Multimedia Analysis + Visual Analytics = Multimedia Analytics

Nancy A. Chinchor ■ ChinchorEclectic
James J. Thomas and Pak Chung Wong ■ Pacific Northwest National Laboratory
Michael G. Christel ■ Carnegie Mellon University
William Ribarsky ■ University of North Carolina at Charlotte

Multimedia analysis has focused on images, video, and, to some extent, audio and has made progress in individual media types. It hasn’t focused on multiple media types or text analysis. Visual analytics has focused on user interaction with data during the analytic process plus the fundamental mathematics, and has continued to treat text as did its predecessor, information visualization. (Generally, we use “analytics” to mean the science of human analysis.)

This tutorial addresses combining multimedia analysis and visual analytics to deal with information from different sources, with different goals or objectives, and containing different media types and combinations of types. The resulting combination is multimedia analytics.

**Historical Perspective**
To begin, here’s a brief history of multimedia analysis and visual analytics, noting these distinct fields’ significant progress through the years.

**Multimedia Analysis**
Modern multimedia information retrieval (MIR) is rooted in computer vision, digital image processing, and pattern recognition, which started in the late 1970s to early 1980s. Since then, new technologies have continued to emerge in the multimedia R&D community.

In the 1980s, when digitized images weren’t an archival medium for the general public, research frequently covered edge finding, boundary and curve detection, region growing, shape identification, feature extraction, and so on, of individual images or frames of images.

In the 1990s, when digital video and images became part of our everyday experience, content-based image retrieval (CBIR) and content-based video clip retrieval were among the most important R&D accomplishments. Robust shot boundary detection and database information query were two of the most active research topics in academic and industrial research labs. The 1990s also saw the World Wide Web’s arrival. The Web brought large amounts of multimedia information directly to our desktop computers and further stimulated the rapid growth of the multimedia and entertainment industries. The first ACM Multimedia International Conference, which included MIR as a major topic, was in 1993.

The MIR community’s primary R&D goal in the 1990s was to develop computer-centric technologies for researchers’ use only. In contrast, the current primary goal is developing human-centric technologies that bridge the gap between general users and the technologies delivering multimedia information to them. We see more attempts to retrieve not just video or image information but
also audio information. However, we haven’t seen successful cases of multimedia information fusion in either the academic literature or patent applications. Overall, audio information retrieval hasn’t played a major role in MIR’s evolution. Even less important has been text information retrieval. Only a few studies have covered analysis of documents containing images and text or any other truly mixed-media forms.

The arrival of handheld mobile devices and the wide popularity of multimedia message services further encouraged industry to develop better indexing technologies to organize multimedia information and better browsing and summarization technology to access the information. The past decade also saw the first ACM International Conference on Multimedia Information Retrieval, in 2008. MIR has finally established its own identity and is no longer an R&D track of the multimedia community.

Visual Analytics
A relatively new suite of technologies has emerged from visual-analytics R&D. Visual analytics aims to provide technology for human-centric analysis through dynamic, active visual interfaces for all forms of data, to deal with scale-independent analytics. It employs fundamental mathematics to represent information and transform it into computable forms and employs knowledge sciences to represent multidimensional information.

Visual analytics involves developing a high-dimensional analytic space to enable detection of the expected and discovery of the unexpected during analytical thinking. Visual-analytics researchers envision a highly engaging intuitive visual interface based on cognitive principles that enables a thought process for analyzing multimedia information across multiple applications. This vision developed out of the natural growth of computer graphics and visualization.

Computer graphics started in the 1970s, focusing on animation, realization, and computer-aided design and engineering, primarily for the automotive and aircraft industries. There was also broad interest in developing and applying computer graphics technologies for scientific domains. A core publication setting an R&D agenda spurred interest in visualization’s potential for scientific computing. Consequently, many fundamental research programs in scientific visualization and the IEEE Visualization Conference were launched in the mid-1980s. Although nonscientific applications of visualization were also of interest, a clear focus on scientific domains emerged. This focus stimulated research funding for visualization in chemistry, biology, astronomy, atmospheric sciences, and many other fields, significantly increasing their capabilities.

