Editor: Nevenka Dimitrova
Phillips Research

The Holy Grail of Content-Based Media Analysis

Shih-Fu Chang
Columbia University

Tools and systems for content-based access to multimedia and—image, video, audio, graphics, text, and any number of combinations—has increased in the last decade. We've seen a common theme of developing automatic analysis techniques for deriving metadata (data describing information in the content at both syntactic and semantic levels). Such metadata facilitates developing innovative tools and systems for multimedia information retrieval, summarization, delivery, and manipulation. Many interesting demonstrations of potential applications and services have emerged—finding images visually similar to a chosen picture (or sketch); summarizing videos with thumbnails of keyframes; finding video clips of a specific event, story, or person; and producing a two-minute skim of an hour-long program. (*Audio–visual skims* are condensed media clips that summarize information in the content.)

There's much excitement and buzz created by these fancy applications. But people are always asking, What will be engineering's Holy Grail for content-based media analysis in practical applications? My response is that the answer is complex. It's less about a specific algorithm or service than a rigorous methodology to formulate and evaluate content-based analysis research.

Content chain

To evaluate content-based research methodologies, we must first consider who the intended users are and whether alternative solutions exist. Hence, it's important to consider each solution within the context of the content chain, the process starting with acquisition, followed by production, processing, and finally consumption. Figure 1 shows a diagram depicting the content chain and the relationships among different components. Each piece exists in part of the

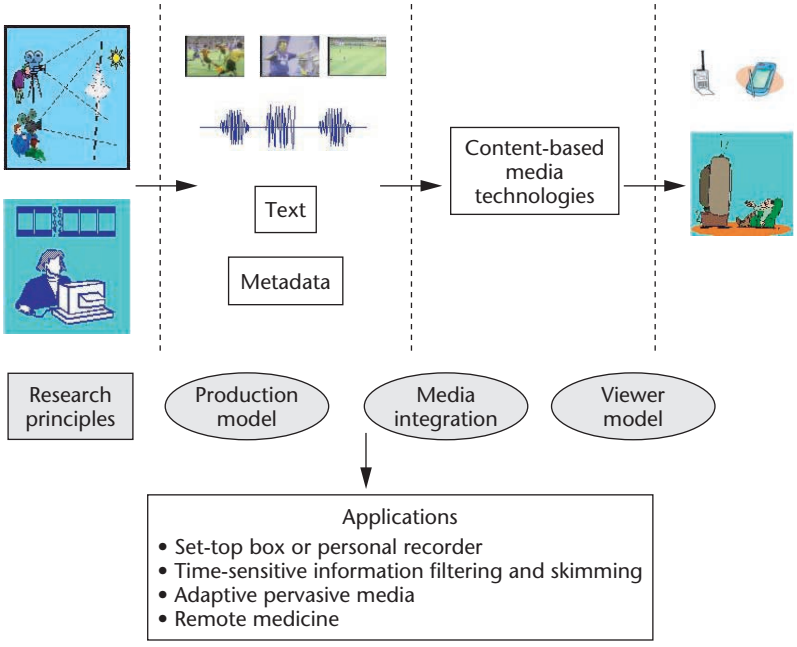

Figure 1. The end-to-end content chain, content-based research principles, and potential applications.

1070-986X/02/$17.00 © 2002 IEEE

chain and is produced by specific production sources and methods, associated with multiple media modalities (audio, images, video, graphics, text, and so forth), and intended for specific user groups. It's important to remember that most often users don't deal with raw content just coming off the sensing devices—many content processing stages have already occurred. Figure 1 also highlights important principles critical for effective content analysis:

- understanding production models,
- fusing multimedia data, and
- exploring perceptual viewer models.

Areas of research

By content-based media analysis, I refer to research in the following areas:

- *Reverse engineering of the media capturing and editing processes*. Works in this area attempt to reverse engineer the capturing and editing processes, and recover the constituent content components such as shots, scenes, and structural elements (dialogues, anchors, and so forth) in the video. By breaking videos into atomic entities at different levels, we can develop intuitive and efficient access tools.

- *Extracting and matching objects*. Similar to breaking documents into words or phrases, we can decompose images and videos into objects, from which we can derive comprehensive attributes. Much work has been pursued to define adequate features and criteria for matching audio–visual content based on audio–visual properties and spatiotemporal relationships.

