ACM SIGMM Retreat Report on Future Directions in Multimedia Research

(Final Report March 4, 2004)

Lawrence A. Rowe
Computer Science Division – EECS
University of California
Berkeley, CA 94720-1776

Ramesh Jain
School of Elec. & Comp. Eng.
Georgia Institute of Technology
Atlanta, GA 30332-0250

Abstract

The ACM Multimedia Special Interest Group was created ten years ago. Since that time, researchers have solved a number of important problems related to media processing, multimedia databases, and distributed multimedia applications. A strategic retreat was organized as part of ACM Multimedia 2003 to assess the current state of multimedia research and suggest directions for future research. This report presents the recommendations developed during the retreat. The major observation is that research in the past decade has significantly advanced hardware and software support for distributed multimedia applications and that future research should focus on identifying and delivering applications that impact users in the real-world.

The retreat suggested the community focus on solving three grand challenges: 1) make authoring complex multimedia titles as easy as using a word processor or drawing program, 2) make interactions with remote people and environments nearly the same as interactions with local people and environments, and 3) make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment. The focus of multimedia researchers should be on applications that incorporate correlated media, fuse data from different sources, and use context to improve application performance.

1. Introduction

The word *multimedia* has many definitions, but most people incorporate the idea of combining different media into one application such as an educational title that uses text, still images, and sound to explain or demonstrate a concept. Another example is a distributed collaboration that allows people at different locations to work together on a document or jointly operate a remote instrument (e.g., a telescope) using streaming audio, video, and application-specific data.

The first computer-based use of multimedia occurred in the early 1960's when text and images were combined in a document. Soon thereafter applications incorporated continuous media. Audio, video, and animations are examples of continuous media. These media are combined, in other words synchronized, using a time-line to specify when they should be played.

The multimedia research community is inherently multi-disciplinary in that it includes people with interests in operating systems, networking, signal processing, graphics, databases, human-computer interfaces, cognitive science, and various application communities (e.g., education, architecture, art, etc.). Consequently, several ACM Special Interest Groups joined together to organize the First International Conference on Multimedia in 1993 (MM'93). The conference was co-located with the SIGGRAPH Conference being held that year in Anaheim, California. MM'93 was very successful and lead to the formation of SIG Multimedia (SIGMM) in early 1994 to serve the multimedia research community.

The annual Multimedia Conference was separated from SIGGRAPH in 1994 to encourage more interaction amongst participants. This annual conference is the premier conference for the publication of multimedia research as measured by attendance and the selectivity of the papers published. Most years more than 300 papers are submitted to the program committee of which 15-18% are accepted for publication. The conference has grown over the years to include formal papers, posters, demonstrations, videos, and a dissertation review program.

Discussions at MM'01 in Ottawa, Canada suggested it was time for senior members of the research community to meet and discuss the current state and future direction for multimedia research. Some members of the community believed that multimedia research, as represented by publications at the annual conference, was addressing narrow topics with limited impact rather than addressing major problems that will have wider impact on technology for real and emerging applications. At the same time, people inside and outside the community questioned why the Multimedia Conference has not grown into a major event similar to the SIGGRAPH conference. The belief is that multimedia is such a hot topic that the conference should attract several thousand people rather than the 200-300 people that typically attend.

Professors Lawrence A. Rowe and Ramesh Jain, the past and current SIGMM Chairs, respectively, organized the retreat with advice from the SIGMM Executive Committee. A two-day retreat was held in conjunction with MM'03 in Berkeley, California. The Executive Committee selected the retreat attendees. Twenty-six researchers, shown in Table 1, participated in the retreat. The goal was to include both academic and industrial researchers from a variety of areas as well as young and old members of the community. Each participant was invited to write a short position paper briefly responding to questions about past research successes, future research directions, and the current state of SIGMM and the annual conference. These position papers were distributed to attendees before the retreat and are being published jointly with this report [SIGMM 2003].

The first day of the retreat was dedicated to discussions about future directions for multimedia research, and the second day focused on organizational issues. This report covers the research recommendations developed during the retreat. These recommendations have been modified somewhat after a public presentation and discussion at MM'03. The organizational issues report will be published separately on the SIGMM Website (http://www.acm.org/sigmm).

