Looking Forward 10 Years to Multimedia Successes

LAWRENCE A. ROWE, FX Palo Alto Laboratory

A panel at ACM Multimedia 2012 addressed research successes in the past 20 years. While the panel focused on the past, this article discusses successes since the ACM SIGMM 2003 Retreat and suggests research directions in the next ten years. While significant progress has been made, more research is required to allow multimedia to impact our everyday computing environment. The importance of hardware changes on future research directions is discussed. We believe ubiquitous computing—meaning abundant computation and network bandwidth—should be applied in novel ways to solve multimedia grand challenges and continue the IT revolution of the past century.

Categories and Subject Descriptors: A.m [**General Literature**]: Miscellaneous; C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Distributed applications*; H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms: Management, Algorithms, Human Factors, Performance

Additional Key Words and Phrases: Multimedia research directions

ACM Reference Format:

Rowe, L. A. 2013. Looking forward 10 years to multimedia successes. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1s, Article 37 (October 2013), 7 pages.
DOI: http://dx.doi.org/10.1145/2490825

1. INTRODUCTION

The 20th Anniversary Panel at ACM Multimedia 2012 "Coulda, Woulda, Shoulda: 20 Years of Multimedia Opportunities" allowed senior researchers in the field to discuss changes, successes, and missed opportunities in the past 20 years. Rather than recapitulate the discussions during the session, this article reviews progress since a research retreat held ten years ago at ACM Multimedia 2003 and discusses my view of what multimedia research should be working on in the next ten years.

Past experience suggests that predicting the future is nearly impossible. While the Internet was incubated by the Advanced Research Projects Agency of the U.S. Department of Defense beginning in the late 1960's, the Commercial Internet did not appear until 1995. Who would have predicted at the 1st ACM Multimedia Conference in 1993 that by 2013 Apple Computer, among the highest valued companies in the world, would produce digital media products that defined the digital generation, that Google, founded in 1998, would dominate the 2012 U.S. online advertising market with revenue over $21 billion and will soon pass the total amount spent on TV advertising each year, and that

Author's address: L. A. Rowe, FXPAL, 3174 Porter Drive, Palo Alto, CA 94304; email: rowe@fxpal.com.

DOI: http://dx.doi.org/10.1145/2490825

Facebook, founded in 2004, would have over 1.1 billion active users out of a total world-wide population of 7.1 billion in 2013?

On a personal level, I gave a keynote presentation at the SPIE/ACM Multimedia Computing and Networking Conference in January 2005 titled "Whither Ubiquitous Video" that asked why online streaming video was not widely available given that streaming audio was widely available and the technology required to delivery video was well understood? As it turns out, the week before I gave the keynote, YouTube was founded which revolutionized the Internet video business. Today, 2 billion YouTube videos are played each day on the Internet.

Notwithstanding the challenge of predicting the future, researchers must guess which technologies and products are likely to emerge so they can choose research projects on which to work. The remainder of this article will assess the multimedia field today as compared with recommendations made ten years ago, present ideas about how to look for significant impacts on the future, and discuss my opinion about the direction for the field of multimedia for the next ten years.

2. ACM SIGMM 2003 RETREAT

A meeting was organized at ACM Multimedia 2003 at which 26 researchers from academia and industrial research laboratories participated in a two-day retreat to discuss the future of multimedia research. Each participant was invited to submit a short position paper that formed the foundation for the discussions. After developing a definition for the term multimedia, the group identified unifying themes that underlie the field and proposed three "grand challenges" that researchers should try to solve. These themes and the current state of research on the grand challenges are discussed in this section. More information on the retreat is available in a published report [Rowe and Jain 2005].

Most people understand *multimedia* to mean more than one media combined in an application, title, or experience. The first unifying theme was that a multimedia application incorporates multiple media that are correlated. Media can be discrete (e.g., an image) or continuous (e.g., sensor outputs), and it can be consumed in real time or offline. Most research in 2003 was concerned with distributed or networked multimedia as opposed to media consumed from local storage (i.e., CD playback).