- *Meaning decoding and automatic annotation*. This area has captured much interest in recent years in recognizing the need to provide semantic-level interaction between users and content. Instead of processing the syntactic-level entities (such as shots, scenes, or objects), here we ask, Can we teach the computer to recognize the meaning of the content at the generic semantic level (people, location, and so on) or specific semantic level in constrained domains? If we have the luxury of a large amount of training data, can the computer automatically discover interesting patterns or outliers in the content? Can we develop techniques to automate the painstaking process of taxonomy construction?

- *Analysis and retrieval with user feedback*. This topic closely relates to feature extraction, matching, and semantic classification. The difference is the emphasis on keeping users in the loop of analysis and retrieval. Among the three entities involved—data, learning agents, and users—many interesting issues exist: efficient visualization and manipulation tools for high-dimensional data, adequate concept representation models, and effective learning methods. The goal is to develop a new analysis framework (enhanced by interactive feedback and learning), represent a personalized information target, and hence, facilitate better retrieval performance.

- *Generating time-compressed skims*. Driven by the trends in mobile communication and personal media applications, several interesting systems have come along that condense long video programs into short skims. Research in this area involves issues broader than pure analysis—understanding user-perception characteristics and matching output with the user's information needs are also critical. Many important questions arise—How much semantic understanding of the content does the system require for automatic skim generation? Will syntactic-level skims that only explore the production syntax (without recognizing the semantics in the video) be useful in any subset of applications?

This list is by no means complete. This is a highly dynamic and interdisciplinary field. Many important related areas exist, such as efficient indexing for large databases, content adaptation for accessing multimedia over heterogeneous devices, and standards for specifying content description language and schemes like MPEG-7.

Impact criteria

Now that we've discussed the context and types of research, let's reformulate the question posed at the beginning. What are the criteria for content-based analysis research to produce high impact? When mentioning impact, here I'm focusing more on practical aspects and less on the purely intellectual ones.

- *Generating metadata not available from production.* Many of us have worked hard to develop systems for video-shot segmentation, claiming their superiority in terms of accuracy, simplicity, or interoperability. However, for future content produced by digital devices, the work may become irrelevant. Shot boundary, time, location, and even photographer information may be included as output by the new generation of capturing devices. The same view applies to other types of metadata that might be preserved in digital editing tools, such as the object layer information from image editing software and the edit list from video editing software. Hence, content-based analysis shouldn't duplicate metadata that's readily available from capture or production devices.

- *Providing metadata that humans aren't good at generating.* We can come up with thousands of words to describe what we see or hear in the content. However, humans aren't as good as computers at processing, manipulating, or measuring entities at the low level with precision (such as quantitative representations of features or ranking distances of a large set of objects). It probably takes several hours for a person to annotate a one-hour video program with reasonably sufficient details. Asking an operator to annotate live streams at the same speed as real-time encoders is probably a tough task. Hence, viable content-based research should attempt to produce annotations not easily constructed by humans, generate them with better accuracy, and operate at a speed faster than humans.

- *Focusing on content with large volume and low individual value.* In balancing the trade-offs between manual and automatic processes, financial resources play an important role. If the value for each piece of content is high, content owners can simply invest enough financial resources for manual annotation. Hollywood films made with megabudgets are good examples. Rare archival content with historic value is another example that can afford extensive manual annotation.

- *Adopting well-defined tasks and performance metrics.* Automatic systems work in an objective, quantitative way, requiring well-defined tasks and measurable performance. Whether it's information retrieval, summarization, or filtering, the tasks have to be quantifiable and computable. In comparing alternative solutions, practitioners need accepted performance metrics to justify the use of automatic tools in lieu of manual ones. The Text Retrieval Conference (TREC) approach used in the information retrieval community offers an excellent example and shows the importance of well-defined tasks and evaluations.

  The issues for multimedia retrieval, however, are more complex. Compared to textual information retrieval, it's more difficult to collect multimedia content. Content owners tend to guard the intellectual property rights with greater security. The consumption models and user tasks for different genres of multimedia present a greater level of diversity than the text domain. I hope that the advances in algorithms for intellectual property rights management tools, coupled with their greater deployment, can resolve this issue.

