Nevenka Dimitrova
Philips Research

Refocusing Multimedia Research on Short Clips

Peter E. Hart, Kurt Piersol, and Jonathan J. Hull
Ricoh Innovations

There's an old joke about a drunk searching for his car keys at the wrong street corner late one night. When asked why, he explains that it's better lit than the right one.

The multimedia authoring research agenda today, like the drunk's search, has been unduly influenced by "where the light is." To push the analogy a bit further, the field is searching for the "keys"—the proverbial killer application—under the "light" of a large corpus of available (that is, network newscast or sports) footage.

We believe that multimedia's killer app might already be at hand—and it's focused around audio and video clips. But researchers aren't addressing critical open research issues because of the current focus on commercially produced, feature-length videos as an experimental corpus.

Searching for killer apps

The phrase *killer application* has a long history, going back to the early days of PCs. VisiCalc, the first commercial spreadsheet program implemented initially on the Apple II, is the standard example of a killer application. Everyone knows that a killer application must by definition provide such compelling value that users shift behavior and buying patterns to gain access to them. The connotations of the phrase imply a mass market and broad adoption. So when we talk about a killer application for multimedia, we tend to think in terms of some sort of consumer application that composes multimedia. Indeed, one of the ACM Special Interest Group on Multimedia's Grand Challenges is to "make authoring complex multimedia titles as easy as using a word processor or drawing program."[1]

The evident assumption here is that consumers wish to produce complex multimedia artifacts but are unable to do so because of the lack of adequate tools. If only more powerful and easier to use technical tools existed, the implicit argument runs, a mass market for them awaits.

We take issue with this assumption. It's not at all clear to us that a mass market exists for producing complex multimedia titles, any more than a mass market exists for tools to help produce books. You might argue that word processors serve this purpose, but anyone who has written a book can testify that a word processor isn't an ideal tool for book production and layout. Instead, word processors serve as a sort of input system for the more complex professional tools that professional designers use to produce finished books. Word processors are used to prepare memos, papers, and other shorter works much more frequently than to prepare book-length drafts.[2]

Is multimedia's killer app already here?

If we shift attention from a consumer interested in recording the family vacation to a knowledge worker who needs to communicate at a professional level, we can identify a killer app. Multimedia segments are already included in PowerPoint presentations or other text-based documents.

Originally, these documents were a series of relatively static page images. As time passed, animations became more prevalent in such presentations. Today, it's rare for high-stakes presentations to include no video segments at all.

These media segments are typically short, usually 10 to 30 seconds long. They might demon-

Editor's Note

In the ever-daunting quest for a killer application, the focus has mostly been on the application. However, we should stop to ask where the killer data comes from. In this article, the authors draw attention to an important aspect of multimedia: the role of corpora on the multimedia research agenda.

—Nevenka Dimitrova

1070-986X/05/$20.00 © 2005 IEEE

Published by the IEEE Computer Society

strate a product vision or, perhaps, an experimental result. They're typically low quality by the standards of professional videographers and producers but are more than adequate when measured by the standards of a typical PowerPoint slide set. They aren't difficult to create, nor are they difficult to describe or manage. Usually, each segment consists of a single scene with no camera motion or scene transitions. Occasionally, they have two or three scenes, but seldom more.

What intrigues us about these short segments, or clips, is that they're usually integrated into much longer compositions. They're part of an hour-long presentation but aren't composed using a professional videographer's techniques. Instead, they're composed using the techniques of the office professional. PowerPoint, OpenOffice, and Firefox are all killer applications, every one of them suitable for including multimedia elements, every one ready for ordinary office professionals to use.

In his book *The Man Who Mistook His Wife for a Hat* (Touchstone, 1998), Oliver Sacks tells the story of a man who can't complete a gestalt. He can recognize the parts of an object but can't make the leap to identifying the whole. He can recognize that an object has a long green stem with thorns and a red blossom at the end, and yet he can't identify it as a rose. Have we made a similar mistake, being unaware that killer applications already exist that can compose multiple forms of media into a complex title?

We think it's safe to answer with a yes. The emergence of podcasting is a clear example. The practice of producing Web logs (blogs) has been growing for several years now. In October 2004, Adam Curry embedded MP3 audio streams into such blogs' Really Simple Syndication (RSS) feeds, allowing a special client to automatically load such audio clips into an MP3 player, specifically the Apple iPod. Users can aggregate such podcasts into various existing blogs and reassemble them into topic-based collections of clips with typical blogging techniques.

