Multimedia Analysis + Visual Analytics = Multimedia Analytics

Nancy A. Chinchor[1], James J. Thomas[2], Pak Chung Wong[3], Michael G. Christel[4], and William Ribarsky[5]

Introduction

Multimedia analysis has focused on images, video, and to some extent audio and has made progress in single channels. Multimedia analysis has not focused on text analysis. Visual analytics has focused on the user interaction with data during the analytic process plus the fundamental mathematics, and has continued to treat text as did its precursor, information visualization. Generally, we use the term "analytics" to mean the science of human analysis. The general problem we address in this tutorial is the combining of multimedia analysis and visual analytics to deal with multimedia information gathered from different sources, with different goals or objectives, and containing all media types and combinations in common usage.

To begin the tutorial, we provide the history of the two distinct fields – multimedia analysis and visual analytics – noting their significant progress through the years. We then provide a survey in each of these fields. We discuss the importance of the combination of multimedia analysis and visual analytics for dealing with the digital data that is commonly available now.

Next, we introduce combining multimedia analysis and visual analytics into the field of *multimedia analytics*. We describe an example of the potential for this field as indicated by the Informedia project at Carnegie Mellon University. The researchers there noticed that extracting evidence and support materials from large video repositories can be extremely tedious, due to the linear time-dependent nature of audio and video recordings, especially during their linear analog tape legacy. As these recordings became digital, access into the materials improved through synchronized metadata and indexing. The researchers also combined information visualization, library science, and speech recognition with their video analysis to produce effective user interfaces for analysis of mixed video and audio media. Their designs were the result of scientific studies of the ways in which humans analyze multimedia, and, thus, illustrate multimedia analytics.

At the IEEE VisWeek 2009 Workshop on Video Analytics, we noticed a palpable excitement about the future of these two rapidly developing fields being brought together to support analysis of all digital data commonly available. At the workshop, there was a resounding call for data sets; therefore, we have included a section in this tutorial on benchmark data sets and evaluation

[1] ChinchorEclectic LLC
[2] Pacific Northwest National Laboratory
[3] Pacific Northwest National Laboratory
[4] Carnegie Mellon University
[5] University of North Carolina at Charlotte

in both multimedia analysis and visual analytics. We conclude this tutorial with lists of journals, professional societies, and references that will provide additional information for the reader.

Historical Perspective

Multimedia Analysis

Modern multimedia information retrieval (MIR) study is rooted in traditional areas of computer vision, digital image processing, and pattern recognition studies, which started in the late 1970s to early 1980s. Throughout the next 30 years, new technologies have continued to emerge in the multimedia research and development (R&D) community.

In the 1980s, when digitized image was not an archival medium for the general public, the study was frequently about edge finding, boundary and curve detection, region growing, shape identification, feature extraction, etc. of individual images or frames of images.

In the 1990s, when both digital videos and images became part of our everyday experience, content-based image retrieval (CBIR) and content-based video clip retrieval (CBVR) were among the most important R&D accomplishments of the decade. Robust shot boundary detection and database information query were two of the most active research topics within academic and industrial research labs. The 1990s were also characterized by the arrival of the World Wide Web (WWW), which 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 (MM) international conference, which included MIR as a major conference topic, was held in 1993. The primary R&D goal of the MIR community in the 1990s was to develop computer-centric technologies for researchers' use only.

In contrast, the primary goal of the 2000s is about developing human-centric technologies that bridge the gap between general users and the technologies that deliver the multimedia information to the users. We do see more attempts to retrieve not just video or image information but also audio information. However, we have not seen successful cases of multimedia information fusion in either academic literature or patent applications. Overall, audio information retrieval does not play a major role in the entire evolution of multimedia information retrieval study. Of even less importance is text information retrieval study. There are only a few studies of the 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 technology to organize the multimedia information and develop better browsing and summarization technology to access the information. The decade of 2000s also marks the birth of the first ACM International Conference on Multimedia Information Retrieval in 2008. The area of MIR has finally established its own identity and is no longer an R&D track area of the more traditional multimedia community.