Satisfying the impact criteria I listed doesn't automatically guarantee high-impact technological solutions in applications that must meet practical requirements and specific user needs. Here, I stress the importance of keeping a broad view toward applications. Traditional views of content-based technologies focus on search and retrieval—which is important but relatively narrow. I argue that almost every stage of the content chain will benefit from content-based media analysis—from intelligent acquisition, computer-assisted media editing, coding, streaming, quality enhancement, and finally to display. As an example, in the "Columbia Digital Video Multimedia Projects" sidebar, I describe a project about content-adaptive streaming live sports video. Overall, I advocate a broadly defined term—content-based media technology—that encapsulates engineering solutions that explore the synergy provided by automatic or semiautomatic content analysis.

Conclusions

In seeking the Holy Grail for research of automatic media content analysis, I've come up with a list of criteria for checking potential impact. I'm sure the list will be controversial and remain to be validated in practice. (I welcome comments from readers with supporting or contrasting views.) However, I want to stress the underlying ideas and objectives. The impact criteria's formation is closely driven by the concept of empha-

Columbia Digital Video Multimedia Projects

Several projects are currently underway in the Digital Video Multimedia (DVMM) lab at Columbia University as examples validating the criteria listed in the "Impact criteria" section. (To learn more about these projects, visit http://www.ee.columbia.edu/dvmm.) Although some projects might not fare as well as others, they're equally exciting from a scientific research perspective.

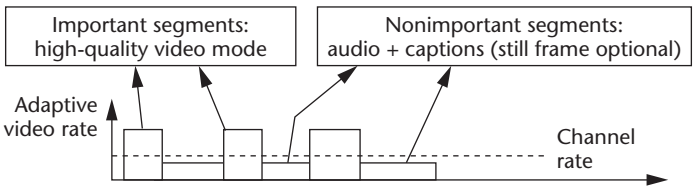

Figure A. Content-adaptive streaming of live video.

Content-based live sports video filtering

Sports video is a big attraction internationally, with massive production and a huge audience base. The per-piece value of sports video is high. However, it drops significantly after broadcasting the event and knowing the outcome. One opportunity for automatic solutions is real-time detection of events and highlights from live video. In our projects, we developed new algorithms and real-time software to detect fundamental semantic units—such as pitching in baseball and serving in tennis—to recognize the text score box information embedded in the images and to detect the boundaries of plays or breaks in the soccer games.

Off-the-shelf solutions exist that use manual approaches for collecting sports statistics. These solutions can manually generate metadata about the games on site (such as score information). However, real-time performance and the potential for significant cost savings make the automatic systems attractive. Additionally, the automatic system can recognize direct information in the visual content (including richer information than just a few fixed annotations made by the annotator).

In the future, engineers may incorporate digital facilities into the sports production processes so that the processes preserve information like game statistics at the production's origin point (such as production trucks). In such cases, the automatic analysis systems' potential impact will greatly diminish.

Real-time sports video information is useful in many applications. We recently advocated the concept of content-adaptive streaming, which allocates bandwidth unequally within a video stream according to the importance of the content in different segments.[1] For example, in wireless streaming of baseball, we can send high bitrate full-quality video during segments showing pitching and important follow-up activities but change to low-bandwidth mode using audio only (and maybe adding keyframes) during unimportant video segments. Figure A shows a diagram illustrating such a concept. By doing so, we can reduce the average stream bandwidth, thus making the video suitable for mobile applications. The performance gain comes from the low percentage of important segments in the video program (fewer than 20 percent for baseball). Just imagine how much time we can save by cutting portions of the video that don't include important activities.

Medical video indexing and summarizing

Some researchers conjecture that the medical domain is a promising area for content-based technologies. In this domain, tasks are defined well—doctors and health care workers want to find specific images or videos for the purpose of diagnosis, prognosis, or research. The data semantics are also defined thoroughly from a clinical perspective—that is, they define what's normal versus abnormal in specific clinical categories. A significant skill threshold exists for personnel qualified for the annotation process. Finally, unlike large-budget commercial films, the domain has voluminous data but each piece of content typically doesn't have significant value (at least for the health-care providers).