The remainder of this report is organized as follows. Section 2 presents background on multimedia research over the past decade. Section 3 presents unifying themes, which underlie the field. Section 4 presents three *Grand Challenges* identified as the problems that multimedia researchers should be trying to solve and funding agencies should be supporting. Finally, section 5 discusses topics mentioned at the retreat and in public and private discussions since the initial findings were presented.

Table 1: SIGMM Retreat Participants

Sid Ahuja (Lucent)
Brian Bailey (UIUC)
Dick Bulterman (CWI)
Shih-Fu Chang (Columbia)
Tat-Seng Chua (Singapore)
Marc Davis (UC Berkeley)
Nevenka Dimitrova (Philips Research)
Wolfgang Effelsberg (TU Mannheim)
Jim Gemmell (Microsoft Research)
Forouzan Golshani (Arizona State U)
Nicolas Georganas (U Ottawa) *
Ramesh Jain (GaTech) *
Martin Kienzle (IBM Research)
Wolfgang Klas (U Vienna)
Joseph Konstan (U Minn) *
Dwight Makaroff (U Saskatchwan) +
Ketan Mayer-Patel (U North Carolina)
Klara Narhstedt (UIUC) *
Arturo Pizano (Siemens SCR)
Thomas Plagemann (U Oslo)
Lawrence A. Rowe (UCB) *
Henning Schulzrinne (Columbia)
Ralf Steinmetz (TU Darmstadt) *
Michael Vernick (Avaya)
Harrick Vin (U Texas)
Lynn Wilcox (FX PAL)

* Member SIGMM Executive Committee
+ SIGMM Information Director

2. Multimedia Research Background

Multimedia research through the middle 1990's focused on the development of infrastructure to support the capture, storage, transmission, and presentation of multimedia data. Researchers and product developers worked on I/O devices, scheduling algorithms, media representations, compression algorithms, media file servers, streaming and real-time network protocols, multimedia databases, and tools for authoring multimedia titles. Driving applications included CD-ROM playback, non-linear audio/video editing, videoconferencing, multimedia content analysis and search, lecture webcasting, video-on-demand (VOD), and video games. While many companies focused on stand-alone multimedia applications (e.g., CD-ROM playback), the research community recognized early on that the most important and difficult applications involved distributed multimedia, sometimes called "networked multimedia," and multimedia database applications. Examples are VOD, videoconferencing, and algorithms to analyze and search music and image databases.

Research on compression algorithms, which began in the 1950's, has lead to the development of standards for low bandwidth audio and video coding that support video conferencing applications and wireless telephony. Low-latency coding is important for these applications because human communication requires bounded end-to-end delay. Compression standards were developed in the 1980's and early 1990's for low-bandwidth, high-quality audio coding to support telephony and for high-quality video coding to support home entertainment applications (e.g., satellite receivers and personal video recorders) and transmission of broadcast programming (e.g., delivering live news and sporting events from anywhere in the world). This research on coding has yielded many algorithms that operate at different points in the space, time, bandwidth, and computational complexity space. While research will continue on further improvements in coding algorithms, many researchers believe dramatic improvements in coding will require significant breakthroughs.

Computer network research has been an active area of multimedia research since the 1980's. New protocols were developed with standard wire formats for packet audio and video that enabled continuous media players to recover when packets are lost. Significant changes to the standard Internet model were explored to support bounded delay protocols and resource management to insure that time-critical streaming media packets are delivered before less time-critical packets (e.g., FTP). Multicast protocols were designed, implemented, and deployed to support broadcast and small group collaboration applications. Today researchers are developing new protocols for wireless networks. Considerable progress has been made on systems and protocols for media streaming, but resource management, scalable multicast protocols, and wireless networking continue to be a challenge.

The conversion to digital media, whether still images taken by a digital camera, an mp3 song downloaded from a music archive, or an mpeg video captured by a desktop video camera or cellphone, and the development of large media databases, which were enabled by the dramatic increase in storage capacity over the past two decades, has lead to research on algorithms to automate the analysis, indexing, summarization, and searching of this content. Internet search engines that operate on text data have proven extremely valuable. Next generation search engines will incorporate other media. While some limited successes have been achieved in multimedia analysis and search, digital asset management that solves real-world problems continues to be a challenge.