The rate of adoption is striking. Feedburner, a blog traffic-monitoring company (http://www.burningdoor.com/feedburner/archives/001029.html), recently noted that more than 1,700 podcast feeds appeared within the first six months after the invention of podcasting. This is approximately six times the rate of adoption seen in the Web's early days, even though the technical barriers to entry appear similar.[3] The archive notes that a podcast's average size is approximately 7 Mbytes, roughly 7 minutes long. Many are longer, many shorter. This is another example of users adding multimedia clips into an existing killer application.

Corpora's influence

First-time attendees at a multimedia research conference might be surprised to find that their attendance is an effective way of gaining access to newscast footage. Several interrelated reasons explain why experimentalists choose this subject material:

- this genre generates so much content that it's easy to get a large corpus;
- such content is attractive because of its relatively simple and regular format; and
- perhaps most importantly, most newscasts have high-quality metadata in the form of closed-caption information that researchers can use either as a primary search pathway or for evaluating the results of other experimental retrieval methods.

Readers need only skim recent multimedia publications to appreciate how the choice of corpus and genre influences the identification of research topics and the reported results.

Of course, there are corpora that aren't primarily composed of news footage. Some widely used corpora provide video and audio of soccer games, movies, presentations, lectures, and interviews. Yet, a certain uniformity still exists in these corpora. They focus on long titles, with somewhat rigid formats. They're captured in controlled environments with good acoustics. A skilled team, using good if not professional production techniques, captures the contents. Examples of this are available at the TREC Video Retrieval Evaluation (TRECVID) homepage at http://www-nlpir.nist.gov/projects/trecvid/.

Such footage is easy to assemble, by design. For example, if we expect to produce a half hour or an hour of quality video each day, then we must limit the choices to a fairly narrow range of options to succeed. A simple wrapper format, with an announcer stringing short film clips together, streamlines the production process.

Indeed, if we choose video as the output format for a multimedia title, then a newscast would be the rough equivalent of a word processor template or a business form: Fill in appropriate material to produce the desired finished

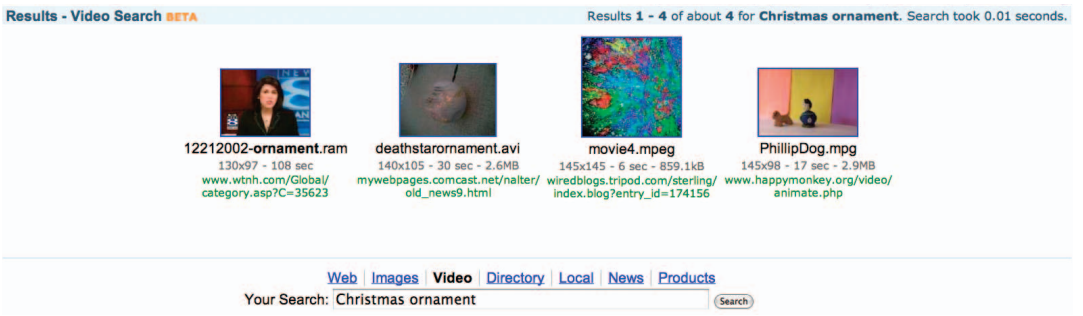

Figure 1. Video search results using the phrase "Christmas ornament."

product. Is it surprising that we find such video easy to disassemble and analyze?

The uniformity of our corpora, combined with the assumption that composition of complex titles is the killer application, suggest that multimedia researchers might not be focusing on the right problems. We're busy attempting to make it easy for people to assemble and disassemble lengthy movies. What we should be doing, instead, is trying to make it easy for people to effectively produce long documents of any type that include typically brief video and audio clips.

World of clips

So what are the implications of our observations for multimedia research? Are we suggesting a new research agenda or being merely querulous?

A world centered on clips is one in which finding the right clip for your purpose is of paramount importance. It's easy to imagine that someone has already solved this problem. After all, numerous Web sites devoted to clip art and stock photographs exist, as well as a smaller number that offer stock footage. But from our experience, none of these sites makes it particularly easy to find a desired clip.

The standard retrieval technique used at the Google and AltaVista sites, for example, analyzes HTML content surrounding multimedia. This can lead to unsatisfactory results: Figure 1 displays the results of searching Yahoo for video using the phrase "Christmas ornament." These results are uninspiring, and while such sites can produce better ones, this isn't a contrived example. We should certainly be able to do better.

