The MIR 2005 Panel

Multimedia Information Retrieval: What is it, and why isn't anyone using it?

Alejandro Jaimes[1], Mike Christel[2], Sébastien Gilles[3], Ramesh Sarukkai[4], and Wei-Ying Ma[5]
[1]FXPAL Japan, Corporate Research Group, Fuji Xerox Co., Ltd., Japan
[2]Carnegie Mellon University, USA
[3]LTU Technologies, France
[4]Yahoo Inc., USA
[5]Microsoft Research, China

ABSTRACT

In this paper, the participants of the panel at the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval answer questions about what multimedia is, how MIR is different from other kinds of retrieval, the most important technical challenges in MIR, killer applications, opportunities, and future directions.

Categories and Subject Descriptions

H.3.3 [Information Search and Retrieval]: Retrieval models.

General Terms: Algorithms, Management, Performance, Design, Human Factors, Theory.

Keywords: Multimedia Information Retrieval, Video Retrieval.

1. INTRODUCTION

Multimedia Information Retrieval has become one of the most active sub-fields of multimedia research. While progress has been significant in research, however, there has been little progress in the development of applications for widespread use.

The purpose of this panel is to get an overview of the main research issues, research and market directions for Multimedia Information Retrieval.

2. PANEL QUESTIONS

2.1 The term Multimedia is currently used to refer to different types of media (images, audio, video, and even text). How do you define multimedia and do you think the term should encompass different independent media?

Mike Christel. Multimedia is combining different media, e.g., text, imagery, video, animation, and sound, into one application, presenting these multiple media in an integrated way to communicate a message. Video is interesting because by itself it can be considered a multimedia presentation. Consider a news broadcast that includes text overlays identifying people and place names, images of maps and photos, video sequences of studio anchorperson narration, scene settings, and interviews, and sounds of human speakers as well as environmental sounds from field reports. Hence, video processing tools tend to work across media elements represented in the video and so by definition are multimedia tools, and in my discussion I will focus on video information retrieval, as that is the area that I have researched over the past 15 years.

Wei-Ying Ma. I would consider text, image, and audio (speech) as media, and video and any other presentation that contains more than one type of media as multimedia. In a broader sense, a web page with both text and image can be considered as multimedia.

Ramesh Sarukkai. Multimedia should refer to any media that encapsulates multiple "independent" information attributes within the source data. Thus video segments have visual attributes, audio features, text/meta-content and hence multimedia. Scanned documents are composed of visual images and meta-data text content, and hence multimedia in that sense. In this presentation, I will focus on video, and argue that the same fundamental models are applicable for all media.

Sébastien Gilles. Literally, "multimedia" means "several media". As with most words, several definitions can be worked out, depending on context. As a noun, the term designates different media types collectively. As an adjective, the term can qualify the structure of a document: a multimedia document really is an electronic document containing different, non-independent, intertwined media. But when used to qualify a device or software ("multimedia PC", "multimedia search engine"), the term qualifies the data types supported by the device or software. So "multimedia" is also for qualifying a device, apparatus or software application that manipulates, processes or displays different media transparently for the user.

2.2 How is Multimedia Information Retrieval (MIR) different from retrieval of non-multimedia information?

Wei-Ying Ma. I think one of the challenges for multimedia information retrieval is a simple but effective way of forming a query. In CBIR, query-by-example has been used as a method to search image databases. However, in many real world applications, it is hard to find an example to describe the user's information need. Because it is more intuitive to use keyword (text) to describe information need, most of current commercial image search

MIR'05, November 10-11, 2005, Singapore.
ACM 1-59593-244-5/05/0011

engines are text-based. Another challenge for MIR is how to automatically annotate multimedia using text (or pre-defined vocabulary) so that users can use text to search multimedia. Because this problem is unlikely to be solved in a near term, researchers have been trying to develop all kinds of new solutions to work around this issue.