Many researchers and companies have developed tools for authoring multimedia content. Content examples are video games, web-based hypermedia (i.e., media data with links between components), and CD-/DVD-ROM titles. Non-linear audio and video editors are notable successes. However, creating multimedia content and using it in everyday applications (e.g., email, documents, web titles, presentations, etc.) is still not possible for most users. For example, many colleges and universities regularly webcast lectures of various sorts (e.g., classes, seminars, conferences, etc.). Using this material in an assignment or creating a study guide that includes links to selected clips with annotations is difficult. Better tools are also needed for professional content authors. Specifically, current tools poorly serve artistic content and multi-player game authors.

While early multimedia systems required special-purpose hardware to decode and play continuous media, regardless of whether it was streamed across a network or read from a local storage system, the continuing improvement in semiconductor technology, the introduction of special-purpose instruction sets (e.g., Intel MMX), and the addition of special-purpose processors on graphics adapters has made multimedia playback and media processing a software application available on all modern PC's. Software media processing coupled with the deployment of broadband networking suggests that distributed multimedia applications will become increasingly important over the next decade.

In summary, research over the past several decades has focused on the "nuts & bolts" infrastructure required by multimedia applications. These applications are inherently *real-time*, which means events or processes must respond within a bounded time to an event (e.g., an I/O interrupt), and *isochronous*, which means processing must occur at regular time intervals (e.g., decode and display a video frame 24 times per second). Two fundamental principals were developed: *statistical guarantees* and *adaptation*. Because continuous media has a presentation time-line, a video frame or audio block that is not available at the scheduled playout time is worthless. Early researchers studied resource allocation algorithms that could guarantee on-time availability. However, guaranteed service requires the reservation of too many resources to prepare for an infrequent worst-case scenario. The development of statistical guarantees allows improved utilization of resources while providing the user a high-quality experience. This high-quality experience is possible because applications can adapt to lost data and limited resources. For example, a decoder can fill-in data that is lost by using redundant information sent in previous packets or by constructing plausible values for missing data. The term *quality-of-service* (QoS) refers to the allocation of resources to provide a specified level of service. QoS management and the development of algorithms and technologies to produce the highest user-perceived quality experience is an important contribution of multimedia research.

3. Unifying Themes

The multimedia field, as mentioned above, is inherently multi-disciplinary. Few researchers identify multimedia as their primary research area. More often, researchers identify themselves as being in signal processing, computer systems, databases, user interfaces, graphics, vision, or computer networking. This list of areas ignores the content side of multimedia, whether it be artistic, entertainment, or educational, which must also be considered part of the multimedia research community. One goal of the retreat was to identify the unifying or overarching themes that unite the multimedia field. These themes help to inform us about the nature of multimedia research.

Many important unifying themes were identified during discussions at the retreat. These themes can be organized into three areas. First, a multimedia system or application is composed of more than one media that are correlated. The media can be discrete (e.g., an image or text document) or time-based (e.g., weather samples collected by a sensor network or a video).[1] Different media are correlated but not necessarily time-based or co-located. For example, an artist might put together a still image and a video to evoke a particular response in the viewer. Or, musicians at different locations playing together are using multiple streams of time-based media (audio) created at different geographic locations. Someone listening to the performance hears one sound, most likely from a stereo or multiple channel surround sound system. Notice that this example is multiple streams of the same media type. A virtual clock correlates the different streams. The representation of time and synchronization of time-based events is a fundamental concept for multimedia applications.

The second unifying theme is integration and adaptation. Any distributed application with user interactions must deal with end-to-end performance and user perception. Multimedia applications are cross-layer (e.g., network protocols, software abstractions, etc.) and multi-level (e.g., high-level abstractions down to low-level representations). For example, streaming media requires application-level framing, that is, the application is best at deciding how to pack media data into network packets because only the application knows enough about the data to make framing decisions that will enable recovery when packets are lost. A simple example of multi-level media is the use of different sized images in an application (e.g., thumbnail images in a summary view and large images in a detailed view). Similar ideas have been applied to other media (e.g., video summarization, audio skimming, etc.) and to different applications (e.g., hierarchical coding). Distributed multimedia applications should provide transparent delivery of dynamic content. Content must adapt to the user's

[1] While analog data can be continuous, digital data is always discrete. Digital data can be sampled frequently to capture an acceptable representation of the continuous data.

environment. For example, content displayed on a PDA might look and behave differently than content displayed on a large projection screen in a classroom or theatre.