The second unifying theme was integration and adaptation. All multimedia applications must deal with end-to-end performance and user perception. Put simply, human perception is paramount, information is conveyed by the relationship between media as well as the media itself, interaction is ubiquitous, and context is an important property of an application. Every multimedia application must deal with these issues.

The third unifying theme was that for most researchers, multimedia applications are multimodal and interactive. Several people noted that human-computer interactions in the future would be more like human-to-human communication. Of course, although it was not discussed, retreat participants understood that humans would become assistants to computer algorithms (e.g., Amazon Mechanical Turk) and computers would grow more important as personal assistants to humans (e.g., Apple Siri).

The three grand challenges identified as problems that multimedia researchers should try to solve were:

(1) to make authoring complex multimedia titles as easy as using a word processor or drawing program,

(2) to make interacting with remote people and environments nearly the same as interacting with local people and environments, and

(3) to make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment.

An assessment of where we are with respect to solving these challenges is discussed in the following paragraphs.

2.1 Multimedia Authoring

Representations and authoring tools for multimedia content have definitely improved over the past ten years. It is relatively easy to capture images, audio, and video using smart phones and incorporate them into applications (e.g., photo-books, websites, social media networks, and documents). This improvement is gratifying because the most frustrating problem I had to solve when creating the "Whither Ubiquitous Video?" keynote was trying to play different video formats using different players within PowerPoint, Web Browser, Adobe Acrobat Reader, or another presentation package. Today, it is easy to incorporate audio and video into popular presentation tools and webpages. In fact, we were asked to produce a slide to introduce ourselves at the 20th Anniversary Panel that incorporated multiple media. I was able to create a slide with text, images, animations, and video the day before the session using my smart phone and laptop. It may not seem like a big deal, but that illustrates the progress made in the past ten years.

The development of HTML5, in particular video tags, is likely to lead to further improvements in supporting continuous media. Several nonlinear audio and video editors have been developed that run either locally or in the cloud. Last summer I captured and edited a short video on my smart phone while on a trip to Singapore. It was cumbersome and lacked many features expected in YouTube-quality video, including the lack of content imagination by the author, but it was easier to do than I expected. However, it is still difficult to share a clip from a continuous media title with a friend or to create a derived work that uses multiple media, including clips from different video titles represented in different formats. Dick Bulterman, who was also on the panel, has an interesting example of how people want to share experiences by taking clips, annotating them, and publishing or sharing them (e.g., Twitter, email, Facebook, etc.).

Another area of authoring progress is the recognition of different types of video content and the development of templates for authoring and/or consuming the content. Good examples include authoring music video [Muvee 1999], lecture videos [Mukhopadhyay and Smith 1999], how-to videos [Branham and Carter 2012; Howcast 2013], and hypervideos [Shipman et al. 2003]. Also, several Augmented Reality toolkits are being actively developed for authoring interactive experiences (e.g., Qualcomm Vuforia and Artoolkit from the HIT Lab at U. of Washington).

My personal interest has been the creation and use of lecture videos for learning. I built the lecture webcasting system at Berkeley [Walsh 2011] in the early 2000's. FXPAL developed and deployed the TalkMiner search engine for lecture webcasts [Adcock et al. 2010]. More recently, the interest in Massive Open Online Courses (MOOC) has led to the development of new paradigms for teaching and new tools for authoring course material composed of video lectures, quizzes, assignments, and exams. Steve Jobs famously observed, "People do not read anymore." That might contribute to the success of MOOC's, because a series of course lectures and assignments is easier to produce and consume than a book that covers the same material. In other words, a MOOC is a "book" for the digital generation. An interesting side note in early MOOC experimentation was the authoring of a book specifically for a course [Fox and Patterson 2013]. I believe that when this new form of multimedia content for learning is further developed, it will include text material and interactive multimedia content. At that point, content will be edited and reviewed the way books are produced today. Content quality matters. There is no substitute for review and editorial assistance.