Mike Christel. MIR can leverage from correlated sources of information. MIR can take advantage of input streams of data that often are highly redundant within small time windows but where the temporal element can be used to refine processing, e.g., guesses on where text appears in a video stream can become more accurate with the processing of a series of video frames rather than processing a single image. MIR when dealing with visual or aural information deals with data that humans directly sense with their eyes and ears, and so can allow for approaches that open up the multimedia data for more intuitive inspection by the user than is possible with non-multimedia information like text. For example, videos from multiple synchronized cameras can be used to generate three-dimensional representations of sporting events or surveillance of an area from multiple angles.

Sébastien Gilles. First, text needs to be singled out from other modalities. Since non-textual media have no alphabet, their content is not explicit, so intermediary low-level features capturing the relevant information have to be computed first.

Then, the difference between MIR and non-multimedia retrieval is that MIR comprises another layer of complexity: fusion. Multimedia to multimedia search requires data fusion at some point, during either feature, metric or ranking computation.

Ramesh Sarukkai. One of the key differences in multimedia information retrieval is a higher level of "perceptual gap" between end user consumption of the media versus current system analysis of the data. In text retrieval systems, this gap is limited due to lesser dimensions of ambiguity.

2.3 Can you provide some examples in which searching for multimedia data requires more than one modality?

Sébastien Gilles. People can only become accustomed to something they have been exposed to. Today, very few, if no popular web service are offering true multimedia search. Rather, all retrieve multimedia using text. The great advantage of text versus other- or multi-media search is that users are not restricted to queries by-example, which are impractical when no example is available at hand.

This being said, the need for multimedia queries exists in specialized domains (police/defense forces, librarians, researchers, health care, etc.) where users are interested in a more through level of analysis, because they are after specific visual and audio cues.

For instance, intelligence/forensics people might be interested in this query: "show me all videos containing parts similar to my sequence and with gun shot noises in the audio track". In medical imaging, a practitioner might want to query a medical video database with an ultrasound heart video, search the database for heart sequences with similar movements and noises, and seek to view the corresponding diagnoses.

Mike Christel. Aural and visual modalities are both needed to find reaction shots of a politician's face when there is jeering in the audience, scenes showing what people are doing in the streets when a bomb explodes, or video sequences showing two or more children together laughing. When the visual modality is the target, synchronized alternate modalities often improve retrieval effectiveness over what is possible with just the visual channel. For example, text overlays and news reporter narration identify shown people in news footage, and home team scoring events in sports sequences can be found when crowd cheers rise dramatically. Also, integrating cues from multiple modalities can improve multimedia summarization interfaces. For example, video skims are more successful when they assemble visual sequences by breaking during silence, and fail completely when audio no longer matches the visuals: aural processing improves the skim presentation [2].

Wei-Ying Ma. One example is using a camera-phone to take the picture of a real world object (e.g. a building or physical store) for which we would like to find information on the Web. The picture together with a few words describing the user's information need can be jointly submitted to a search engine, and then the engine returns web pages containing information about that real world object in the right context. This multimodal query (photo + keyword description) retrieves web pages that are more than one modality based on both image and text similarity.

Ramesh Sarukkai. One reason for this current limitation is the client device capability itself. As devices (e.g. wireless) get more sophisticated, many multi-modal input opportunities arise. For example, we can envision an application of image matching technology when a user takes a picture of an object on a mobile phone and searches for objects that are similar to that image, refining the search based on audio descriptions.

2.4 What are, in your opinion, the three most important technical problems in multimedia retrieval (and suggested directions)?

Mike Christel. First, how can we address the semantic gap between low-level features and high-level user information needs for MIR, when the corpus is not well structured and does not contain narration audio documenting its visual contents? As Alex Hauptmann notes in his CIVR keynote talk [4], and echoed by others (e.g., see [6]), MIR researchers have successfully harvested low-hanging fruit: clever tricks and techniques of using speech transcripts and broadcast genres with detailed well-understood structures to identify the contents of news and sports broadcasts. The challenge now is to transform these techniques "into a serious body of science applicable to large-scale video analysis and retrieval" [4]. Some directions include inferring media content from spatio-temporal context and the social community of media capture [3], the Large Scale Concept Ontology for Multimedia (LSCOM), understanding work to reliably detect hundreds of intermediate semantic concepts like face, people, sky, and buildings across corpora [4], and working with less structured collections rather than just news or sports [6]. A second key problem for multimedia retrieval against video is demonstrating that techniques from the computer vision community scale to materials outside of the researchers' particular test sets. I strongly believe in the value of community benchmarking activities like NIST TRECVID which support the statement from the 2003 ACM retreat report that "repeatable experiments using published benchmarks are required for the [MIR] field to progress" [7]. Third, how can we best leverage the intelligence and goals of human users in accessing multimedia contents meeting their needs, rather than overwhelming them with exponentially expanding amounts of irrelevant materials? Directions include applying lessons from the human computer interaction and information visualization fields and being pulled by user-driven requirements rather than just pushing technology-driven solutions.