Media integration means that information is conveyed by the relationship between media as well as by the media itself. A simple example is a video composed of a sequence of audio blocks synchronized with a sequence of still images. The user requires both sequences to understand the video. Either media by itself is insufficient. Much of the current research in analysis, compression, and organization considers these sequences separately. Media must be considered separately and jointly to address emerging problems.

A second facet of integration and adaptation is ubiquitous interaction with multiple media. One retreat participant cited the following example. A user should be able to enter a room and interact with various devices and sensors in that space. For example, the user's laptop computer or PDA should sense or query the environment to locate cameras, microphones, printers, presentation projectors and the applications available to manage and use them. It should be easy to access, display, annotate, and modify the media. Contrast this situation with the reality today. A user must explicitly configure his or her computer to tell it what devices and sensors exist and how to use them. Unfortunately, retrieving data from a large collection for display to remote participants or capturing a reference to a subpart of this media along with an annotation currently requires detailed knowledge about media representations, networking, and other components of the system infrastructure. The focus should be on "ease of use" to solve a problem, not on system configuration and operation.

A third facet of integration and adaptation is the emphasis on using multiple media and context to improve application performance. Early research on multimedia content analysis, summarization, and search focused on one media type (e.g., still image or music archive query) and limited context. Researchers are now exploring systems that use information derived from correlated media and context. For example, executing a query to find information about the election of a state governor might involve restricting the search to TV news programs and identifying segments in which a person is shown in the video stream who uses the words "election" and "governor" in the audio stream. Using the type of program (e.g., a chemistry lecture, baseball game, etc.) is an example of using context to guide the search and improve the results. Research on parsing news programs and producing synchronized text transcripts has been very successful. The challenge now is to extend this research to less well-structured environments where transcripts are not provided and the speaker uses different vocabularies. Lecture webcasts or discussion seminars are examples of less well-structured content in an education setting.

The third unifying theme is that multimedia applications are multi-modal and interactive. The conventional interface to a desktop or laptop computer, that is, a two-dimensional windows, mouse, and keyboard interface, is being replaced with new interface modalities (e.g., pen, voice, gesture, touch, smell, etc.) and multiple devices (e.g., PDA's, active badges, tablet computers, projectors with embedded computers connected directly to the network, etc.) and smart spaces. Most applications that use these interfaces will be interactive and require coordination between different ways of specifying an operation (e.g., pushing a button on a computer, gesturing with your hand or speaking a command might specify "move to the next slide" in a presentation). These applications are also inherently multimedia because they incorporate traditional continuous media (e.g., audio and video). Human-computer interactions as well as communication among humans using computer-based applications (e.g., Voice-over-IP, video conferencing, immersive environments, etc.) are important themes for research on multimedia applications. Several people noted that human-computer interactions in the future would be more like human-to-human communication.

4. Grand Challenges

The primary goal of the retreat was to identify a short list of "grand challenges" that multimedia researchers should solve. The idea was to raise awareness of the importance of research that involves "multiple media" and impacts real users as opposed to narrow results that can be published in a journal.

Three grand challenges were identified. The first challenge is to *make authoring complex multimedia titles as easy as using a word processor or drawing program.* Content authoring is expensive and difficult. Most groups that produce hypermedia content use teams of experts supervised by producers and directors. Specialized tools are used for different media (e.g., a word processor for text, a non-linear editor for audio and video, an image editing tool for still images, a 3D modeling system for a animations, etc.). These content elements are then combined to produce the title. Combining the different media with rules for synchronization and user interaction requires a programmer and yet another collection of tools. Producing the title, that is, coding the material and physically publishing it (e.g., pressing a DVD or uploading the material to one or more servers) is time-consuming and complex. Moreover, different versions of the title are typically produced for different environments (e.g., TV set-top box, game console, desktop computer, PDA, etc.), which is itself a challenge. Few people have the experience required to use these tools and produce multiple versions of a title.