Finally, it is a common belief that there are no solved problems within the MIR community, which includes the more traditional image and video retrieval community. "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." [Lew et al. 2006]

Visual Analytics

A relatively new suite of technologies has emerged from visual analytics R&D. Visual analytics aims to provide analytic reasoning technology facilitated by human interactions through dynamic and active visual interfaces for all forms of media to deal with scale-independent analytics approaches. Visual analytics has a deep history in computer graphics and visualization and grew out of information visualization. Visual analytics brings in fundamental mathematics for represention and transformation of information into computable forms. It also brings in knowledge sciences to represent multidimensional information. In visual analytics, a high-dimensional analytic space is developed to enable the *detection of the expected and discovery of the unexpected* during analytical thinking. Visual analytics researchers envision a highly engaging intuitive visual interface that is based on cognitive principles and 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 with a focus on animation, realization, and computer-aided design and engineering, primarily for the automotive and aircraft industries. There was also a broad interest in developing and applying computer graphics technologies for scientific domains. A core publication setting an R&D agenda spurred interest in the potential impact for scientific computer graphics [McCormick et al. 1987]. As a result, many fundamental research programs in scientific visualization and the IEEE Visualization forum were launched in the mid-1980s. While non-scientific applications of visualization were also of interest, a clear focus on scientific domains emerged, stimulating research funding for visualization in chemistry, biology, astronomy, atmospheric sciences, and many other fields, significantly increasing their capabilities.

In the early 1990s, a group from the U.S. government 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 very difficult and had little mathematical foundation. In tackling text visualization, researchers focused on visualizing 200-2000 documents in a relatively simple format. This was accomplished as a prototype in early 1994 and highlighted on the cover of the proceedings for the first IEEE Symposium on Information Visualization [Gershon and Eick, 1995]. This field of study rapidly grew as many saw the opportunities in information visualization in the late 1990s. The Spatial Paradigm for Information Retrieval and Exploration (SPIRE) technology [Wise et al. 1995] formed the basis of much of this work.

The 2001 terrorist attack on the United States stimulated a re-look at technology to reduce the risk of another attack through effective analysis of all forms and types of information. Also, analysts were swamped with an ever-increasing amount and complexity of information. This situation stimulated the U.S. Department of Homeland Security to establish the National Visualization and Analytics Center (NVAC) at the Pacific Northwest National Laboratory (PNNL) in 2004 to consider a new approach for visualization. The PNNL-developed IN-

SPIRE™ technologies [Hetzler and Turner, 2004] were proof that new visualization approaches were possible. In 2005, a team of ~40 individuals from industry, academia, government, and national laboratories developed a research agenda for *visual analytics*, published in the book *Illuminating the Path: The Research and Development Agenda for Visual Analytics* [Thomas and Cook eds., 2005]. This R&D agenda was the foundation for visual analytics and included new thinking on multi-modal visual analytics.

Full multimedia analytics has been slow to develop to this point, so we are attempting to bring attention to the critical new suite of technologies required to analyze image, text, video, geospatial, 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 medical, economic, social media, and security applications.

Surveys

Multimedia Analysis

Information retrieval in multimedia includes topics from user interaction, data analytics, machine learning, feature extraction, information visualization, and more. The Multimedia Information Retrieval (MIR) community, in general, does not classify text or document information as a kind of multimedia data. Additionally, image media represents the majority of the work of the MIR community. Video media is often studied, but audio media only shows up in a handful of applications. A more difficult problem is dealing with multimedia information gathered from different sources and with different goals or objectives. The following survey of surveys presents short descriptions of peer-reviewed survey papers on various multimedia topics:

Aigrain et al. [1996] address image and video information retrieval. The paper covers traditional 1) video analytics topics from color, texture, shape, spatial similarities; 2) video parsing topics such as temporal segmentation, object motion analysis, framing, and scene analysis; and 3) video abstraction topics such as skimming, key-frame extraction, content-based retrieval of clips, indexing, and annotation.

Rui et al. [1999] describe a rather complete taxonomy of classic image information retrieval techniques developed in the early years. The discussion mainly covers 1) feature extraction techniques such as color, texture, shape, color-layout, and segmentation; 2) image indexing techniques such as dimensional reduction and multidimensional indexing; and 3) image retrieval systems developed in the 1990s. Many techniques covered by this survey paper have become the foundational technology for many multimedia systems and applications today.

Smeulders et al. [2000] have a strong focus on image processing, pattern analysis, and machine learning. The paper starts with a discussion on basic components such as color and texture. It then visits 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. The paper also covers machine learning topics of similarity matching and semantic interpretation as well as database topics such as image indexing, storage, and query.