In the early 1990s, a US government group asked several scientists in research centers to consider visualization of unstructured text documents. At the time, many researchers were visualizing biological sequences for drug discovery; however, developing visualizations for text analysis seemed difficult and had little mathematical foundation.

The 2001 terrorist attack on the US stimulated a relook at technology to reduce the risk of another attack through effective analysis of all forms of information.

The Need for Multimedia Analytics
Full multimedia analytics has been slow to develop, so we’re attempting to bring attention to the critical new suite of technologies required to
analyze images, text, video, geospatial data, audio, graphics, tables, and other forms of information. Multimedia analytics is a critical need for a broad range of applications, including, but not limited to, medicine, economics, social media, and security.

Surveys

A glimpse of some of the major peer-reviewed surveys on multimedia-analysis and visual-analytics topics lays a foundation for the field of multimedia analytics.

**Multimedia analytics is a critical need for a broad range of applications, including, but not limited to, medicine, economics, social media, and security.**

**Multimedia Analysis**

MIR includes topics from user interaction, data analytics, machine learning, feature extraction, information visualization, and more. The MIR community generally doesn’t classify text or document information as multimedia data. Also, image media represents most of the MIR community’s work. MIR researchers often study video media, but audio media shows up in only a handful of applications. A particularly difficult problem is dealing with multimedia information from different sources and with different goals or objectives.

Philippe Aigrain and his colleagues addressed image and video information retrieval. They covered traditional

- video analytics topics related to color, texture, shape, and spatial similarities;
- video-parsing topics such as temporal segmentation, object motion analysis, framing, and scene analysis; and
- video abstraction topics such as skimming, keyframe extraction, content-based retrieval of clips, indexing, and annotation.

Yong Rui and his colleagues described a complete taxonomy of early, classic image information retrieval techniques. They mainly covered

- feature extraction techniques related to color, texture, shape, color layout, and segmentation;
- image-indexing techniques such as dimensional reduction and multidimensional indexing; and
- image retrieval systems developed in the 1990s.

Many of the covered techniques have become the foundational technology for multimedia systems and applications today.

Arnold Smeulders and his colleagues focused on image processing, pattern analysis, and machine learning. They started by discussing basic components such as color and texture. They then visited the more advanced topic of features, which can be extracted from an image and form a hierarchy of global features, salient features, signs, shapes, and object features. They also covered machine-learning topics of similarity matching and semantic interpretation as well as database topics such as image indexing, storage, and query.

Cees Snoek and his colleagues described video indexing as a hierarchy that groups different index types. This hierarchy characterizes different genres (such as news, sports, movies, or commercials) and subgenres (such as basketball and ice hockey) in terms of their most prominent layout and contents. They split the hierarchy into named events (such as football games and tennis matches) and logical units (such as car chases and violence).

Michael Lew and his colleagues covered image, video, and audio information retrieval from data gathered from different sources and stored in an archive. They focused more on human-centric topics that bridge the semantic gap between users and their multimedia information and less on traditional computation-centric topics such as similarity search. They attempted to bridge the semantic gap by “translating the easily computable low level content-based media features to high level concepts or terms which would be intuitive to the users.”

Although they discussed audio-related problems, most problems they discussed were related to video and image information retrieval. They didn’t discuss multimedia fusion or retrieval of combined multimedia concepts.

Ritendra Datta and his colleagues addressed CBIR. Their survey had a strong data-mining flavor and covered all aspects of knowledge discovery of image databases. Besides technical topics such as signature extraction, clustering, categorization, visualization, and similarity matching, they discussed nontechnical issues such as aesthetics, security, the Web, and storytelling.

In other notable surveys, Rainer Lienhart focused on shot boundary detection in video, Ming-Hsuan Yang and his colleagues focused on face detection, and Johan Tangelder and Remco Veltkamp focused on content-based 3D shape retrieval.