The multimedia research community should develop the algorithms, heuristics, and tools that will allow average users to produce compelling multimedia content. Users need tools to support creation of different types of content. Some examples are:

1. A teacher needs tools to prepare educational material that includes video demonstrations to show an object and simulations and animations to illustrate dynamic behaviors. Good educational material allows students to explore the underlying principles and objects by modifying the input parameters to a simulation and examining related objects.
2. A travel agent needs tools to prepare material showing places and experiences potential customers might want to visit. This material might include pictures, videos, sounds, and other immersive experiences. It might include live interactions with people at a remote location. It might also include links to material authored by someone who has taken the trip. This title might be composed of slide shows, videos, trip summaries, and links to detailed information about places visited, artifacts seen, places to stay, and methods of transportation.
3. A family member needs tools to prepare material that documents a significant life event such as a birthday, wedding, or birth of a child. For example, the title created for a wedding might include a time-line that relates different events (e.g., proposal, engagement, wedding ceremony, and celebration), formal and informal pictures of the participants and the event, audio and video captured at various stages in the event, and detailed information about the clothes worn by the bride and groom and the locations of the ceremony and reception.

Some excellent tools exist either for particular media (e.g., Photoshop for images, Dreamweaver for websites, Premiere for audio/video, etc.) and particular applications (e.g., PowerPoint for presentations, iMovie for home movies, FrameMaker and Word for documents, etc.). The problem is that the tools are not integrated, do not encourage content re-use, run on different platforms, and are targeted at different user communities. For example, Photoshop is an excellent tool for graphic design experts, and iMovie is an excellent tool for less sophisticated tech-oriented end-users. The problem is that expert-user tools require too much learning and end-user tools are typically too restrictive. Authoring tools and systems are needed that can incorporate editors for different media, perhaps different versions of the editor depending on user experience or application requirements. These tools must work together seamlessly with content acquired from different sources (e.g., uploaded from a capture device, downloaded from an archive, created during the authoring process, derived from other content, etc.). And lastly, the tools must incorporate features to support production of different versions of the title and on-going enhancement and bug fixing.

Tools can make a significant difference. PowerPoint is a good example. Before the development of this end-user tool, expensive custom-designed commercial systems were used to produce 35mm slides. These systems provided expert-user interfaces that ran on specialized computer systems. The introduction of PowerPoint and similar tools in the 1980's made the creation of formal presentation material available to nearly every PC user. The grand challenge is the development of the systems and tools required to support widespread authoring of multimedia content.

This grand challenge may sound like an engineering problem, that is, build a wonderful software package. Our intent is not an engineering solution, rather the challenge to the research community is to develop new user-interface paradigms, software abstractions, media processing algorithms, display presentations and operations for editing media, and media databases that will significantly reduce the effort required to produce high-quality multimedia titles. In other words, make all data types first-class citizens in our applications. To begin with, it might make sense to carefully observe and measure how titles are created today. Research on multimedia authoring will likely include better understanding of media aesthetics, storytelling, and how people relate to a multimedia experience.

The second grand challenge is to *make interactions with remote people and environments nearly the same as interactions with local people and environments*. This grand challenge incorporates two problems: distributed collaboration and interactive, immersive three-dimensional environments. Videophones and videoconferencing have been around for a long time. RCA conducted early experiments with video telephone booths in the 1930's, and A.T.&T. demonstrated a videophone at the 1960 World's Fair in New York. While notable successes have been achieved, for example, small group videoconferencing using H.323 systems (e.g., Polycom), web-based on-line meeting services (e.g., WebEx), person-to-person video chats (e.g., NetMeeting), and webcasting audio or video programming, the telephone is still the dominant medium for remote collaboration. It is still too complicated and expensive to incorporate n-way collaboration using a variety of media into our day-to-day lives. Many problems can be identified including: 1) the difficulty of setting up and operating the equipment, 2) the cost of bandwidth required for high-quality n-way communication is too expensive, 3) the poor support for flexible and scalable multicast services, 4) service limitations (e.g., H.323 does not allow multiple people to view each other and carry on parallel conversations at the same time nor does it scale to large groups with different meeting interaction styles such as lectures, town meetings, etc.), and 5) collaboration tools, that is tools to encourage working together on a document, viewing results produced by remote equipment (e.g., a telescope or CAT scanner) or meeting to review project schedules, are inadequate. But, the grand challenge is more than solving the distributed collaboration problem. The promise of interactions with remote people, places, and virtual environments, may dramatically change the way we live. New sensors (e.g., touch, smell, taste, motion, etc.) and output devices (e.g., large immersive displays and personal displays integrated with eye glasses) offer the opportunity for more intimate and sensitive interaction with a remote environment. And, continued development of semiconductor technology will bring real-time three-dimensional virtual environments to every computing and communication platform. As one participant said, "interacting with a remote environment should be better than being there." The grand challenge is to understand the opportunities these new hardware technologies offer and to develop user interfaces and interaction paradigms that allow seamless communication and interactions with remote and virtual environments.