Snoek et al. [2005] describe the video indexing process as a hierarchy that groups different index types, characterize different genres (news, sports, movies, commercials) and sub-genres (such as basketball and ice hockey under sports) in terms of their most prominent layout and contents, and splits the hierarchy structure into named events (such as NFL football and tennis games) and logical units (such as car chase and violence).

Lew et al. [2006] claim to cover image, video, and audio information retrieval from data gathered from different sources and stored in a data archive. The paper focuses more on human-centric topics that bridge the semantic gap between the users and their multimedia information and less on traditional computation-centric topics such as similarity search. The authors attempt 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 audio-related problems are included in the discussion, the majority are video- and image-related information retrieval problems. There is no discussion on multimedia fusion or retrieval of combined multimedia concepts in the survey paper.

Datta et al. [2008] address content-based information Retrieval (CBIR). The paper has a strong data mining flavor that covers all aspects of knowledge discovery of image databases. In addition to technical topics such as signature extraction, clustering, categorization, visualization, and similarity matching, the paper discusses non-technical issues such as aesthetics, security, web, and storytelling.

Other peer-reviewed survey papers of note include Lienhart [2001] focused on shot boundary detection in video, Yang et al. [2002] focused on face detection, and Tangelder and Veltkamp [2004] focused on content-based 3D shape retrieval.

Two very recent papers are excellent for summing up the state of the art in Multimedia Analysis for the beginning reader. Snoek and Smeulders [2010] try to answer the question "Visual-Concept Search Solved?" in their paper in *IEEE Computer*. In *ACM Communications*, Grauman [2010] reviews new algorithms that provide the ability for robust but scalable image search in her article "Efficiently Searching for Similar Images."

Visual Analytics

The major publications that survey the field of visual analytics have been produced by large groups of researchers. The first Research and Development Agenda appeared in [Thomas and Cook eds., 2005]. Five years later, a special issue of the *Journal of Information Visualization* looked to the past and the future of visual analytics. For an overview, we recommend [Kielman et al. 2009]. In addition, the first article of that special issue discussed five success stories demonstrating the value of the early new technologies.

Combining Multimedia Analysis and Visual Analytics: The Beginnings of Multimedia Analytics

Human communication is known to be multi-modal. 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 in spoken language contain significant syntactic and phonological information. Even in our electronically connected world, we find it more satisfying to communicate in person. The telephone, email, or even simply a curtain or darkness between us makes it difficult to read the other person.

Analysis of multimedia is no different. Multiple media types collected for many different purposes are regularly presented to us digitally for interpretation. To accomplish our analysis quickly, we need the computer to be able to access these records to suit our immediate needs, utilizing the many rich connections among the media types that humans placed there for communicative purposes. In a document, there can be figures, images, and even video clips that enhance the text and the ways in which these types of media are combined to form the overall message are not well understood. Visual analytics has been a boon to dealing with large collections of data. 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. We know that these fields have much to offer each other. We describe the Informedia project as an example of the synergy among them.

The Informedia research group at Carnegie Mellon University established the use of speech recognition, image processing, and language technologies to derive synchronized metadata in order to facilitate better navigation into and across video collections. Benchmarking forums such as the National Institute of Standards and Technology (NIST) Text Retrieval Conference (TREC) video track (TRECVID) evaluation forum charted progress through the years.

Informedia work focused not only on automated metadata production and organization but also on the interfaces leveraging from such metadata in order to deliver efficient, effective retrieval from multimedia corpora [Christel 2009]. Rather than predetermine how the user should view a large story collection, the Informedia interface provides a multiplicity of views, which are covered more thoroughly in [Christel 2009]. These views draw from work done in the fields of information visualization and library science.

For materials with an audio narrative track, such as documentaries or broadcast news, speech recognition can provide a searchable transcript for navigation. The accuracy of transcripts produced by automatic speech recognition (ASR) for video retrieval was first investigated by Informedia researchers Alex Hauptmann and Michael Witbrock, showing that the information retrieval effectiveness can be adequate despite the transcription mistakes by ASR.

The issue was investigated further by the NIST TREC Spoken Document Retrieval (SDR) track from 1997 to 2000, which ended in 2000 with the conclusion that retrieval of excerpts from broadcast news using ASR for transcription permits relatively effective information retrieval, even with word error rates of 30%. In cases where accurate transcripts are available, automated speech systems can still be used to add value. ASR engines can provide tight word-time alignment, which supports the pinpoint navigation of elements of interest within a broader audio or video.

