Media Semantics: Who Needs It and Why?

Chitra Dorai, Andreas Mauthe, Frank Nack, Lloyd Rutledge,
Thomas Sikora, Herbert Zettl
dorai@watson.ibm.com, andreas.mauthe@gmx.de, frank.nack@cwi.nl, lloyd.rutledge@cwi.nl,
sikora@hhi.de, hzettl@sfsu.edu

Organizers: Chitra Dorai and Frank Nack

Introduction

As pointed out in the keynote address at the 2001 ACM Multimedia Conference [3], the current major goal of multimedia research is directed towards provisioning information for pervasive access and use. To achieve this, what will become important are technologies that help sift useful nuggets of information from torrents of media data, which can be turned into valuable knowledge just in time and need, and tools that help provide access to these nuggets in anytime - anywhere - any device mode to everyone ranging from enterprise customers to independent consumers. Further we need to treat various media on an equal basis in environments that provide multimedia-based interactions, where they ultimately add value to users, whatever the nature of the interactions may be and whatever the preferred mode of media access may be. Thus, there is a fundamental need to investigate the means to elucidate, sublimate, or rationalize information and knowledge from media data. However, current user expectations are far from being met owing to generic low-level content metadata available from automated processing that deal only with representing perceived content, and not the semantics of it.

The challenges in modeling and extracting media-intrinsic as well as extrinsic semantics are as complicated as the attendant problem of matching them with user needs in various domains in which we anticipate robust media usage [4]. Moreover, once information is gathered from various repositories in a federated fashion, we also need mechanisms to automatically process the disparate data for generating visually pleasing media presentations, and for need-oriented and device-appropriate media playback. The tools for designing and developing such technologies are in their infancy and their development depends very much on a better understanding of user requirements, domain needs, objective measurements of media items as well as subjective interpretations.

One promising approach at bridging the semantic gap and building high-level semantic descriptions for reliable content location, access, and navigation services is founded on an understanding of media elements and their roles in synthesizing meaning, manipulating perceptions, and crafting messages, with a systematic study of media productions [1]. The two broadest attempts for a standardized description of content, the Semantic Web from the W3C [5], and MPEG-7 from ISO [2], also provide some insight into these problems but only briefly touch on the problem of underlying semantics needs in different domains. The goal of this panel is to explore, discuss and come to a better understanding of the following issues:

- How do we relate the issue of semantic gap in media research with the market place need to build an effective “Media-Google”?
- What kinds of semantic descriptions do domains with real business needs require? Or in other words what are the appropriate levels of semantic descriptions to satisfy those needs of users?
- How do content creators craft semantics in their material, i.e., what are their processes to compose meaning?
- Which tools do they use and what do they still need?
- Which presentation/discourse tools do we need and which level of semantic descriptions do they require?
- How should the current state of the art in Semantic Web and MPEG-7 standards be extended in a way that makes user, domain, and presentation-specific ontologies applicable for describing multimedia content?
- How should domain and multimedia ontologies be related to ensure that terms from multiple ontologies could be used within a single instance?
- How far can the new approaches such as production knowledge and film grammar-based content analysis be pushed?
- How do we build technologies for production-time (just-in-time) annotation tools, thus avoiding annotating material as an afterthought in the process of content creation?
- Which technologies do we need to finally allow general public, the access to annotated media material for the purpose of discourse in encyclopedic spaces?

The members of this panel are stakeholders representing the whole multimedia value chain, and involved in content creation, authoring, media production, content management, distribution, interactive knowledge spaces, media standards, and presentation. Each of these fields presents various facets of the issues outlined above that will help us to explore the need and importance of media semantics in the broadest sense, identify fruitful research directions and propose some exciting solutions.

Panel Statement: Andreas Mauthe

Computer based systems to support the (digital) production and archiving of content are only slowly and with much hesitation being adopted by the broadcast and media industry. Reasons for this are a general apprehension towards new technologies and a fear of misdirected investments into the "wrong" system. But why are these systems considered to be insufficient? The solutions currently being available have two major drawbacks, viz. i.) they use proprietary largely non-open technologies and ii.) the applications do not support the search and handling of media in a natural way. The latter is mainly due to a lack of understanding and consideration of media semantics.

Content is described either using ridged frameworks and special terminologies only experts can understand, or using free text and keyword annotations that capture only a fraction of the actual semantic of the object. Although visual tools such as key frames are used for content retrieval and representation, the search for content in large content pools is still tedious and success is a matter of chance.