Many research problems are incorporated into this grand challenge including exploring the use of multiple streams of data, whether it be images, sounds, or sensor readings, and developing interaction hardware and software that allow humans to use this data. For example, users interacting with educational or entertainment programs, whether it be live sporting events, lecture webcasts, or broadcast programs want a variety of services that allow them to locate interesting or important events, view program summaries (e.g., a lecture outline or baseball game summary) with links to allow them to watch detailed segments of the program, skim through stored programs rapidly, record material for viewing at a different time (i.e., time shifting), and translations that allow viewing a program on a different platform (e.g., a TV or cellphone). Moreover, it should be easy to create derived works from the content. These services and many more will be required in the future for interactions with remote people and environments.

The third grand challenge is to *make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment.* The widespread adoption of digital cameras and the emergence of cellphones with built-in video cameras are adding to the information glut. Of course, increases in storage capacity and reductions in cost make it possible to store massive amounts of this data. The challenge is to make it useful. Past research has addressed multimedia database models, algorithms to analyze media data, and algorithms to search for relevant or interesting data (e.g., query a music archive by humming a tune or find pictures with similar color palettes). However, significant challenges remain before the multimedia storage and search problems are solved. For example, the following problems cannot be solved today.

1. Search an archive of radio broadcasts to find an interview with a particular individual and a picture archive to find a photo of the person visiting a particular city. Text-to-speech requires context to disambiguate the words being spoken (e.g., technical terms interspersed in a news broadcast are often misunderstood) and identifying where a particular photo was taken might require extensive image analysis or automatic capture of metadata when the photo was taken (e.g., geographic location of the camera at the time the picture was captured). The problem is complicated by the fact that the data in the broadcast archive is not fused with the photo archive.

2. Find lectures by a particular person published on the web. This problem might be solved by looking at the text associated with a streaming media file published on a web page. However, it may be difficult to identify the text associated with a video clip if the web page is generated dynamically. Problems arise too because most commercial webcasting systems use proprietary media coding, storage representations, and network packet formats.
3. Who is that person across the room? The idea is to point your cellphone camera at the person and have it tell you the name of the person. Solving this problem takes context and data fusion as well as connecting to a shared database and a processing server. The obvious solution is to do face matching on the person using the captured image. But, this approach might return too many possible matches or take too much time. What the system should do is use the context of the situation (e.g., a holiday party for a company or a workshop at a conference) to restrict the candidate matches to people who might actually be at the event.
4. Make the billions of hours of home video currently stored in shoeboxes useful. People shoot video but there are no good tools to organize and store it in a form so a user can say, "show me the shot in which Jay ordered Lexi to get the ball." The solution to this problem may require developing semi-automatic analysis tools coupled with powerful tagging and indexing to organize data so it is easily accessible using unified indexes.

Again, these problems might look like engineering problems because they are drawn from applications that people might actually want to use. However, the grand challenge is to work on the fundamental algorithms (e.g., query planning, parallel search, media-specific search and restriction, combining partial results, unified indexing, and tagging multimedia data) so the problem can be solved sufficiently well that a system could be built and deployed that people will use.

Finally, underlying this last grand challenge is the problem of digital rights management. While this topic is not directly related to multimedia, it will have a dramatic impact on the development and use of content. Discussions at the retreat identified the need for access and propagation rights particularly for fair-use and educational-use rights, the need to track the source of a media asset, and the need for an economic model to pay content owners and creators.

5. Discussion

During the retreat several topics were discussed that merit comment. First, people asked whether research on text and images (e.g., a web browser) or research on analyzing or querying a single media (e.g., an image archive) is acceptable multimedia research today. The answer is "no" to text and image research unless it contributed something new to our understanding of combining these media. Ten years ago such research was definitely valid. The sense at the retreat is that we should raise the bar. Music query is a more difficult question. On the one hand, it is a single media, which suggests "no." Truly innovative research on a single media will always be of interest. But, the contribution must be greater than "here is my nifty algorithm to query still images using color histograms and frequency domain filtering." Repeatable experiments using published benchmarks are required for the field to progress. Several times during the retreat two ideas were emphasized: 1) compare new algorithms with previously published algorithms using published benchmarks, and 2) use software developed by other researchers. For example, while some people raised issues about specific details, the development of the video TREC benchmark is viewed as an important contribution to the research community [NIST 2003].