Sébastien Gilles. I guess the number one technical problem is the semantic gap, for instance in image retrieval: from low-level image features, how do we infer image semantics? We know that abstract keywords are almost impossible to correlate to image content. Descriptive keywords, such as object names can be successfully inferred to a certain extent with classification and case-based reasoning techniques.

Equally, the issue of generalization poses some serious challenges. The high dimensionality of multimedia features makes the feature space very sparse, leading to generalization errors. A rule of thumb is to use as much training data as possible, but in some applications, training data is itself sparse. This advocates the use of reinforcerment-learning systems, that can cope with concept drifts (eg: spam filters).

Finally, search speed and scalability is a central issue in projects such as national digital libraries, but also in personal multimedia. Distributed computing is a solution to this problem in 3-tier architectures (e.g: search engines). For personal multimedia, the question is purely algorithmical: design low search-complexity MIR algorithms.

Ramesh Sarukkai. The three most important problems:

a) Proper modeling & application of contextual constraints.

b) Dynamic selection & fusion of the right modalities for inference.

c) Enabling & leveraging large scale media meta-content.

The suggested approach is to move away from a "one model fits all" approach to multimedia information retrieval. For instance, let us take the problem of video retrieval: do we rely on meta-data, visual or auditory features as the dominant factor? Rather than have a fixed model for retrieval, having a more dynamic, and flexible model can better leverage the different technologies: If the source is a mobile device with geo location, then that takes precedence to visual features. If the source is a sports media production source, then visual attributes (e.g. motion cues) are weighted more based on this context.

Wei-Ying Ma. The three most important problems:

1. Large scale image classification (e.g. able to classify images into more than 1000 categories that are defined based on the analysis of image search log). The needed labeled image data can be potentially obtained from the Web. Researchers also need to think of a clever way of leveraging the collective efforts from web users to annotate the photos. The learning algorithms should also leverage the large amount of unlabeled data on the Web.

2. Relevance feedbacks.

3. UI for presenting multimedia retrieval results.

2.5 Do you think there will be a killer app in multimedia, a particular domain which drives multimedia retrieval, or will it simply (incrementally) become integrated into multiple domains and applications?

Mike Christel. The killer functionality which crosses domains and applications is transforming our capability to produce and store massive amounts of multimedia materials into a benefit. How can vast stores of photographs, audio, and video clips become accessible so users are not overwhelmed by all the data, but instead can transform the data into personally meaningful information and knowledge? How can users ideally contribute back into the data repository both passively (e.g., their interactions) and actively (e.g., a "folksonomy" of annotations) additional metadata that makes the multimedia repository more valuable? "Value" encompasses *effectiveness* (can I get the right stuff), *efficiency* (can I get it quickly), and *satisfaction* (is it enjoyable/worth my effort/would I do it again). The purpose of high value multimedia repositories varies from training to entertainment, with compelling goals including the communication of a sense of time travel and digital immortality.

Ramesh Sarukkai. The makings of the killer application is already here: Web Based Video Search. This content will flow into wireless devices, the digital home and transform the evolution of digital media (production and consumption). The key domain that will alter the nature of this space is aggregation and application of very large volumes of data: this data can be acquired from end users, or inferred. Other implicit attributes can be leveraged. This is akin to the breakthroughs in speech recognition where large amounts of acoustic and statistical language model text data were aggregated and applied to have a tremendous impact.