Two very recent papers excellently summed up the state of the art in multimedia analysis for beginning readers. Snoek and Smeulders tried to answer
and Kristen Grauman reviewed new algorithms providing robust but scalable image search.\textsuperscript{16} Finally, the common belief is that there are no solved problems in the MIR community, which includes the more traditional image and video retrieval community. As Lew and his colleagues put it, "In some cases a general problem is reduced to a smaller niche problem where high accuracy and precision can be quantitatively demonstrated, but the general problem remains largely unsolved."\textsuperscript{10}

**Visual Analytics**

The major publications surveying visual analytics have been produced by large groups of researchers. As we mentioned earlier, the first R&D agenda appeared in 2005.\textsuperscript{5} In December 2009, a special issue of the *Journal of Information Visualization* looked to visual analytics’ past and future. For an overview, we recommend the guest editors’ introduction in that issue.\textsuperscript{17} That introduction also discussed five success stories demonstrating the early technologies’ value.

**Multimedia Analytics’ Beginnings**

Human communication is multimodal. Linguistic studies of sign language in the late 1970s and early 1980s concluded that visual language is just as rule-based and creative as spoken language. However, researchers also noticed that visual cues accompanying spoken language contain significant syntactic and phonological information. Even in our electronically connected world, we find it more satisfying to communicate in person. A telephone, email, or even simply a curtain or darkness between us makes reading the other person difficult.

Analysis of multimedia is no different. Media use multiple modalities to communicate. For example, a video rarely has no sound track, and most reports have some graphics, whether they’re tables or images. Multiple media types collected for many different purposes are regularly presented to us digitally for interpretation. For us to accomplish our analysis quickly, the computer must be able to access these records to suit our immediate needs, utilizing the many rich connections among the media types that humans have placed there for communicative purposes. In a document, figures, images, and even video clips can enhance the text, and the ways in which these types of media combine to form the overall message aren’t well understood. Visual analytics has been a boon to dealing with large data collections. Multimedia analysis has made significant strides in analyzing each type of media. Computational linguistics also has come a long way in extracting meaning from large collections of text and speech. These fields have much to offer each other.

Carnegie Mellon University’s Informedia project (www.informedia.cs.cmu.edu) provides an example of the synergy among these fields. The project’s products result from scientific studies of how humans analyze multimedia; thus, they illustrate multimedia analytics.

Informedia researchers noticed that extracting evidence and support materials from large video repositories can be extremely tedious, owing to the linear time-dependent nature of audio and video recordings, especially those stored on tape. To facilitate better navigation into and across these recordings, Informedia employed speech recognition, image processing, and language technologies to derive synchronized metadata and indexing. Benchmarking forums such as the US National Institute of Standards and Technology Text Retrieval Conference Video Retrieval Evaluation (NIST TRECVID) track have charted Informedia’s progress over the years.

Informedia research has also focused on interfaces that leverage metadata to deliver efficient, effective retrieval from multimedia corpora.
In the following examples, two demonstration corpora illustrate how Carnegie Mellon University’s Informedia project (www.informedia.cs.cmu.edu) has employed speech, image, and language processing to improve navigation in a video corpus. The HistoryMakers African-American oral history digital archive (www.idvl.org/thehistorymakers) is 913 hours of 18,254 stories, with one shot per story. The US National Institute of Standards and Technology Text Retrieval Conference Video Retrieval Evaluation (NIST TRECVID) 2006 broadcast news test set is 165 hours of US, Arabic, and Chinese news sources comprising 5,923 stories, with 146,328 shots.

Further field and evaluation work on these corpora appears elsewhere, along with a complete set of references. These examples aim to show the potential of multiple coordinated views of a multimedia space in the hands of an intelligent human operator for video exploitation.

The HistoryMakers Archive

HistoryMakers users are interested in story access, with much information contained in the audio. Consider a user interested in music—specifically, gospel, jazz, rap, and rock. Querying this dataset returns 2,372 stories that match one or more of the terms.

Figure A shows a scatterplot of points across a time line, and a visual information browsing environment (VIBE) plot. The VIBE plot, first developed through the University of Pittsburgh Library and Information Science Program, lets users see and manipulate query terms’ contributions to the result space. The initial view shows that “music” is the dominant term, appearing in 2,152 of the results. Unchecking the “music” query term shows that 823 stories remain, discussing one or more of gospel, jazz, rap, and rock. Drawing a bounding box in the VIBE plot, the user can drill down to the 23 stories discussing two or more of these terms.

