Multimedia Information Extraction Roadmap

Thom Blum

Muscle Fish, a division of Audible Magic Corporation
2550 9th Street, Ste. 207-B, Berkeley, CA 94710 USA
thom@musclefish.com
http://www.audiblemagic.com

Efforts to automatically extract salient features from rich media documents (audio, video, still images, speech, etc.) have been underway and maturing for the last 15 to 20 years; longer than that if you consider speech recognition research. However, there is still a paucity of real world applications that leverage the technology, and scaling the technology for situations "outside the lab" have posed mighty challenges. (This is not the case, of course, with the technologies and applications developed by my company, Muscle Fish, and its parent company, Audible Magic Corporation.) In general scalability, reliability/robustness, and accuracy have presented formidable obstacles to finding, and then reaching, a market.

The following lists sketch what I believe are the major technical challenges and gaps in state of the art multimedia information extraction (MMIE), as well as existing approaches and resources that can be quickly tapped for research and development.

Critical Technical Challenges

What are the critical technical challenges in multimedia information extraction (MMIE)?

- Source separation in audio (and perhaps other media), enabling:
  - audio "scene analysis" - identification of layered sounds within an audio "scene"
  - musical semantics - understanding the musical structure, on multiple levels, of music input.
- Reliable and robust video feature extraction, enabling:
  - query, comparison, and identification of video
- *More* reliable and robust speaker independent (and speaker dependent) speech recognition
- We need to think of new, useful and interesting applications that leverage the information being extracted. New capabilities of extracting and querying MM info are fantastic, but applications that "put the data to work" are lacking and are vital to continued efforts. MMIE is still, to some extent, a technology in search of applications, and that poses a challenge to all who work in this field; albeit a "creative," more than a "technical," challenge, but a challenge nonetheless.

Existing Approaches

What are the important existing methods, techniques, data and tools that can be leveraged?

DISCLAIMER: The following is a very partial list of existing resources – mostly, open source – to assist with the task of multimedia information extraction, retrieval, identification, etc. I do not vouch for these resources, as I have not explored them in depth. I can say that many of the resources, below, do enable one to get started in the field relatively quickly, but the smattering of personal experience that I've had with them lead me to the conclude (perhaps too quickly) that they suffer from the scalability, accuracy, and reliability issues that I mentioned in the "Critical Technical Challenges" section, above. Also, I am not writing here to promote the technologies of my own company, Audible Magic Corporation. Suffice it to say, the reader may want to study patents and documentation available on our Web site, at http://www.audiblemagic.com/company/patents.asp.

Methods and Techniques

This is a partial list of audio/speech information extraction APIs and other resources which are all (or mostly) open source.

Audio Fingerprinting
MusicIP - http://www.musicip.com
MusicBrainz – http://musicbrainz.org and http://wiki.musicbrainz.org
Relatable – http://www.relatable.com/tech/trm.html
MusicURI – http://semedia.deit.univpm.it/musicuri

Speech Recognition
Sphinx (Carnegie Mellon University) - http://cmusphinx.sourceforge.net/html/cmusphinx.php

DataSets

Online Music Databases
http://en.wikipedia.org/wiki/List_of_online_music_databases

Speech Models
http://www.speech.cs.cmu.edu/sphinx/models/

Tools

See above, plus APIs, song and video databases, identification services, etc. from Audible Magic Corp at
http://www.audiblemagic.com

See also, various video and audio logging, tagging, and retrieval applications and tools available from Virage (an Autonomy Systems Limited company)
http://www.virage.com

Remaining Gaps

What key technology gaps remain that require focused research? [If possible, forecast when you believe these gaps will be filled in terms of 1 to 5 or 10 years in the future]

- "Scaling up" for real world applications of MMIE:
    - moving beyond "the lab" (generally small databases of "knowns" against which to test retrieval accuracy)
- Perhaps "standardized" input data test sets (i.e., collections of known sources for movies, broadcast news, music, print news, etc.) are needed:
    - benchmarks for retrieval against these sets
- Source separation - decomposition of "scene" components, for example:
    - recognition of multiple layers/levels/components within a given medium (e.g., the component sounds/layers within a natural soundscape, analysis of foreground/background within a video, multiple audio "samples" – that is, fragments from existing songs – when played simultaneously or overlapped, etc.)