Wei-Ying Ma. Video search has a huge potential as it is tightly related to TV advertising. Video search also needs new techniques to enable easy navigation of search results (e.g. hierarchical video browser) and a more scalable way of streaming video over the Internet as the user may be searching for a specific segment of video instead of the entire video sequence. As more photo forums are appearing on the web, a community type of search for these photos could be interesting. Also, as the web search is moving towards vertical areas. There could be a similar opportunity for multimedia search as well.

Sébastien Gilles. Multimedia will probably roll-out several killer applications, because the domain is so vast and exciting.

Personal multimedia is driving innovation today. Strong innovation also comes from specialized domains (defense, medical, etc.). The bottom line is that multimedia content volume keeps increasing at a faster rate while human capacity for digesting information remains fixed. So there is a throughput problem, which is a good condition to witness a technological breakthrough involving MIR.

With soon-to-come real-time interactive mobile technologies, new ways for social, personal or business interactions will emerge, creating opportunities to see killer applications.

At present, a natural positioning for MIR technology is as a "technological brick" because it is a horizontal technology useful to many market verticals. Killer applications that blend multimedia with other great technologies (geo-localization, P2P, etc.) can also see the light in this context.

2.6 Certain conditions must be met for a scientific discipline to flourish. This includes, among others, a theoretical framework, measurable (and duplicable) results. Where do you see the state of multimedia information retrieval in this respect?

Sébastien Gilles. Much remains to be done with respect to the theoretical framework. MIR is a cross-discipline area, federating

efforts in formerly independent domains (text, image, audio). Each research community has developed nice theoretical frameworks but no unifying framework for MIR really exists. This might also be due to the lack of real-world data made available for research.
Regarding the repeatability and measurability of results, initiatives like TRECVID will be beneficial to the MIR community, like TREC has been to the text community or FERET to the face recognition community: the development of benchmarking programs drives research forward by setting a common measure tool and fostering emulation.

Wei-Ying Ma. Similar to IR research, MIR also needs evaluation methods and a standard benchmark so that different techniques can be compared.

Mike Christel. As mentioned earlier, I see great value in community benchmarking activities like TRECVID (see [5]), with the need to expand such benchmarks into representative corpora beyond those having great structure and a correlated text transcript like broadcast news and sports. One example I am currently working on is dealing with surveillance-style video in nursing home environments to promote better quality of care and quality of life, but the difficulty of course is dealing with privacy and intellectual property rights to share such data for benchmarking. In the absence of community repositories and benchmarks, it is too tempting for a researcher to work on a carefully crafted corpus (e.g., Corel professional image subset) and report outstanding results that are not achievable in representative domains (e.g., home photos and movies).

Ramesh Sarukkai. We have pretty good training and evaluation sets from initiatives such as TRECVID. The computer vision, audio analysis, and pattern recognition technologies are fairly well researched and have a number of techniques that show promise in restricted domains/applications. On the theoretical front, more work needs to be done to model the effects of fusion of different, varied information sources for video, especially taking into account very large volumes of data.

2.7 MIR is inevitably a human activity (multimedia is produced by humans and accessed by humans). What should be the role of the user and what techniques do you see as being the most important?

Mike Christel. Back in 1999 Shahraray noted that "well-designed human-machine interfaces that combine the intelligence of humans with the speed and power of computers will play a major role in creating a practical compromise between fully manual and completely automatic multimedia information retrieval systems" [1]. We need to incorporate the intelligence of the user through HCI techniques, and information visualization strategies, with my research work [8] very much focused on this theme.

Sébastien Gilles. MIR is a computer-assisted human activity. In particular, an MIR system should not be seen as a decision-making system but rather as a smart search tool operated by a user. Thus, the user has a central role in MIR.
In terms of interaction, there is no best technique, but rather a best technique given a user context. So the system should offer a range of tools for interacting with it, that the user can choose to use when needed. Navigation features such as browsing are useful but clearly query by-text and query by-example are key to a good MIR system. Tools for selecting/pasting document parts or objects are also very important to access to inside-document content.