We might approach this by recognizing the relationship of the clips to one another and to the collection. Consider how users of normal clip art often have their own personal favorites. Shared information about these subsets could provide an Amazon-like feature that says, "People who chose this clip also chose that clip." This could be useful in composing new compound documents. For example, the clip showing someone opening a door and walking into a building could be coupled with the clip showing that person walking out of the building and closing the door.

This particular approach assumes a shared database rather than a personal collection, but it might be possible to share metadata about clip usage within documents without sharing the clips themselves. We know that access to shared video footage isn't always practical. If Peter adds the opening door clip from his own database, how can Paul take advantage of the relationship information without access to the clip? He might use the metadata embedded in each clip to initiate a search of his own database for an appropriate door-closing sequence.

Including cross-modal information is another useful way to consider clip interrelationship, especially when targeted for a single user's collection. For example, a particular video clip could automatically suggest a musical clip (for background music) from the user's personal collection. This matching process could then use visual characteristics of the video (quick scene changes at the beginning, slow toward the end) to find musical clips with a matching tempo.

These modest ideas can be extended to form the basis for a more comprehensive research agenda.

Casual curators

Commercial footage and clip art sites often produce more appealing results in response to search queries than the Yahoo example in Figure 1. Such sites have staffs of professional curators who organize and annotate collections to enable effective retrieval. Of course, their focus is on commercially viable collections. Still, the idea of a more personal form of curation deserves further thought.

Curating is a difficult task. Curators must not only organize, but explain, because we can't fore-

see all the potential uses of a particular object. A user might not be able to find items in a collection easily (if at all) without an overview of what's available. Good curators provide basic education along with the organization so that people can discover the valuable gems available in a collection of material. They synthesize, draw conclusions, and provide a context to the items they curate. They have an editorial point of view.[4]

Many people assume the role of curator at some point in their life. As more people produce short video segments with media recorders, smart phones, and screen capture utilities, we'll find ourselves with large personal or group libraries of content. Who else can provide an essential curator's work, turning a pile of media into a reusable set of valuable content?

If we could provide an assistant to the casual curator, we would greatly simplify the task. Media metadata are clearly a key, but are we searching for the kinds of metadata that make curatorship easy? A casual curator doesn't need to manage vast collections of information, on the scale of Google or a similar general search engine. Categorization is more important to curators than retrieval because, once they've specified a category, it's quite manageable to examine each element in a suitably sized collection.

How might we attack video categorization in this context? It's well known, for example, that general face recognition in video is a hard problem. However, we can build a trainable recognizer that's good at recognizing only a few faces with about 90 percent accuracy.[5] Such a recognizer is probably useless for security purposes and useless to a search engine company. But the curator of the Albert Einstein Archives might find great use in a system that lets him test films easily for the presence of Einstein's face and extract valuable clips. In a wider adoption context, we might use it to test a home video collection for all video of a particular family member.

Voice-print recognition—again a much simpler problem than general speech recognition—would allow a curator to categorize based on the presence of particular voices. Indeed, voice-print recognition could also point out the presence of common and unidentified voices. This begins to approximate a curator's behavior, pointing out a collection's most important properties.

We can suggest several properties that a casual curation system might possess. It must provide categorization. It should, ideally, provide an overview of a collection, a summary of its contents, and a basic organization. It should point out contexts and relationships, which might not be obvious to the casual observer. Curators often consider deep reorganization of the content in their collections, so the system should make this easy. Another key function of the curator is to "publish" the results of their efforts as exhibits, which attracts attention to the whole.

These goals would likely be technically too ambitious if we situated them within a Web-scale application, but they seem quite approachable today if situated within an individual or small-group context. Of course, it's important to keep the user's task in mind, including the time available, intention, and so on. Clearly, the idea of a personal or casual curator, taken seriously, begins to suggest the outlines of a substantial research agenda.

Shareable, multimedia clip corpus

Research progress in experimental fields such as multimedia has always been greatly facilitated by the availability of shareable experimental data sets. More than 40 years ago, optical handwriting recognition research benefited greatly from the work of W.H. Highleyman,[6] which facilitated great leaps forward in the field. This is where the issue of corpora and their construction comes into play. W. John MacMullen[7] suggests a scientific basis for designing corpora based on several criteria—the goal being to produce a reservoir of data useful for replicating experiments. His first criterion is representativeness, which he describes as critical.

