns.purdue.edu/UNS/html4ever/2004/040330.Ramani.shape.html> [Aprilth, 2005].