To best incorporate human activity parameters, we must build models of human activity in the context of MIR. Relevance feedback is interesting because it can specifically model short-term memory when searching for information, but long-term memory, user profiling and subjectivity are best addressed with an explicit user preferences mechanism.

Ramesh Sarukkai. For web based media search, the user is central to the retrieval task. Thus, we should leverage appropriate techniques such as relevance feedback, user models, and memory wherever applicable.

Wei-Ying Ma. I believe relevance feedback will become very important for MIR, even for commercial media search engines. MIR is different from web page search because of the challenge of formulating an unambiguous query. I believe an effective media search engine will be an interactive one. It should allow the user to mark the returned objects as positive or negative in order to tell the engine more about what the user is looking for.

2.8 The camera-phone phenomenon is having a huge impact on the creation of content. Can you identify a technology likely to have a strong impact on MIR?

Wei-Ying Ma. Any technology that facilitates the instant and seamless sharing of photos and video across the Internet and devices will have a huge impact on MIR.

Mike Christel. I believe that Google, Yahoo, and others' search interfaces for web-based image and video materials working with text metadata shows that for a huge collection of materials, when looking for a single instance of something (currently images, soon video), you have a good chance of finding it in the top 20 returned items. Someone somewhere annotated the containing web page and/or image/video with enough text to allow it to be retrieved; this approach leverages successfully the collective effort of millions of internet contributors. When users want precision of at least one relevant item in the top 20 images/video clips, they will see MIR as solved (with the solution provided by virtually ignoring all cues outside of the text modality). A problem that remains for MIR researchers is addressing those cases where greater recall is desired, e.g., finding most/all instances of a person falling in a nursing home setting, or most/all people seen with a suspect under surveillance.

Sébastien Gilles. Broadband and cheap access cost will definitely drive MIR innovation forward. Indeed, since mobile phone network architectures comprise a central MMSC server connected at fast rate to mobile terminals, a large transmission speed means that mobile terminals can virtually have the same multimedia processing power as a MMSC server. Today, UMTS is just emerging, but real-time mobile interactive multimedia applications will become a reality, and should be a very exciting domain. This should also be a major technical and societal revolution. As a consequence, means for searching multimedia information will also be radically changed.

Ramesh Sarukkai. Mobile phones that are all-in-on digital camcorders, handheld computers, and intelligent communication devices are changing the production of media. Consumption is still not easy and it's not clear where the real breakthrough lies: in the device technology, user interface, or application. Digital Home devices integrated with media information retrieval capabilities will have a clear impact on the consumption of media in a highly networked environment. The publish, share, subscribe, and download to your favorite device style of consumption will also have a strong influence on video search.

2.9 Many social and legal issues can be addressed with the aid of technology, e.g., standards and Digital Rights Management for privacy and copyright protection. Are we doing enough work in these areas to make MIR a success in real-world applications?

Mike Christel. I don't do any research into this area, and so will simply point the audience to MPEG-21 and DRM schemes in web video players for additional information.

Sébastien Gilles. Initiatives and standards for inter-operability between networks (phone, internet) and devices have been key to the development of a true "multimedia network". This was the first step: enabling multimedia content circulation with good connectivity. The second step is to get users and content providers to trust this network. Likewise secured payment has boosted online shopping, secure content distribution and repurposing, as well as user privacy protection techniques should boost MIR applications. These domains are generally pioneered by companies which are along the content production chain (from production to distribution).

Wei-Ying Ma. The success of web search today does not rely on standards and DRM. It's driven by the growing online advertising business. I think it also applies to MIR. As long as we could build a successful business model around it, MIR does not necessarily need more standards or better DRM to make a success in real world applications.

Ramesh Sarukkai. As a media research community, we are not doing enough work to make MIR a real success in real-world applications. This is mainly because while standards come first, technology follows later with market adoption. As long as content producers don't leverage these standards, the value of building out compatible systems immediately is low. However, with market growth, application trends in supporting the standards will follow rapidly.

2.10 Why are we not seeing, yet, true multimedia (content-based) retrieval and when do you think MIR will really hit the masses?