Much of Informedia research has emphasized using the multiple modalities of text, image, and speech to compensate for deficiencies in fully automated processing. Interface work brings in the user as well to overcome problems in automated processing, such as a textured set of trees being misrecognized as a crowd or the term “Sax” being misrecognized as a person in named entity tagging. Through such work, and benchmarking activities like TRECVID, the user will undoubtedly have more tools at hand with which to deal with greater volumes of information. Active learning has been used by which the user can mark mistakes so that the system can learn from them in building future classifiers.

Another direction has been to use computationally expensive approaches that can deliver improved recognition at the expense of requiring more processing time. For example, MoSIFT is a technique to better recognize activities in surveillance video by exploiting continuous object motion explicitly calculated from optical flow, integrated with distinctive appearance features [Chen 2010]. The value of these techniques for video retrieval are assessed through continued participation in international benchmarking forums, such as NIST TRECVID which chart progress on tasks related to video analytics.

Benchmark Data Sets and Evaluation

Multimedia Analysis

NIST TRECVID, under the coordination of Alan Smeaton, Wessel Kraaij, and Paul Over, has charted the progress of a number of video retrieval tasks through the years, including shot detection, semantic indexing, and fully automatic and interactive retrieval performance. Shot detection decomposes a video narrative into component shots allowing for higher-level processing to classify shots with attributes and to allow a visual table of contents into the video to be constructed through thumbnail image representations for shots. A storyboard of shot thumbnails serves a similar purpose as ASR indexing of audio: a means to survey and navigate into a linear video presentation. Shot detection was performed at greater than 90% accuracy by most participating systems by 2006, with the task retired in that year by TRECVID to focus on more challenging issues.

Semantic indexing, the automatic assignment of semantic tags to video sequences (such as shots), can be fundamental technology for filtering, categorization, browsing, search, and other video exploitation. Through the years, results from NIST TRECVID experiments have shown that some visual concepts can be automatically tagged to video with excellent accuracy, such as the presence of faces or text, while others, such as bridge, bus, or flower, remain challenging. In light of the varying accuracy of automatic classification for different semantic tags, the Informedia group developed interfaces allowing for the user to drive whether greater precision (seeing fewer candidates with anticipated higher accuracy) or greater recall (seeing more candidates in order not to miss anything of relevance) should be exercised for a given task and tag. Allowing the user interactive control over storyboard interfaces into shots and which tags to apply as filters and to what degree has consistently led to better performance on video retrieval tasks through the years. That is, NIST TRECVID experiments have confirmed the value of a person in the loop for shot-based retrieval tasks, where a human operator as part of the visual analytics operation significantly outperforms fully automatic searching. Interestingly for

TRECVID 2009, there were 3 topics of 24 where the best automated search system outperformed the best interactive search submission. As some visual indexing schemes mature, such as face detection, those tasks that can best leverage from such schemes will also show dramatic improvements (like finding crowds of people or people at desks, two of the three topics that did so well for the automated search). NIST TRECVID documents the progress for important sets of activities that aid visual analytics.

Visual Analytics

When the field of visual analytics was formed in 2005, a new effort in evaluation started. The common issue heard by researchers in visualization was lack of data. Therefore, a new project was started at NVAC to create synthetic data sets very close to real data without issues of classification or personally identifiable information. This project now creates the data sets for the annual IEEE Visual Analytics Science and Technology (VAST) Challenges (http://hcil.cs.umd.edu/localphp/hcil/vast10/index.php). The synthetic data set project was originally expected to last about 10 years, starting with relatively easy to analyze data that would become more complex data over the years. Today, these data sets are publicly available and are being used in classes, industry, and government-funded research programs. Each data set is a "ground truth" scenario that approximates real situations but is completely open for analysis.

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

Recent progress in the field of visual analytics has been manifested in presentations given in large meetings of U.S. government analysts. It has become established wisdom that visualization for data discovery and visualization for illustration of evidence have significantly different characteristics. Visual analytics has achieved so much success that analysts may now look for more data rather than bemoaning the problem of too much data.