If we want to produce effective systems for the casual curation of multimedia content, we need to understand what kind of content is representative of our existing killer applications. The most reused content is likely rich in information content but not in structure. The most reusable video is probably almost devoid of the structure we find in newscasts, interviews, or commercial films. A valuable research topic would be to analyze the contents of included media in a large collection of ordinary business documents to determine whether these assertions about structure are true.

Given a better understanding of representative content, a second valuable research task would be to construct a representative corpus. This would be a corpus of media segments that represents reusable content widely used in office environments. We'll need to consider many issues of intellectual property, privacy, and sampling. However, without such a corpus, we'll

Additional Research Areas

Besides the role of the corpora in multimedia research and how it affects its agenda, we feel three other areas also merit special attention.

The first major area is printed representations of compound documents that include time-based media. The widely adopted productivity applications of today are in many respects primitive in their integration of video and audio. The printed representation of dynamic media has been a topic of recent research. Video Paper[1] is a good example of such a representation (see Figure A). How can we use such summarization techniques in a compound document setting? Normal productivity applications have printing models that aren't well suited for including extensive summaries of this sort, and yet we feel strongly that users will want printed representations of such documents.

Figure A. Video paper description for a recorded meeting.

The second area for future consideration is integrating spatial and temporal elements in compound documents. For some time, we have been satisfied with drawing a border around the edges of a video display box and calling the result media integration. This appears to be an area ripe for future examination. For example, can we conceive of a spatial equivalent to the transitions used to stitch together video sequences? Simple alpha blending is one possibility, but probably many more exist. The techniques employed by Apple's Exposé system of window management, with live windows moving apart to make everything visible, show us that many things are possible using animation and scaling effects. Can we think of new editing and interaction techniques that integrate time-based media into the typical flows, tables, and sequences of common static document types?

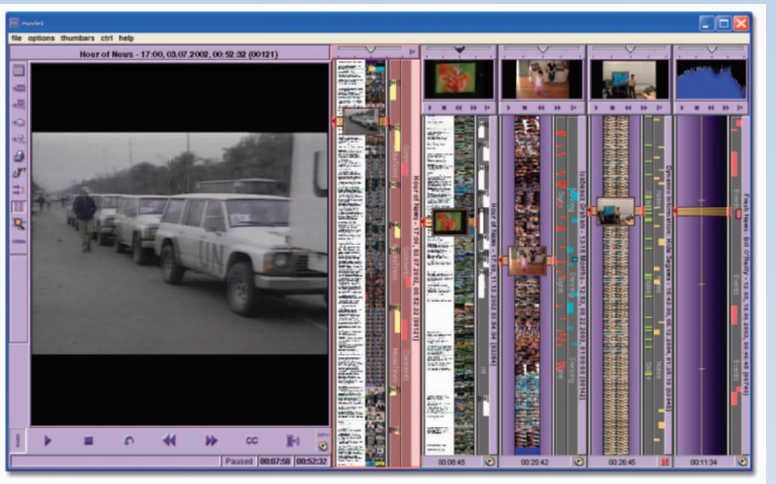

Figure B. MuVIE interface showing multiple clips of different modalities.

The third major area is user interfaces for curating collections of clips. Research in this area would provide easy-to-use tools for organizing, browsing, and skimming multiple clips simultaneously. Our own work on the MuVIE Multimedia Visualization Environment is one effort in this direction (see Figure B).[2] Future work should consider how its multitrack-feature-based visualization could support a searching strategy that would, for example, let a user find a clip that shows two business people shaking hands while there's loud clapping on the audio track.

References

1. J. Graham and J.J. Hull, "The Video Paper Multimedia Playback System," *Proc. ACM Multimedia 2003*, ACM Press, 2003, pp. 94–95.
2. J. Graham, *MuVIE: A Multimedia Visualization and Integration Environment,* Ricoh CRC Tech. Report 0451, 14 Feb. 2005; http://www.rii.ricoh.com/~jamey/crc-tr-0451.pdf.

have to rely on much less rigorous methods to evaluate our system's effectiveness.