Mike Christel. With respect to visual, aural, and text modalities, we can do text IR relatively well and thus far have defaulted MIR to make use of text IR wherever possible. Visual and aural automated processing techniques are difficult and hence researchers often begin with tuned test data sets where results do not generalize. Some notable exceptions include face detection which has worked well across different corpora. By capturing more and better metadata at creation time, using folksonomies and aggregated usage statistics, developing intermediate semantic concepts through LSCOM efforts, and rewarding research working with representative benchmark sets, we can slowly migrate MIR toward better performance levels enjoyed by text retrieval. Given the difficulty of automating MIR for general corpora, the role of the human user to guide and filter retrieval should be optimized. MIR will really hit the masses when it lets us see our growing multimedia data sets as useful assets rather than information glut.

Sébastien Gilles. The fact is that text-based search (for instance when combined to automatic speech recognition and close captioning) is particularly well-adapted for supporting searches about named entities and proper names, a very popular type of searches with internet users. Major search engines also favour this approach because it enables them to leverage existing architectures for text-based search.

MIR is a promising and exciting domain of research and industry, with many technical challenges ahead. It could well hit the masses within 5 or 10 years when broadband will be fast and cheap enough to enable real-time mobile interactive multimedia interactions for millions of users.

Ramesh Sarukkai. The trend we are seeing is that media is transformed through information layers that add/generate more data (whether it be content-based, community-based or other media based sources): thus media generated a few years from now will be lot more than just the video/audio data. We are clearly seeing the standards (such as MPEG-7) trending towards that. With this fundamental shift at the media production sources, a migration towards augmenting the meta-data based systems with more content analysis techniques applied in a contextual manner will follow. I am optimistic that 2007 will be the year of significant application of content-based media retrieval to improve meta-data based MIR systems.

Wei-Ying Ma. General internet users are unlike us. They are very inexperienced in search, so the media search technologies for the mass have to be very simple, intuitive, and easy to use. Current content-based retrieval techniques are still too complex to use. Text-based method still works best because it is the most natural way of describing the information need. I think we are going to see some major breakthrough in MIR in the next 10 years, but the breakthrough may not come directly from the technologies. MIR may hit the mass because of a simple but yet effective integration of media sharing, annotation, search, and management across desktop and the Web.

3. OBSERVATIONS

In this section, the first author, who organized the panel and formulated the questions, makes observations on the panelists' responses.

On multimedia. The definitions given by the panelists seem to be in general agreement. Although there is a strong focus on video analysis, the term multimedia is applied to scanned documents. Furthermore, it is pointed out that the term can apply to the structure of a document, a device, or software.

How MIR is different. Several aspects are mentioned including the query, integration of multiple features in time, user's intuition, the lack of "alphabet," and the perceptual gap.

Examples of MIR. Examples include image and text, as well as videos with different sounds. Challenges mentioned include the client device, and the fact that people are not exposed to MIR so they cannot properly use it.

Most important problems. The semantic gap continues to be one of the major issues, along with scalability, benchmarking, search speed, generalization in Machine Learning, modeling, selection and fusion of modalities, meta-content, use of humans' intelligence and user interfaces.

Killer application. Web-based video search is mentioned by two of the panelists. But there also seems to be agreement that important applications will emerge in other areas such as community search, personal multimedia, and specialized domains.

Scientific discipline. The panelists agree that more work needs to be done in developing a unifying theoretical framework, on the importance of benchmarks like TRECVID, and on studying issues related to fusion of multiple media sources.

Role of the user. Relevance feedback, incorporating user intelligence in the search process, human memory, and models of human activity are considered important issues. It is also noted that different approaches should be used depending on the particular application.

Technology likely to impact MIR. Several technologies are mentioned, but particular emphasis is placed on broadband and any technologies that allow instant and seamless sharing of multimedia content.

Standards and DRM. On one hand, it is mentioned that secure content distribution and repurposing, as well as user privacy protection techniques should boost MIR applications. On the other hand, it is said that MIR does not necessarily need more standards or better DRM to make a success in real world applications, and if content producers don't leverage standards, the value of building compatible systems is low.