My assessment about multimedia authoring is that we have made good progress, but the challenge remains to be solved. In particular, support for linking different elements with additional material and supporting interactivity are still difficult to author.

2.2 Distributed Collaboration

The second grand challenge was to support remote collaboration that is practical and effective. As with multimedia authoring, significant progress has been made. CODECS that produce higher-quality video at lower bandwidths and high definition (HD) video standards have been developed that produce better quality at lower cost. Moreover, modern computers, including desktops, laptops, tablets, and smart phones, have the computational power to encode and decode video streams simultaneously so any device can participate as an end-point in a video collaboration. Broadband connection costs have decreased dramatically and are more widely deployed than ten years ago. Many free video chat applications have been produced (e.g., Apple FaceTime, Google Hangout, Microsoft Lync/Skype, etc.). Most of these applications support n-way collaboration (i.e., group chats with up to 10 participants), screencasting, document sharing, and bridging for telephone participants. Vendors usually charge a fee to use some of these features. Several 3D virtual environments have been developed that can be used for collaboration (e.g., 2nd Life, OpenQwaq, etc.) but these systems do not provide enough value to be widely used for business collaboration in spite of the hype and attention given to this technology over the past decade.

Given this progress, why is it still hard to use video conferencing for distributed collaboration? The primary problem is ubiquity and complexity. Everyone you want to talk with has a phone, and you can call them given a phone number. Video conferencing does not provide this level of ubiquity. A common addressing scheme does not exist, and the chance that you can setup a video conference with someone requires detailed information about conferencing technology (e.g., Skype, H.323, and FaceTime, do not interoperate) and network issues (e.g., firewalls, bandwidth constraints, etc.). Most video conferencing systems are closed and proprietary. They typically use the same call setup protocols, media and packet formats, and CODECS. But the vendors do not want to interoperate. Second, the setup and operation of a video conferencing system still requires trained technicians and operators. Simple requests like putting video on one display and presentation material on another display are nontrivial unless both people are running exactly the same hardware and software.

Other problems include the lack of collaboration-aware applications and time zone issues. Sharing presentations and documents is challenging, particularly in companies with strict Internet use restrictions (e.g., some companies do not allow employees to access a website that provides file sharing, such as DropBox). And no matter what you do, a regular conference with someone on the other side of the world (i.e., 12 hours ahead or behind) is nearly impossible to schedule. We need improved methods for asynchronous collaboration with occasional synchronous communication. Social networks (e.g., chat, blogs, twitter, etc.) provide support for asynchronous communication, but how do they interact with video discussions. In other words, how can a meeting happen over a multiple-day time period during which individuals join and leave the session multiple times.

I understand many of these issues are engineering, not science, problems. The SIP and ITU protocols have the features required for a common addressing system and interoperability that could be used to develop a common system. At present, vendors of video chat and conferencing systems do not want to support interoperability for competitive reasons. I suggest they read the history of the Kingsbury Commitment that settled the conflict between AT&T and the independent telephone companies [Brooks 1975]. Interoperability will grow your business more rapidly than attempting to establish a proprietary standard.

My assessment is that we are likely to see improved video conferencing and collaboration systems over the next decade. Customer demand will force video conferencing vendors to support interoperability. Recently, innovative research has been done in this area, particularly the work on 3D teleimmersion [Yang et al. 2010]. But it is hard to fund this research probably because commercial products

are viewed as having solved the problem, which is not true. Globalization and the benefits of exploiting talented humans located throughout the world are too important for future economic growth. Companies and funding agencies must support research on conferencing and collaboration.

2.3 Capturing and Using Digital Media

I am not an expert on multimedia retrieval. The panel had two researchers (e.g., Dick Bulterman and Ramesh Jain) who are more knowledgeable about the field, so I will defer to their judgment on this topic. The 2003 Retreat Report identified four examples that illustrated the types of problems to be solved: (1) speaker diarization (i.e., identifying the speaker in an audio clip), (2) finding a lecture by a particular person, (3) identifying a person across the room, and (4) making use of home videos. Good progress has been made on all of these problems but the general problem has yet to be solved.