New retrieval techniques such as Image Similarity Retrieval and Search by Example promise better results closer to what users are actually looking for. However, since they are largely based on mathematical functions that do not consider media semantics the result of such an operation is often amusing or plainly annoying. This is similar for automatic analysis, indexing and documentation tools that are quite good as long as statistical operations can be performed but mostly fail to consider context and semantics.

Hence media professionals use these technologies to support their work to a certain extent but rely for most critical tasks that require knowledge of context and semantics on human input and interaction. In future this might mean that some of the ever-increasing amount of content will become inaccessible because of a lack of human resources to perform the task of describing and handling it.

Biography

Dr. Andreas Mauthe holds a Master (Dipl-Wirtsch-Inf.) in Applied Economics and Computer Science from the University of Mannheim, Germany and a PhD in Computer Science from Lancaster University, UK. His research area has been multimedia systems, group and multipeer communications, and QoS. In 1997 he joined the Content Management Division of tecmath (now blue order) originally working on two European projects (CARO and OPAL) dealing with the problems of content exchange and management, collaboration and Web presence of radio and television broadcasters. Subsequently he became Chief Development Officer in the same company responsible for the development of media archive, a distributed content management system. In this capacity he also oversaw a number of large-scale customer projects at major public and private broadcasters throughout Europe. From January 2001 until March 2002 he was responsible for the UK Operations of the same company. As member of the division board he was also involved in the strategic planning of the product development and business development of the company. Since leaving blue order in March 2002 he has been pursuing research in the area of Content Delivery Networks, Peer2Peer Systems and Distributed Content Management Platforms.

Panel Statement: Lloyd Rutledge

The main bottleneck in media semantics is the human effort required. There has been much progress in automated feature-based and structure-based indexing and retrieval. However, the media annotation that remains, that which can't be automated and thus requires human-creation, is often much more valuable. So how do we break this "human bottleneck"? One part of the solution is to minimize the effort required by people, making precious human-hours spent on annotation go further. Another part is to attract, and better train, more humans to perform this annotation.

Both optimizing and attracting human effort is helped by the development of standards for media annotation and the technologies for them. Standards, particularly those from the W3C, usually stimulate the development of tools for them that are readily obtainable, integrable and of low-cost. As communities using these standards and tools grow in size, so will the availability and understanding of techniques and methodologies for using them. As the application of these technologies gains more of a presence and market with the community at large, the motivation to become part of this market increases. The development of semantic standards and technologies is currently the focus of much effort in the W3C. There are several W3C working groups working on top, and a sister organization has been spun off from the W3C devoted to developing the Semantic Web.

So while the Semantic Web and related activities provide political and economic impetus for developing a human media semantic authoring community, several questions arise. One is how the nature and technical details of the emerging formats and technologies will impact this community, how it operates, and how it cooperates. Another is what choices can we make now in how these efforts proceed to optimize their impact and influence. And finally, as these efforts proceed as planned, what else is needed to attract people to media annotation and to make their work easier?

Biography

Dr. Lloyd Rutledge is a researcher at CWI, the Dutch national center for computer science and mathematics research. His research involves the Semantic Web, adaptive hypermedia and hypermedia standards. He received his Sc.D. from the University of Massachusetts, Lowell. Dr. Rutledge is a member of the W3C working group that developed SMIL. He is also a co-author of "SMIL: Interactive Multimedia on the Web", published in May by Pearson Education.

Panel Statement: Thomas Sikora

There is no doubt: Understanding semantics and meaning of media is urgently needed. Audio, speech, images and video floods the Internet, our TV systems, desktop computers, electronic handheld PCs and mobile phones. Finding the bits of interest becomes increasingly difficult. For many users finding media chunks has become a cumbersome, frustrating, and time consuming task. In the future this will develop into a significant cost factor for existing and emerging industries. What is needed to accommodate complex media search is an efficient (Internet) search engine geared towards media information: the "Media-Google" – a machine that helps finding media chunks based on semantics in media itself.

Semantics in sentences is sufficiently understood - but what is semantics in media that our Media-Google can extract and search for? Fact is: Media is so rich in its content variety that it will never sufficiently be described by text or words. Thus, searching for media described by words will never be sufficient - besides, who will actually take the time to annotate the media chunks? Most importantly, semantic meaning of media is application and user dependent. There are rarely two applications or users who would describe media content by the same semantics entities.

