

# Multimedia Information Retrieval: Promises and Challenges

Nicu Sebe  
Faculty of Science,  
University of Amsterdam  
nicu@science.uva.nl

The explosion of multimedia content in databases, broadcasts, streaming media, etc. has generated new requirements for more effective access to these global information repositories. Content extraction, indexing, and retrieval of multimedia data continues to be one of the most challenging and fastest-growing research areas. A consequence of the growing consumer demand for multimedia information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, we need robust techniques to index/retrieve and compress multimedia information, new scalable browsing algorithms allowing access to very large multimedia databases, and semantic visual interfaces integrating the above components into unified multimedia browsing and retrieval systems.

The aim of these systems is to handle general queries such as “find outdoor pictures or videos of an interview with James Cameron discussing the making of the Titanic film.” Answering such queries requires intelligent exploitation of both speech and visual content. For multimedia retrieval, the combination of multiple integrated media types increases the performance of content-based retrieval. Available content analysis and retrieval techniques tailored to a specific media are therefore not adequate for queries as the one mentioned above. Clearly, Multimedia Information Retrieval is a very broad area covering both structural issues (e.g. framework, storage, networking, client-server models) and intelligent content analysis and retrieval. These all need to be integrated into a seamless whole which involves expertise from a wide variety of fields.

## Challenges

*Multimedia Input Analysis.* Many research challenges remain in areas such as inter-media segmentation, partial input parsing and interpretation, and partial multimedia reference resolution. New interactive devices (e.g., force, olfactory, and facial expression detectors) need to be developed and tested to provide new possibilities, such as human emotional state detection and tracking. Techniques for media integration and aggregation should be further refined to ensure syner-

gistic coupling among multiple media, managing input that is impartial, asynchronous, or varies in level of abstraction. Algorithms developed for multimedia input analysis have proven beneficial for multimedia information access [2]. The crossover of algorithms from input analysis to artifact processing and vice versa will continue to be an area of further research opportunity.

*Multimedia Output Generation.* Important questions remain regarding methods for effective content selection, media allocation (e.g., choosing among language, non-speech audio, or gesture to direct attention), and modality selection (e.g., realizing language as visual text or aural speech). In addition, further investigation remains to be done in media realization (i.e., choosing how to say items in a particular media), media coordination (cross modal references, synchronicity), and media layout (size and position of information) [3].

*Multimedia Collaboration.* Discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired. In a multimodal collaboration environment many questions remain: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual (or group) attention and intention across media. Careful and clever instrumentation and evaluation of collaboration environments [4] will be key to learning more about just how people collaborate.

*Agent Interfaces.* Agents are present in learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural interaction. For example, they might be able to adapt sessions to a user, deal with dialog interruptions or follow-up questions, and help manage focus of attention. Agents raise important technical and social questions but equally provide opportunities for research in representing, reasoning about, and realizing agent belief and attitudes (including emotions). Creating natural behaviors and supporting speaking and gesturing agent displays [8] are important user interface requirements. Research issues include what the agents can and should do, how and when they should do it (e.g., implicit versus explicit tasking, activity, and reporting), and by what means should they carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior and who is responsible when things go wrong.

*Machine Learning.* Machine learning of algorithms using multimedia promises portability across users, domains, and environments. There remain many research opportunities in machine learning applied to multimedia such as on-line learning from one medium to benefit processing in another (e.g., learning new words that appear in newswires to enhance spoken language models for transcription of radio broadcasts). A central challenge will be the rapid learning of explainable and robust systems from noisy, partial, and small amounts of learning material [1]. Community defined evaluations will be essential for progress; the key to this progress will be a shared infrastructure of benchmark tasks with

training and test sets to support cross-site performance comparisons.

*Neuroscience-inspired Models.* Observations of child learning and neuroscience suggest that exploiting information from multiple modalities (i.e., audio, imagery, haptic) reduces processing complexity. For example, researchers [5] have begun to explore early word acquisition from natural acoustic descriptions and visual images (e.g., shape, color) of everyday objects in which mutual information appears to dramatically reduce computational complexity. This work, which exploits results from speech processing, computer vision, and machine learning, is being validated by observing mothers in play with their pre-linguistic infants performing the same task.

Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuroanatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

*Resource Requirements.* To assess research effectively in multimedia retrieval, task-related standardized databases on which different groups can apply their algorithms are needed. In text retrieval, it has been relatively straightforward to obtain large collections of old newspaper texts because the copyright owners do not see the raw text being of much value, however image, video, and speech libraries do see great value in their collections and consequently are much more cautious in releasing their content. While it is not a research challenge, obtaining large multimedia collections for widespread evaluation benchmarking is a practical and important step that needs to be addressed. One possible solution is that task-related image and video databases with appropriate relevance judgments are included and made available to groups for research purposes as is done with TREC [6]. Useful video collections could include news video (in multiple languages), collections of personal videos and possibly movie collections. Image collections would include image databases (maybe on specific topics) along with annotated text - the use of library image collections should also be explored.

## Concluding remarks

Multimedia analysis is an emerging research area that has received growing attention in the research community over the past decade. Though modeling and indexing techniques for content-based image indexing and retrieval domain have reached reasonable maturity [7], content-based techniques for multimedia data, particularly those employing spatio-temporal concepts, are at the infancy stage. Content representation through low-level features has been addressed fairly, and there is a growing trend towards bridging the semantic gap. Monomodal approaches have proven successful to a certain level, and more efforts are being put for fusion of multiple media. As visual databases grow bigger with advancements in visual media creation, compaction, and sharing, there is a growing

need for storage-efficient and scalable search systems.

## References

- [1] I. Cohen, N. Sebe, F.G. Cozman, M.C. Cirelo, and T.S. Huang. Semi-supervised learning of classifiers: Theory and algorithms for bayesian network classifiers and applications to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.
- [2] N. Dimitrova. Multimedia content analysis: The next wave. In *International Conference on Image and Video Retrieval*, pages 9–18, 2003.
- [3] N. Dimitrova, H-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *IEEE Multimedia*, 9(3):42–55, 2002.
- [4] M. T. Maybury, editor. *Intelligent Multimedia Information Retrieval*. AAAI/MIT Press, 1997.
- [5] D. Roy and A. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.
- [6] A. Sematon and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *International Conference on Image and Video Retrieval*, pages 19–27, 2003.
- [7] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [8] G. Wei, V. Petrushin, and A. Gershman. From data to insight: The community of multimedia agents. In *International Workshop on Multimedia Data Mining*, pages 76–82, 2002.