We suggest designing the corpus with several goals. The data contained should consist of the actual contents of authentic office documents, taken from publicly available sources. If possible, it should mimic the contents of a typical user's hard drive, with collections of music, images, and video as well as documents. We need not mimic a hard drive's structure—merely include appropriate

quantities and mixtures of media content. An ideal, but difficult task, would be to obtain rights to the actual content of a series of user disks.

The content should be readily accessible and free of intellectual property encumbrances. The Creative Commons (http://www.creativecommons.org) provides a good model for licensing such material for appropriate uses. It's extraordinarily difficult, in the current climate of digital rights debates, to actually obtain content for research that we can legally share with others. One solution would include contributions from public-spirited citizens (such as those Dick Bulterman[8] mentions), who would authorize use of their data for research purposes.

The Caviar project (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/) makes a good example corpus available. The subject matter would only be useful to a small subset of the ordinary office user population, but the data set consists of clips of the proper length and the expected number of scenes. Its data set is multivalent and has rich metadata for each clip.

Search where the keys really are

We hope we've given a convincing argument that a focus on clips and their integration into existing applications would be valuable. We focused on the notion of clips, and most especially their retrieval, as a key enabler for the wide adoption of multimedia. Other researchers have had complementary ideas in recent years, which bear some mention. Shih-Fu Chang's ideas about impact criteria are interesting when applied to a collection of clips, leading to conclusions similar to those we present here.[9] Nevenka Dimitrova suggests an enhanced emphasis on memory and context, and many of her ideas for clip curation rely on just such added information.[10]

Retrieval, however, is certainly not the only area that merits additional thought. Indeed, it's one of four areas that we feel are important to consider. (See the "Additional Research Areas" sidebar for a more detailed discussion of the other three.)

Conclusion

So let's return to the drunk looking for his car keys—the killer app—under the lamppost of currently popular corpora. Our argument is that we can already dimly perceive the outline of the car keys—multimedia clips within a larger, though otherwise conventional, document. But we won't see the keys in detail until we move the lamppost or, better yet, construct an entirely new one. Our proposed new lamppost is personal clip libraries and supporting technologies, curated by individuals, who enable these killer apps. **MM**

Acknowledgments

We acknowledge the substantive guidance and comments of Nevenka Dimitrova in the preparation of this article, which helped to improve it significantly.

References

1. L. Rowe and R. Jain, *ACM SIGMM Retreat Report on Future Directions in Multimedia Research*, http://bmrc.berkeley.edu/~larry/sigmm-retreat/sigmm-retreat03-final.htm.
2. E.D. Lara, D. Wallach, and W. Zwaenepoel, "Opportunities for Bandwidth Adaptation in Microsoft Office Documents," *Proc. 4th Usenix Windows Systems Symp.*, Usenix Assoc., 2000; http://www.usenix.org/publications/library/proceedings/usenix-win2000/delara.html.
3. R.T. Griffiths, *History of the Internet, Internet for Historians* (*and Just About Everyone Else*), http://www.let.leidenuniv.nl/history/ivh/frame_theorie.html.
4. G.E. Burcaw, *Introduction to Museum Work*, 3rd ed., AltaMira Press, 1997.
5. J. Huang, B. Heisele, and V. Blanz, "Component-Based Face Recognition with 3D Morphable Models," *Proc. Int'l Conf. Audio- and Video-Based Person Authentication*, Springer-Verlag, 2003, pp. 27-34.
6. W.H. Highleyman, "Data for Character Recognition Studies," *IRE Trans. Electronic Computers*, vol. EC-12, April 1963, p. 135.
7. W.J. MacMullen, *Requirements Definition and Design Criteria for Test Corpora in Information Science*, SILS Tech. Report 2003-03, School of Information and Library Science, Univ. of North Carolina at Chapel Hill, 2003.
8. D.C.A. Bulterman, "Is It Time for a Moratorium on Metadata?" *IEEE MultiMedia*, vol. 11, no. 4, pp. 10–17.
9. S.-F. Chang, "The Holy Grail of Content-Based Multimedia Analysis," *IEEE MultiMedia*, vol. 9, no. 2, 2002, pp. 6–10.
10. N. Dimitrova, "Context and Memory in Multimedia Content Analysis," *IEEE MultiMedia*, vol. 11, no. 3, 2004, pp. 7–11.

Readers may contact Peter Hart at hart@rii.ricoh.com.

Contact Visions and Views department editor Nevenka Dimitrova at nevenka.dimitrova@philips.com.