State-of-the-art extraction tools today analyze image, speech, sound and music - to extract basic low-level non-text-descriptions of the particular media chunks (i.e. colors and texture in images, rhythm and harmonics in music, pitch and spectral shape of general sound). Currently much research effort is geared towards bridging the semantic gap - to extract semantic content from these low-level descriptors. Promising new technologies for clustering music and image files into predefined categories already provide more efficient media-specific descriptions of content.

Even though results are promising these technologies will not be sufficient to solve real world problems. To account for the variety of what users actual look for, search engines need to adopt to users preferences. Search engines need to be developed that learn and understand about what they see or hear - based on users continuous feedback using the above low level descriptions. This is similar to how babies learn when they grow older. It is the users (parents) task to train the machine based on positive and negative feedback - in an iterative process until the preferences are sufficiently understood by the search engine. Call this the “man-machine-gap” or the “generation-gap”. Tackling this problem will be vital for developing the efficient Media-Google.

Biography

Prof. Thomas Sikora is the chair of the Communication Systems Department at Technical University Berlin, Germany. He received the Dipl.-Ing. degree and Dr.-Ing. degree in electrical engineering from Bremen University, Germany, in 1985 and 1989 respectively. In 1990 he joined Siemens Ltd. and Monash University, Melbourne, Australia, as a project leader responsible for video compression research activities in the Australian “Universal Broadband Video Codec” consortium. Between 1994 and 2001 he was the director of the “Interactive Media” Department at Heinrich-Hertz-Institute (HHI) Berlin GmbH, Germany. Dr. Sikora is co-founder and director of 2SK Media Technologies and Vis-a-Pix GmbH, two Berlin-based start-up companies involved in research and development of audio and video signal processing and compression technology.

Dr. Sikora has been involved in international ITU and ISO standardization activities as well as in several European research activities for a number of years. As the chairman of the ISO-MPEG video group (Moving Picture Experts Group), he was responsible for the development and standardization of the MPEG-4 and MPEG-7 image and video algorithms. He also served as the chairman of the European COST 211^{ter} video compression research group. He frequently works as an industry consultant on issues related to interactive digital audio and video. He is an appointed member of the Advisory and Supervisory board of a number of German companies and international research organizations.

Dr. Sikora is recipient of the 1996 German ITG award (German Society for Information Technology). He is appointed as an Associate Editor for a number of international journals including the IEEE Signal Processing Magazine and the EURASIP Signal Processing: Image Communication and EURASIP Signal Processing journals. He currently serves as the Editor-in-Chief of the IEEE Transactions on Circuits and Systems for Video Technology. He is a member of ITG and a senior member of IEEE.

Panel Statement: Herbert Zettl

Contextual Media Aesthetics as the Basis for Media Computing: The problem is to close the semantic gap when using computers for efficient and effective input and retrieval of specific media content. Whereas the computer is quite good at counting, storing, and retrieving concrete data, it seems less than adequate at identifying and interpreting meaning.

My contention is that this problem is less the fault of the machine and its programmers than the apparent apathy of media scholars to get involved in the study of the fundamental aesthetic elements of the moving image – light, space, time-motion, and sound – how they interrelate and what they communicate in a variety of contexts.

I believe that one of the more successful approaches to developing algorithms for detecting meaning might not be the construction of progressively finer grids for the analysis of existing films and video fare, but a careful study of how the basic aesthetic elements are applied in production. Specifically, this means that we should try to solve the problem not only by analytical techniques with which we seek to detect deeper meaning and semantic subtleties, but primarily by examining the process of synthesis: how and why various aesthetic elements are used to engender the desired perceptual effects.

Hence, I advocate a model that shows the hierarchical, yet interdependent relationship of (1) the five aesthetic media elements, light and color, two-dimensional space, three-dimensional space, time-motion, and sound; (2) their perceptual effects when structured within their specific aesthetic fields; (3) the aesthetic context (the agenda-setting framework in which we perceive the visual and aural cues), (4) the associative context (their semantic agenda-setting role), (5) cognitive mental maps (making sense of where things and people are and move to), and (6) affective mental maps (aesthetic cues that translate intuitively and almost instantly into feelings).

If such a model seems somewhat contrived, let me remind you that all good television and film directors apply it routinely in their daily work.