Usage of MIR. The panelists mention the ease of use of text, and the difficulties in formulating MIR queries as possible reasons for low use of MIR. One panelist mentions the value of multimedia as an important factor, and the aggregation of media is also mentioned. Integration of media sharing, annotation, search, and management across desktop and the Web is seen as important in boosting MIR.

The panel covers a wide range of topics and serves as a mere starting point of discussion. One observation is that the focus continues to be video and the web. Although examples of MIR are given, video continues to dominate. The term multimedia, however, includes integration of many other types of media coming from motion sensors, haptic sensors, and others. Another important issue, mentioned by the panelists, is the user's difficulty in formulating a query, and although relevance feedback is mentioned as an important technique, it does not necessarily address the problem: each iteration is in fact a query. Memory is mentioned as important in considering the role of the user. But other factors such as cultural differences, context, subjectivity, and emotion are not mentioned. Sharing of multimedia content is discussed as an important area, with brief mentions of annotation. Recent trends in the use of the web suggest that community annotation is likely to play an important role in multimedia data aggregation. But there are many open issues related to annotation of multimedia, including annotator consistency, accuracy, reliability, and others. This is strongly related to the use of ontologies and use of machine learning for automatic annotation.

In a way, the discussion seems to imply that the use of text for retrieval hinders the retrieval of multimedia using other media. In other words, if text is available, it still remains the easiest and most effective means of MIR. One question is, then, whether the goal should be to extract features and develop techniques to map multimedia data to textual information so we can leverage existing text retrieval techniques—one could argue that this is how most researchers are addressing the semantic gap. A more promising direction, however, (in the opinion of the first author) is to take a human-centered approach to MIR and work on more effective techniques for browsing and multi-modal retrieval (not only text). This requires a new approach to MIR for which there is a large number of open issues and many technical questions. Discussions such as this one, however, contribute to our understanding of the field and provide a pause to re-examine the approaches we are taking to solve the important problems in MIR.

4. CONCLUSIONS

The panelists have presented diverse points of view on what multimedia retrieval is, what the main research problems are, and what the future killer applications might be. As the answers indicate, the outlook for MIR remains positive, but many challenges remain, particularly in user-related issues: query formulation, usage of MIR, and the semantic gap.

5. ACKNOWLEDGMENTS

Mike Christel: my experience derives from ongoing Informedia research efforts, (http://www.informedia.cs.cmu.edu), work supported by the National Science Foundation under Grant No. IIS-0205219 and by the Advanced Research and Development Activity under contract numbers H98230-04-C-0406 and NBCHC040037.

The panel organizer wishes to thank the panelists for participating, and their institutions for their support. We would also like to thank the workshop co-chairs.

6. REFERENCES

[1] Chang, S.-F., moderator. Multimedia Access and Retrieval: The State of the Art and Future Directions. In *Proc. ACM Multimedia '99* (Orlando FL, Nov. 1999), 443-445.

[2] Christel, M., Smith, M., Taylor, C.R., and Winkler, D. Evolving Video Skims into Useful Multimedia Abstractions. In *Proc. CHI '98* (Los Angeles, CA, April 1998), 171-178.

[3] Davis, M., King, S., Good, N., and Sarvas, R. From Context to Content: Leveraging Context to Infer Media Metadata. In *Proc. ACM Multimedia '04* (New York, NY, Oct. 2004), 188-195.

[4] Hauptmann, A. G. Lessons for the Future from a Decade of Informedia Video Analysis Research. In Proc. CIVR '05 (Singapore, July 2005), LNCS 3568: 1-10.

[5] Hauptmann, A., and Christel, M. Successful Approaches in the TREC Video Retrieval Evaluations. In Proc. ACM Multimedia '04 (New York, NY, October 2004), 668-675.

[6] Hart, P.E., Piersol, K., and Hull, J.J. Refocusing Multimedia Research on Short Clips. IEEE Magazine 12(3): 8-13.

[7] Rowe, L.A. and Jain, R., ACM SIGMM Retreat Report on Future Directions in Multimedia Research, http://www.sigmm.org/Events/reports/retreat03/sigmm-retreat03-final.pdf, March, 2004.

[8] http://www.cs.cmu.edu/~christel/MyPubs.html