Using software developed by others has been a continuous refrain in the multimedia research community. Everyone recognizes the need to avoid re-implementing known algorithms. The problem is that funding to support a widely adopted common toolkit is too expensive for traditional funding sources. One participant remarked on the investment in Unix by A.T.&T. or Java by Sun. Building an acceptable software toolkit for distributed multimedia applications will likely cost $20M or more with on going support costs. It is difficult to fund these types of projects without a strong financial incentive for the funding organization. Meanwhile, many researchers build incomplete toolkits and targeted platforms and they spend too much time trying to use a limited commercial toolkit to test one idea. The community collectively spends many times the money required to build an open source portable toolkit re-implementing common components. The lack of software toolkits also causes researchers to forego possible experiments because the effort required to do them is too high.

Several interesting comments were made during the authoring tools discussion. The trade-off between tools for expert versus novice users is well known in the human-computer interface community. Experts want tools that provide more features and finer-grain control over the presentation and behavior of the content being created. However, more features and finer-grain control typically leads to more complex user-interfaces (e.g., they have more options and operations), which makes the authoring tool more difficult for less experienced people to learn and use. On the other hand, a tool designed for inexperienced people might be missing the features required for an expert. The Holy Grail is a system that adapts to user experience, but that remains a challenge for the research community to achieve. The notions of incorporating critics or agents to watch user behavior and either automatically change the object or suggest a change to the user to improve it is an appealing idea. But, more research is needed on these types of interactive environments before they can be made to work.

A common metric used in multimedia applications is QoS. This metric optimizes an internal parameter related to the application (e.g., network delay, CPU cycles dedicated to decompression, etc.). Discussions during the retreat supported the observation that *Quality of Experience* (QoE) is more important than QoS because it relates the user-perceived experience directly rather than the implied impact of QoS. QoE is related to QoS, but it might be a complex function of several parameters, including human perception, rather than easily quantified engineering parameters. The multimedia research community should focus on QoE as the primary metric to be optimized. Research that incorporates the user is more difficult because human behavior is so variable. Nevertheless, the goal of nearly all multimedia applications is to solve a problem for a user, so user perception must be incorporated in an evaluation metric for the algorithm or application.

The last three topics that received considerable attention and generated interesting discussions during the retreat relate to new media, user interfaces, and system configurations. Looking at new media (e.g., haptic, smell, and other sensors) will encourage researchers to think about new ways that humans can interact remotely with people or equipment or with the computer itself. Multimedia research during the past decade has focused on audio and video media. It is time to explore other media.

The windows, mouse, and keyboard human computer interface using a two-dimensional output display has been the standard UI for the past twenty-five years. Many new interface devices and metaphors have been explored during this time including three-dimensional output displays and pen, speech, gesture, and tangible input interfaces. An important theme is multi-modal interfaces that allow the user to interact with the system using several media (e.g., pen and speech). The user should use different devices for an operation (e.g., gesture or mouse) depending on the situation. The multimedia research community should participate in these research activities, even though they are primarily UI problems, because the multimedia community understands the underlying time-oriented media and multimedia processing.

Lastly, many participants discussed the ubiquitous computing metaphor for human-computer interaction. The idea is that many sensors and smart devices with embedded computers will be present in our environment either carried by the user or permanently located in the space. Applications should be written to exploit this collection of devices. They should adapt to the availability of equipment and processing to solve a user's problem. Distributed multimedia is inherent in this new world.

The past decade has seen significant progress in multimedia research. Now is the time to raise expectations for the future. The focus should be on incorporating new media and devices and exploiting multiple media to create applications that solve an important problem and produce high quality user experiences.

6. References

[NIST 2003] Digital Video Retrieval at NIST, http://www-nlpir.nist.gov/projects/trecvid, 2003.

[SIGMM 2003] ACM SIG Multimedia Strategic Retreat Participant Position Papers, http://www.acm.org/sigmm/, December 2003.