In the digital echocardiography video project,[2] we developed techniques to detect and recognize constituent views in each video, extract clinically important frames, and generate time-compressed clinical summaries (we show the system architecture in Figure B, next page). Such tools facilitate many exciting applications. In remote medicine, we can send clinically meaningful time-compressed clips, showing only important information to remote experts. In computer-aided diagnosis, we can build systems to discover prior medical cases that may have similar spatiotemporal attributes revealing important clinical information.

continued on p. 10

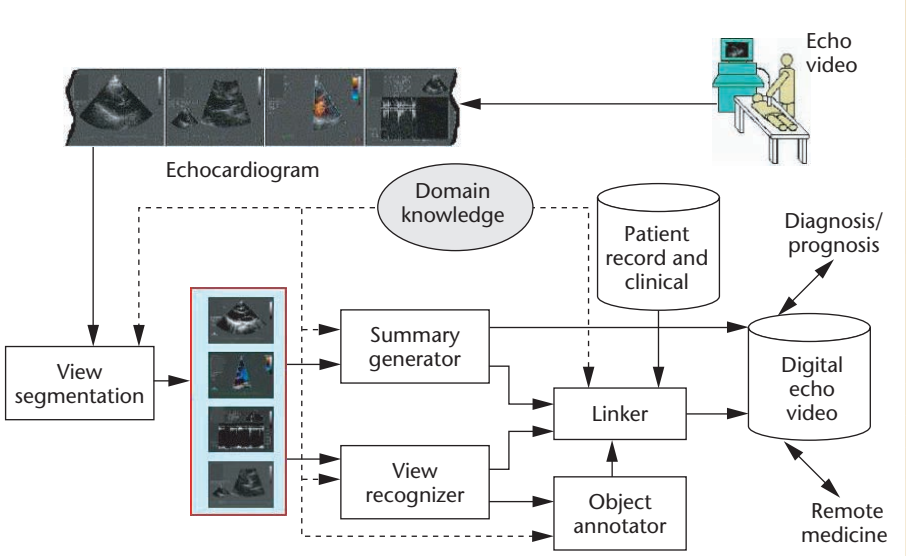

Figure B. System architecture of Columbia's content-based medical video system.

Computational parsing and skimming of films

Films violate most of the impact criteria. A movie audience typically has some knowledge about directors and content in the film. Annotating the meanings of a film might not be such an unwelcome task for humans. More importantly, when necessary, financial resources always exist to fund manual and extensive annotation processes.

Nonetheless, films offer a great domain for research in content analysis because of the many genres and production styles. Mapping various theories in film production to computational models and investigating effective tools for detection and classification provides an exciting avenue of scientific research. For example, we've developed technologies for computational scene analysis and summarization.[3] Many interesting issues arise. For instance, can we teach computers to

- measure the Hitchcockian level of a newly produced film,
- detect every time a particular cut style appears, or
- recommend some predefined video editing templates for nonprofessional users?

The last idea is perhaps too ambitious but appealing from the end user's point of view. For example, wouldn't it be great if I could quickly convert the videos I shot at my daughter's school and her music recitals into a short documentary-style film about her early school years? Such empowerment by providing better tools of video storytelling is undoubtedly compelling.

References

1. S.-F. Chang, D. Zhong, and R. Kumar, "Real-Time Content-Based Adaptive Streaming of Sports Video," *Proc. IEEE Workshop Content-Based Access to Video and Image Libraries*, IEEE CS Press, Los Alamitos, Calif., 2001.
2. S. Ebadollahi et al., "Echocardiogram Video Summarization," *Proc. SPIE Medical Imaging,* SPIE Press, Bellingham, Wash., 2001.
3. H. Sundaram and S.-F. Chang, "Determining Computable Scenes in Films and Their Structures Using Audio Visual Memory Models," *Proc. ACM Multimedia*, ACM Press, New York, 2000.

sizing the end-to-end content chain and the many issues evolving around it. What's the best way to integrate manual and automatic solutions in different parts of the chain? What's the adequate computational model for content production and human's perception in each domain? How should we combine different media modalities in analysis and generation? Engineers who attempt to answer these questions will benefit from applying the criteria list to verify the potential impact in any specific domain. Hopefully they can identify more adequate objectives and fruitful technical directions as well. **MM**

Acknowledgments

I sincerely thank Nevenka Dimitrova for her encouragement and discussion in shaping some of the views described in this article. I also appreciate Alejandro Jaimes and Hari Sundaram for contributing their valuable comments.

Readers may contact Shih-Fu Chang at the Digital Video and Multimedia Research Lab, Dept. of Electrical Engineering, Columbia Univ., New York, NY 10027, email sfchang@ee.columbia.edu.

Contact Visions and Views editor Nevenka Dimitrova at Phillips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510, email nevenka.dimitrova@philips.com.