Biography

Dr. Zettl is a Professor Emeritus of Broadcast and Electronic Communication Arts, College of Creative Arts at San Francisco State University. Dr. Zettl’s research emphasis is media aesthetics and video production. He also acts as advisor to the Institute of International Media Communication. Prior to joining the San Francisco State University faculty, he worked at several professional television stations, including KPIX, the CBS affiliate in San Francisco, where he was a producer-director. He participated in numerous CBS and NBC network television productions, such as Edward R. Murrow’s Person to Person. While at KPIX, he won an Emmy Award of the San Francisco Chapter of the National Academy of Television Arts and Sciences (with two other people) for innovation in entertainment shows. He has been inducted into the prestigious Silver Circle of the National Academy of Television Arts and Sciences, Northern California Chapter, for outstanding contributions to the television profession.

He has presented numerous papers on media aesthetics and video production for a variety of academic and professional media conventions here and abroad, and has been asked to consult for several American and international universities and professional broadcasting institutions. He is one of the founders of the U.S. Annual Visual Communication Conference. Dr. Zettl was a visiting professor at Concordia University, Montreal, Canada, Heidelberg University, Heidelberg, Germany, and the Institute for Television and Film, Munich, Germany. He spearheaded various experimental television productions, such as dramas for simultaneous three-screen presentation, and various interactive multimedia programs.

He has written numerous articles, many of which were translated into foreign languages and/or published abroad. His books on television production and aesthetics, all published by Wadsworth Publishing Co., include: Television Production Handbook, 8th ed., 2003; Television Production Workbook, 8th ed., 2003; Sight Sound Motion: Applied Media Aesthetics, 3d ed., 1999. Video Basics 3, 3d ed., 2001; Video Basics Workbook 3, 3d ed., 2001. The TV Handbook, Sight Sound Motion, and Video Basics 3 were translated into several foreign languages, and are used in key television production centers and universities around the world. He also developed with the Cooperative Media Group, San Francisco, an interactive multimedia program Zettl's Video Lab 2.1, published by Wadsworth in 1995. Dr. Zettl retired from active classroom teaching in June 2000.

Organizer Biography

Dr. Chitra Dorai is a Research Staff Member at the IBM T.J. Watson Research Center, New York, where she leads the Media Semantics and e-Learning Media projects. She also serves as the IBM Research Relationship Manager for the media sector. Her research interests are in the areas of multimedia systems and digital video analysis, computer vision, pattern recognition and machine learning. Her current research focuses on developing technologies for digital media analysis in various domains such as education and training media and motion pictures, that are useful in content-based structuralization, annotation and search, and smart browsing. She received her Ph.D. from the Department of Computer Science at Michigan State University, where she was a recipient of the Distinguished Academic Achievement Award from the College of Engineering. Her work has received awards and recognition such as the Top-Ranked Paper at the 2002 Asian Conference on Computer Vision, the Best Paper Prize at the 2001 IEEE Pacific-Rim Conference on Multimedia, the Best Industry-related Paper Award at the 2000 International Conference on Pattern Recognition, and Honorable Mention in the 24th Annual Best Paper Award Contest of the Pattern Recognition Journal, 1997. She is a senior member of the IEEE and a member of the ACM.

Dr. Frank Nack is a senior researcher at CWI, currently working within the Multimedia and Human-Computer Interaction group. He obtained his Ph.D. in "The Application of Video Semantics and Theme Representation for Automated Film Editing", at Lancaster University, UK. The main thrust of his research is on video content representation, digital video production, multimedia systems that enhance human communication and creativity, interactive storytelling and media-networked oriented agent technology. He is a member of the MPEG-7 standardization group where he served as editor of the Context and Objectives Document and the Requirements Document, and chaired the MPEG-7 DDL development group. He is on the editorial board of IEEE Multimedia, where he edits the Media Impact column.

1. REFERENCES

[1] C. Dorai and S. Venkatesh. Computational Media Aesthetics: Finding meaning beautiful. *IEEE Multimedia*, 8(4):10–12, October-December 2001.

[2] ISO/IEC/JTC1/SC29/WG11. MPEG-7 Overview. http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm, July 2002.

[3] R. Jain. Teleexperience: Communicating compelling experiences. In *Proceedings of the ACM Multimedia 2001*, page 1, Ottawa, Ontario, 2001.

[4] F. Nack. The future of media computing. In C. Dorai and S. Venkatesh, editors, *Media Computing Computational Media Aesthetics*, chapter 8, pages 159–196. Kluwer Academic Publishers, 2002.

[5] W3C. Semantic Web, Web Ontology Working Draft. http://www.w3.org/2001/sw/, July 2002.