Matching problems are easier to solve than search problems. For example, the Shazam music service matches a fingerprint computed from a short clip with a database of known music. The results are excellent. In other domains, the image search problem can be translated into a text search problem (e.g., the TalkMiner lecture video search engine identifies slides in the video and then OCR's the text on the slides to create a search index).

Progress has been made on search in limited domains or when restricted to a smaller number of cataloged items, but search is usually successful if the context is known and the content has either been tagged or otherwise identified by text metadata.

An important improvement over the past decade is the automatic tagging of media (e.g., date/time and GPS coordinates where a picture, audio clip, or video is captured). I believe it will be a long time before a computer-based algorithm will produce acceptable results finding a particular video clip given a text description or sketch when compared to human performance. Nevertheless, the problem is important and warrants further work.

3. SIGNIFICANT IMPACTS ON THE FUTURE

This section discusses how to identify good research problems and areas in which to work. A colleague at UC Berkeley once commented that it seems all important advances in computer science result from changes in hardware. This view may seem disturbing if you work in fields such as software development, algorithm design, or understanding human behavior and perception, because we want to believe research results are produced by our intellect rather than "processors are cheaper and faster." Frankly, I do not think it matters because we know hardware will be cheaper and faster in 5–10 years and, as researchers, we should anticipate those changes and work on problems with solutions that will be enabled by those changes.

For example, researchers at Xerox PARC believed in the 1970's that personal computers with bitmapped displays connected by a local area network would be widely available and inexpensive when they began work on the Alto, Ethernet, and GUI interface systems and applications. This research was driven by the confidence that processors and memory would be less expensive in the future. They were right and so helped define the dominant computing paradigm for the past 30 years. As processor speeds improved and memory costs declined, applications that could trade computation for user productivity made sense, such as a word processor that does continuous spell checking or a search engine that does search term completion. Lastly, low-cost graphics processors enabled game consoles, virtual and augmented reality, and high-quality interactive animations.

Consider the most recent major trends: mobile devices and cloud computing. Mobile devices were enabled by low-cost and low-power processors and radio chips, and cloud computing was enabled by low-cost servers and network bandwidth that allowed low-cost remote access to applications. Cloud

applications are sold by a subscription model that significantly reduces the total cost of ownership. Network streaming of video was enabled by low-cost secondary memory and network bandwidth.

So the question becomes, "what hardware trends in the next 10–20 years will lead to significant changes in computing?" My guess is that computation will be ubiquitous meaning that objects and places will have easily accessible computing to mediate between users, data, and computation. Second, network bandwidth will increase, and the cost will decline dramatically. This bandwidth will allow data collection and efficient movement of data to computation and vice versa. Lastly, display pixels will be everywhere. Users will be able to walk to a display and use it seamlessly to review an image or video sent to your smart phone. All this computation will enable the use of multimodal interfaces that combine speech, touch, gesture, pens, and haptic interfaces. A video produced by Corning titled "A Day Made of Glass 2" contains many examples of interfaces and computation enabled by ubiquitous computing and communication [Corning 2012].

A major point I tried to make at the panel was that an abundance of computation and network bandwidth should be applied to improve remote collaboration. HD video conferencing is noticeably better than half-sized standard definition video conferencing. The question we should be asking is, "what happens when a group of 10 people join a session to collaborate in which there are 20 HD streams and a participant is able to access and manipulate content from anywhere in the world?" We know existing conferencing systems do not work that well. The research question is, "how can we apply computation and bandwidth to improve collaboration?"

Current TV and entertainment trends include video delivery over the Internet, second screen interaction (i.e., using a tablet to lookup information or interact with a program playing on a shared display), and synchronized remote program watching (i.e., make the viewing experience the same whether two people are in the same room or geographically separated). All of these trends will be enhanced as the availability of computation, network bandwidth, and display pixels grows and costs less.

