Introduction to the Special issue on

Semantic Analysis for Interactive Multimedia Services

Yiannis Kompatsiaris	Yannis Avrithis	Noel E. O'Connor
ikom@iti.gr	iavr@image.ntua.gr	oconnorn@eeng.dcu.ie
http://mklab.iti.gr	http://www.image.ntua.gr/~iavr/	http://www.eeng.dcu.ie/~oconnorn/
CERTH, ITI, Greece	NTUA, Athens, Greece	Dublin City University, Ireland

This editorial introduces the Special Issue, which contains a selection of papers that present recent advances in a number of areas relating to semantic analysis for interactive multimedia services. Recent progress in hardware and communication technologies has resulted in a rapid increase in the amount of multimedia information available to users. The usefulness of multimedia applications is largely determined by the accessibility of the content, so the demand for efficient methods for extracting knowledge from multimedia content has led to a growing research community. Technologies for retrieving semantic information from multimedia content potentially enable new applications and services for both commercial and personal end users. Some examples of such applications include access to multimedia content by the growing number of knowledge workers in today's information society, access by consumers to entertainment or educational content in their home or when mobile, and sharing of content by both professional and private content owners. Whatever the application, it is clearly acknowledged within the research community that user access is best served by a deep understanding of the real world semantics depicted in the content.

The Issue attempts to present a representative sample of ongoing research focusing mainly on analysis, retrieval, structuring visualization, interactive search, vision-based perceptual interfaces, representation and interpretation of analysis results and real-time pre-processing for knowledge extraction. All of these can be considered as component technologies that are required for addressing the broader issue of the so-called 'Semantic Gap' currently facing the content-based information retrieval (CBIR) research community.

The Special Issue Call for Papers received a strong response from the community. A total of 16 manuscripts were submitted for consideration. Of these submissions, 7 papers were accepted following a rigorous review process, coordinated by the guest editors. We are grateful to all reviewers for their effort to ensure the highest possible quality in all accepted papers.

Matching local interest point descriptors is becoming an approach of increasing importance and applicability to image retrieval, classification and object recognition problems. This usually involves searching for one to one correspondences among particular descriptors, as well as establishing affinity

among groups of descriptors. *Kosinov, Bruno* and *Marchand-Maillet* express this as an eigenvalue problem, where the principal eigenvector's components render the importance values of individual descriptors, while the corresponding eigenvalue represents an estimate of the overall strength of affinity between images being matched. This yields a spatially-consistent descriptor matching method that is highly efficient in the domain of content-based image retrieval.

Still in content-based image retrieval, relevance feedback is a quite popular way of involving the user in an interactive search process towards modeling the underlying semantics of queries. In this case, new combinations of descriptors are generated during multi-image queries consisting of positive and negative selections. *Arevalillo-Herraez*, *Zazares*, *Benavent*, and *de Ves* use fuzzy sets to model the user's interest in each image in the repository. Positive and negative selections are used to determine the degree of membership of each picture to this set. The attempt is to capture the meaning of a selection by modifying a series of parameters at each iteration to imitate user behavior, becoming more selective as the search progresses.

A similar matching problem extended to video, is to detect similar or near-duplicate (repeated) sequences. Applied to TV broadcasts, this may allow inter-program sequence detection (commercials, jingles, credits, ...), and consequently automatic temporal structuring and extraction of TV programs. *Berrani*, *Manson*, and *Lechat* model the problem by a micro-clustering technique that groups similar audio/visual feature vectors. This is performed in a non-supervised fashion, not requiring any manually created reference database, and continuously analyzing the broadcasts to periodically return analysis results. This approach is at the root of many novel services related to TV broadcast and in particular to TV-on-Demand services.

Jean and *Albu* present an example where a vision-based perceptual interface is successfully used for a real application. The results of visual analysis in this paper are used to drive an interface used in a music application. More specifically, they propose a new perceptual interface for the control of computer-based music production. They address the constraints imposed by the use of musical meta-instruments during live performance or rehearsal by tracking feet motion relatively to a visual keyboard. A real-time, two levels feet tracking algorithm, namely a coarse level for foot regions, and a fine level for foot tips, is used. The output of the tracking is used for the spatiotemporal detection of key-"press" events.