Journals and Magazines

ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCACAP) – **http://tomccap.acm.org/**
IEEE Computer Graphics and Applications
IEEE Multimedia – **http://www.computer.org/portal/web/multimedia/home**
IEEE Transactions on Multimedia – **http://www.ieee.org/organizations/society/tmm/**
IEEE Transactions on Visualization and Computer Graphics
IEEE VAST Symposium and Conference Proceedings. 2007, 2008, 2009.
Journal of Information Visualization (Palgrave)

Professional Societies

ACM Special Interest Group on Multimedia (SIGMM) – http://www.sigmm.org/

IEEE Technical Committee on Multimedia Communications (MMC), IEEE Communications Society – http://committees.comsoc.org/mmc/
IEEE Technical Committee on Visualization and Graphics

References

Aigrain, P, Zhang, H., and Petkovic, D. 1996. Content-Based Representation and Retrieval of Visual Media: A State-Of-The-Art Review. Multimed. Tools Appl. 3, 3, 179–202.
Cees G.M. Snoek and Arnold W.M. Smeulders 2010 Visual-Concept Search Solved? In IEEE Computer.

Chen, M., Mummert, L., Pillai, P., Hauptmann, A., and Sukthankar, R. 2010. Exploiting Multi-Level Parallelism for Low-Latency Activity Recognition in Streaming Video. In Proceedings of the First Annual ACM SIGMM Conference on Multimedia Systems (Phoenix, Arizona, USA, February 22 - 23, 2010). MMSys '10. ACM, New York, NY, 1-12. DOI= http://doi.acm.org/10.1145/1730836.1730838

Christel, M.G. Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation. San Rafael, CA: Morgan and Claypool Publishers, 2009. DOI: 10.2200/S00167ED1V01Y200812ICR002

Datta, R., Joshi, D., Li, J., and Wang J. Z. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, 40, 2, Article 5.

Gershon N., and Eick S., Proceeding of IEEE ’95 Information Visualization, IEEE Computer Society Press, Oct 1995.

Hetzler E. and Turner, A. 2004. Analysis Experiences Using Information Visualization. IEEE Computer Graphics and Applications, 24(5):22-26.

Kielman, J., Thomas, J. and May, R. 2009 Foundations and Frontiers in Visual Analytics. In Journal of Information Visualization Special Issue: Foundations and Frontiers of Visual Analytics vol. 8 No. 4

Kristen Grauman 2010 Efficiently Searching for Similar Images. In Communications of the ACM, June 2010, vol. 53, no. 6. DOI= http//doi.acm.org/10.1145/1743546.1743570

Lew, M. S., Sebe, N., Djeraba C., and Jain, R. 2006. Content-based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, Communications, and Applications. 2, 1, 1-19.

Lienhart, R. 2001. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. International Journal of Image and Graphics 1(3), 469-486.

McCormick B., Defaunti T., Brown M., Visualization in Scientific Computing, ACM SIGGRAPH, Volume 21, Number 8, 1987

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349-1380.

Snoek, C.G.M., Worring, M., Van Gemert, J., Geusebroek, J.M., Koelma, D., Nguyen, G.P., De Rooij, O., and Seinstra, F. 2005. MediaMill: Exploring News Video Archives Based on Learned Semantics. In Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November, 225-226.

Tangelder, J. and Veltkamp, R.C. 2004. A Survey of Content Based 3d Shape Retrieval Methods, In Proceedings of International Conference on Shape Modeling and Applications, Genova, Italy, June 2004, IEEE, New York, 157-166.

Thomas, J.J. and Cook, K.A., eds. 2005 Illuminating the Path: The Research and Development Agenda for Visual Analytics, Los Alamitos, CA: IEEE Computer Society Press.

Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. 1995. Visualizing the Nonvisual: Spatial Analysis and Interaction with Information from Text Documents. Proceedings of the 1995 IEEE Symposium on Information Visualization, IEEE CS Press, Los Alamitos, Calif., pp. 51-58.

Yang, M.H., Kriegman, D.J., and Ahuja. N. 2002. Detecting Faces in Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34-58.

Yong Rui, Thomas S. Huang, and Shih-Fu Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues," Journal of Visual Communication and Image Representation 10, 39-62 (1999).

Acknowledgments

The Informedia work reported here is supported in part by the National Science Foundation under Grant No. IIS-0705491. Some of the work described has been supported by the National Visualization and Analytics Center™ (NVAC™) located at the Pacific Northwest National Laboratory in Richland, WA. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RL01830.