4. RESEARCH DIRECTIONS

Throughout history it has been good idea to have a model or vision for the system or environment we hope to create. Vannevar Bush published a seminal paper in 1945 in which he foretold the possibility of online storage of documents and linking between elements. Doug Englebart presented a live demonstration, now called "The Mother of All Demos," at a Fall Joint Computer Conference in December 1968 at which he demonstrated the computer mouse, live video conferencing, hypertext and hypermedia, and collaborative real-time editing [MouseSite 2013]. These two visions motivated computer science research for the next 20 years. Alan Kay described his vision for a portable computer called the DynaBook in 1970 [Wikipedia 2013]. That vision accurately portrayed laptop computers, tablets, and smart phones. Because I am of the Star Trek generation, the Holodeck has long been my vision for remote collaboration, and the Corning video contains many ideas for a future that incorporates ubiquitous computing, remote collaboration, and a style of authoring that involves combining media (e.g., health records) and capturing live events.

While the past ten years have seen significant improvements, I suspect the next ten years will lead to even more dramatic changes in how we communicate and use computers. I think hardware improvements will be as important a force on research in the next ten years as it was in the past ten years.

REFERENCES

ADCOCK, J., COOPER, M., DENOUE, L., PIRSIAVASH, H., AND ROWE, L. A. 2010. TalkMiner: A search engine for online lecture video. In *Proceedings of the 18th ACM International Conference on Multimedia (MULTIMEDIA'10)*. ACM Press, New York, NY, 241–250.

BRANHAM, S. M. AND CARTER, S. 2012. Who makes, shares Internet how-to videos? In *Proceedings of the Workshop on Social Mobile Video and Panoramic Video*. http://www.fxpal.com/publications/FXPAL-PR-12-682.pdf.

BROOKS, J. 1975. *Telephone – The First Hundred Years*. Harper & Row, New York, NY.

CORNING INC. 2012. *A Day Made of Glass 2.* http://www.youtube.com/watch?v=jZkHpNnXLB0. (Last accessed 5/13).

FOX, A. AND PATTERSON, D. 2013. *Engineering Software as a Service: An Agile Approach Using Cloud Computing*. Strawberry Canyon, LLC, Berkeley, CA.

HOWCAST. 2013. Watch How. (Feb 2008). http://howcast.com/. (Last accessed 5/13).

MUKHOPADHYAY, S. AND SMITH, B. 1999. Passive capture and structuring of lectures. In *Proceedings of the 7th International Conference on ACM Multimedia (MULTIMEDIA'99)*. ACM Press, New York, NY, 477–487. DOI: http://dx.doi.org/10.1145/319463.319690.

MOUSESITE. 2013. http://sloan.stanford.edu/MouseSite/1968Demo.html. (Last accessed 6/13).

MUVEE TECHNOLOGIES PTE. LTD. 1999. Muvee – life expressed! http://www.muvee.com/. (Last accessed 5/13).

ROWE, L. A. AND JAIN, R. 2005. ACM SIGMM retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Commun. Appl. 1*, 1 3–13. DOI: http://dx.doi.org/10.1145/1047936.1047938.

SHIPMAN, F., GIRGENSOHN, A., AND WILCOX, L. 2003. Generation of interactive multi-level video summaries. In *Proceedings of the 11th ACM International Conference on Multimedia (MULTIMEDIA'03)*. ACM Press, New York, NY, 392–401. DOI: http://dx.doi.org/10.1145/957013.957096.

WALSH, T. 2011. *Unlocking the Gates*. Princeton University Press, Princeton, NJ.

WIKIPEDIA. 2013. Dynabook. http://en.wikipedia.org/wiki/Dynabook. (Last accessed 6/13).

YANG, Z., WU, W., NAHRSTEDT, K., KURILLO, G., AND BAJCSY, R. 2010. Enabling multi-party 3D tele-immersive environments with ViewCast. *ACM Trans. Multimedia Comput. Commun. Appl. 6*, 2, Article 7. DOI: http://dx.doi.org/10.1145/1671962.1671963.

Received May 2013; accepted May 2013