Luo, Fan and Sato present a novel scheme for effective analysis, retrieval and exploration of large-scale news video collections by performing multi-modal video content analysis and synchronization. Focusing on large-scale collections, they combine analysis results with a novel hyperbolic visualization scheme to visualize large-scale news topics according to their associations and interestingness. Different sources of available information are exploited, first,

automatic keyword extraction is performed on news closed captions and audio channels to detect the most interesting news topics and second, visual semantic items, such as human faces, text captions, video concepts, are also extracted automatically by using semantic video analysis. The news topics are automatically synchronized with the most relevant visual semantic items. In addition, an interestingness weight is assigned for each news topic to characterize its importance.

A key question that arises when addressing semantic knowledge extraction is how to map the results of multiple content analysis algorithms to an integrated representation that supports inference of complex real-world situations. In their paper, *Fernández, Baiget* and *Roca* propose a system designed to automatically provide high-level interpretations of complex real-time situations in outdoor and indoor surveillance scenarios. They propose a high-level architecture that enables the integration of semantic information and content analysis from multiple camera sources. The resulting architecture is modular, allowing both top-down and bottom-up information flow and has been designed to integrate ontological resources for cooperation with the reasoning stage. The system is also potentially extensible to other content analysis modalities (e.g. audio) and supports multi-lingual functionality.

Much work on semantic knowledge extraction in video is based on the assumption that a camera shot, rather than a video frame, is the atomic unit of access to the content. For this reason, significant work has been reported in the literature on shot boundary detection. Indeed, to large degree it is considered a solved problem in the uncompressed domain when algorithms have access to the raw pixel data. However, many real world applications cannot tolerate the luxury of assuming uncompressed video. For this reason, compressed domain approaches to shot boundary detection are increasingly important. However, the compression community continues to develop new coding schemes with new forms of data present in the associated bitstreams. Thus it is important that real-time compressed domain shot boundary detection keep pace with these new compressed representations. *De Bruyne et al* present a novel shot boundary detection technique that operates completely in the compressed domain using the recent H.264/AVC video standard. The algorithm is tailored to the standard's new coding tools and hierarchical bitstream patterns. Interestingly, as well as increased efficiency, their experimental results also show that the proposed algorithm achieves a high accuracy.

In conclusion, we believe that this special issue includes a selection of papers that present recent exciting results in the field of semantic analysis for interactive multimedia services. From these papers it is clear that there is significant effort being invested by the multimedia analysis research community in a number of different areas and applications. Looking to the future, we believe that continuous exchange of ideas and results between such endeavors is essential in order to ensure that the community can meet the demanding user-centric challenges of

future multimedia services and applications.

Ioannis Kompatsiaris received the Diploma degree in electrical engineering and the Ph.D. degree in 3-D model based image sequence coding from Aristotle University of Thessaloniki, Greece, in 1996 and 2001, respectively. He is a Senior Researcher (Researcher C') with the Informatics and Telematics Institute currently leading the Multimedia Knowledge Laboratory. His research interests include semantic multimedia analysis, indexing and retrieval, multimedia and the Semantic Web, knowledge structures, reasoning and personalization for multimedia applications. He is the coauthor of 6 book chapters, 23 papers in refereed journals and more than 60 papers in international conferences. He is a member of IEEE.

Yannis Avrithis received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens in 1993, the M.Sc. in Communications and Signal Processing (with Distinction) from the Department of Electrical and Electronic Engineering of the Imperial College of Science, Technology and Medicine, University of London, UK, in 1994, and the PhD from NTUA on digital image processing in 2001. He is currently a senior researcher at the Image, Video and Multimedia Systems Laboratory (IVML) of NTUA, coordinating R&D activities in Greek and EU projects, and lecturing in NTUA. His research interests include image/video segmentation and interpretation, knowledge-assisted multimedia analysis, annotation, content-based and semantic indexing and retrieval, video summarization and personalization. He has published 1 book, 13 articles in international journals, 15 book chapters and 65 in conferences and workshops. He is an IEEE member, and a member of ACM and EURASIP.

Noel E. O'Connor is a Senior Lecturer in the School of Electronic Engineering of Dublin City University and a Principal Investigator in the Science Foundation Ireland funded CLARITY Centre for Science and Engineering Technology. Since 1999 he has published over 130 peer-reviewed publications, filed 5 patents and spun off a campus company. He has acted as PC Chair for 3 international conferences. He has guest edited 5 special issues of different journals. His research interests are in the broad field of multi-modal content analysis for knowledge extraction, including work on audio and visual analysis (including stereo and multi-spectral sources), content-based information retrieval, user interface design and hardware for media processing on mobile devices. He is a member of the IEEE, Engineers Ireland and the Institution of Engineering and Technology.