Further interface elements include color-coded match terms, a relevance score bar in the segment grid view with each result, and information brushing across views. In Figure A, the user’s curser hovers over Isaac Hayes in the segment grid view, showing a tooltip and coloring yellow the points related to this story in the VIBE and time line views. A yellow point in the VIBE plot indicates that this story discusses rock, rap, and music. A yellow point in the time line shows that this story discusses the 1970s.

The views offer quick, comparative lenses on the data, emphasizing geography, time, and named-entity relationships, among other things. For example, drilling into the 141 rap stories from the 1940s through the 1980s shows that most 1940s references drop out: 34 references show for rap, with only one in the 1940s. Drilling into the 487 jazz stories and opening the named-entities view produces the results in Figure B.

Through a sequence of interactions, the user provides context with which to build more carefully tuned interfaces to serve his or her needs. Perhaps the user was querying only people, in which case he or she could open the named-entity view to emphasize solely people connections rather than people, places, and organizations. The user can select a node such as “Billie Holiday” and ask to see a video skim, or highlight reel, of all the clips tying Billie Holiday to jazz (in this case, three clips). Work with TRECVID video summarization tasks has proven the difficulty of generating video skims in the general case.
users mark mistakes so that the system can learn from them and apply that learning when building future classifiers.

In addition, Informedia researchers have worked on approaches that can deliver improved recognition but are computationally expensive. For example, MoSIFT (Motion Scale-Invariant Feature Transform) recognizes activities in surveillance videos by exploiting continuous object motion explicitly calculated from optical flow, integrated with distinctive appearance features. Such video retrieval approaches’ value is assessed by international benchmarking forums, such as TRECVID, that chart progress on video analytics tasks.

**Reference**


**Benchmark Datasets and Evaluation**

At the IEEE VisWeek 2009 Workshop on Video Analytics, we noticed a palpable excitement about the future combination of multimedia analysis and visual analytics to support digital-data analysis.
However, there was a resounding call for datasets. So, here we describe benchmark datasets and evaluation in the two fields.

**Multimedia Analysis**

TRECVID, under the coordination of Alan Smeaton, Wessel Kraaij, and Paul Over, has charted the progress of various video retrieval tasks, including shot detection, semantic indexing, and fully automatic and interactive retrieval.

Shot detection decomposes a video narrative into component shots. This decomposition enables higher-level processing to classify shots with attributes and to allow construction of a visual table of contents through thumbnail representations of shots (shot thumbnails). A storyboard of shot thumbnails serves a purpose similar to ASR indexing of audio: a means to survey and navigate a linear video presentation. By 2006, most participating systems performed shot detection at greater than 90 percent accuracy, with TRECVID retiring the task in that year to focus on more challenging issues.

Semantic indexing, which automatically assigns semantic tags to video sequences (such as shots), can be fundamental technology for filtering, categorization, browsing, search, and other video exploitation. TRECVID experiments have shown that some visual concepts, such as “face” or “text,” can be automatically tagged to video with excellent accuracy, whereas others, such as “bridge,” “bus,” or “flower,” remain challenging.

In light of automatic classification’s varying accuracy for different semantic tags, Informedia has developed interfaces that let users control whether to require greater precision (seeing fewer candidates with anticipated higher accuracy) or greater recall (seeing more candidates, to avoid missing anything of relevance) for a given task and tag. Allowing the user interactive control over storyboard interfaces and over which tags to apply as filters and what degree of precision or recall to use has consistently improved performance of video retrieval tasks. For example, TRECVID experiments confirmed the value of a person in the loop for shot-based retrieval. In these experiments, a human using an interactive search system significantly outperformed fully automatic search.

On the other hand, for TRECVID 2009, for three of the 24 search topics, the best automated system outperformed the best interactive system. As some visual-indexing schemes mature, such as face detection, automated tasks that can best exploit them will also improve dramatically. Examples of such tasks are

- finding crowds of people and
- people at desks.

These are two of the three topics on which the automated systems did so well.

**Visual Analytics**

When the field of visual analytics emerged in 2005, a new evaluation effort started. A big issue was the lack of data. So, NVAC started a project to create synthetic datasets very close to real data without issues of classification or personally identifiable information. This project now creates the datasets for the annual IEEE Visual Analytics Science and Technology Challenge (http://hcil.cs.umd.edu/localphp/hcil/vast10/index.php). The project was originally expected to last about 10 years, starting with relatively easy-to-analyze data and then employing increasingly complex data. Today, these datasets are publicly available and being used in education, industry, and government-funded research. Each dataset is a “ground truth” scenario that approximates real situations but is completely open for analysis.

In addition, NVAC established a program to go beyond usability evaluation to utility evaluation. This is a major change. Visual-analytics researchers want to be able to evaluate technologies to show the effectiveness of not only the interface but also the analytic improvement that was the interface’s goal. The overall goal is to develop evaluation processes that enable researchers to scientifically prove the new technologies’ increased analytic value.

Recent progress in visual analytics has been manifested in presentations at large meetings of US government analysts. Visual analytics has been so successful that analysts can now look for more data rather than bemoan the problem of too much data. The established wisdom is now that visualization for data discovery differs significantly from visualization for illustration of evidence.

T

he VAST challenges so far have resulted in researchers seeing their technology more and more as analysts see it. As these technologies become more successful, they’ll incorporate seamless collection of additional data during analysis. Analysts are valued by their reporting. Their reports’ quality is enhanced by their handling of larger quantities of data and their ability to discover the unexpected. However, researchers have done little work to support the generation of illustrations explaining large quantities of data.20

For a list of publications and societies related to multimedia analytics, see the related sidebar.
Acknowledgments
The US National Science Foundation has supported the Informedia research reported here under grant IIS-0705491. The US National Visualization and Analytics Center at the Pacific Northwest National Laboratory has also supported some of the described research. The Battelle Memorial Institute manages the laboratory for the US Department of Energy under contract DE-AC06-76RL01830.

References

Published and Societies Related to Multimedia Analytics

■ ACM Special Interest Group on Multimedia, www.sigmm.org
■ IEEE Communications Society, www.comsoc.org
■ IEEE Technical Committee on Visualization and Graphics, www.cc.gatech.edu/gvu/tccg
Nancy A. Chinchor is founder of ChinchorEclectic. Her research interests are analytics of visual media and multimedia. Chinchor has a PhD in linguistics from Brown University. Contact her at chinchoreclectic@gmail.com.

James J. Thomas is an American Association for the Advancement of Science Fellow and Laboratory Fellow at the Pacific Northwest National Laboratory. He’s founder and past director of the Department of Homeland Security National Visualization and Analytics Center. His research interests include establishing investment directions for visual analytics and information and computing technology, leading major technology initiatives, mentoring staff, and serving as a principal investigator on several major science programs. Thomas has a master’s in computer science from Washington State University. Contact him at jim.thomas@pnl.gov.

Pak Chung Wong is chief scientist at the Pacific Northwest National Laboratory. His research interests are visualization, bioinformatics, data signature, computational science, steganography, and wavelets. Wong has a PhD in computer science from the University of New Hampshire. Contact him at pak.wong@pnl.gov.

Michael G. Christel is a research professor at Carnegie Mellon University’s Entertainment Technology Center. His research interests are edutainment, digital libraries, human-computer interaction, and multimedia analytics. Christel has a PhD in computer science from Georgia Tech. Contact him at christel@cmu.edu.

William Ribarsky is the Bank of America Endowed Chair in Information Technology and the chair of the Computer Science Department at the University of North Carolina at Charlotte. His research interests are visual analytics, 3D multimodal interaction, bioinformatics visualization, virtual environments, visual reasoning, and interactive visualization of large-scale information spaces. Ribarsky has a PhD in physics from the University of Cincinnati. He’s on the IEEE Computer Graphics and Applications editorial board. Contact him at ribarsky@uncc.edu.