Semantic Image and Video Indexing in Broad Domains

I. Introduction

IMAGE and video collections are abundant, but where for document retrieval standard products are widely available, access to image and video collections is still cumbersome. The reason is the semantic gap between what we can derive automatically from the visual data and the semantic interpretation a user has of the same data. To bring semantic access, bridges across the semantic gap have to be developed.

The semantic gap has dictated that solutions to image and video indexing could only be applied in narrow domains using specific concept detectors (e.g., "sunset"). This leads to lexica of at most 10–20 concepts. A number of factors gave the impetus to go beyond such small lexica. New machine learning techniques, such as one-class classifiers, have been developed which deal with the special characteristics of multimedia data with sparse and heterogeneous examples. The use of multi-modal indexing where visual information is effectively combined with textual information can employ the strength of each modality, rather than relying on a single modality. Important factors are also the availability of large annotated information sources e.g. the LSCOM and MediaMill video annotations of news data or the Corel collections of various scenes and object, and the computational resources to use them. These trends have paved the way to increase lexicon size by orders of magnitude (now 100, in a few years 1000). This brings it within reach of research in ontology engineering i.e. creating and maintaining large typically 10 000+ structured sets of shared concepts. The research in data-driven image and video indexing and top-down ontology engineering has traditionally been within different communities. When recent advances in both fields are combined, we are reaching the point where the semantic gap can be bridged for many concepts.

To bring semantics to the user in broad domains both the indexing and retrieval step have to be reconsidered. This special issue brings together a number of high-quality papers which address both steps and its relation to ontologies. Jointly, they set the scene for a big leap over the semantic gap in broad domains such as film, news, sports, and personal albums.

II. Image and Video Indexing

Image and video indexing is based on a lexicon of detectors where each detector gives a measure for the presence of the concept in the visual data. The basis for recognition are features extracted from the visual data. However, supervised recognition methods require large sets of examples annotated by experts for every concept in the lexicon. When the domain is broad and many concepts need to be annotated one has to rely on collaborative annotation to make it feasible. However, when several people are annotating visual data, they might differ in their interpretation leading to inconsistent annotations. Volkmer [6] presents a method where the annotations are analyzed to filter out inconsistent annotations resulting in annotated datasets of higher quality.

The statistics of the features extracted from the annotated data pose an implicit structure upon the different concepts. The paper by Koskela [3] considers several measures of similarity to compare different pictures or shots and provides methods to learn and analyze this implicit structure. This learned structure is only one structure to employ. Ontology engineering takes a top-down approach to the problem and defines structures in a declarative manner. General ontologies such as Wordnet can yield the relations required for broad domains. To use this information, the concepts in the detector lexicon have to be linked to the ontology as done by Snoek [5].

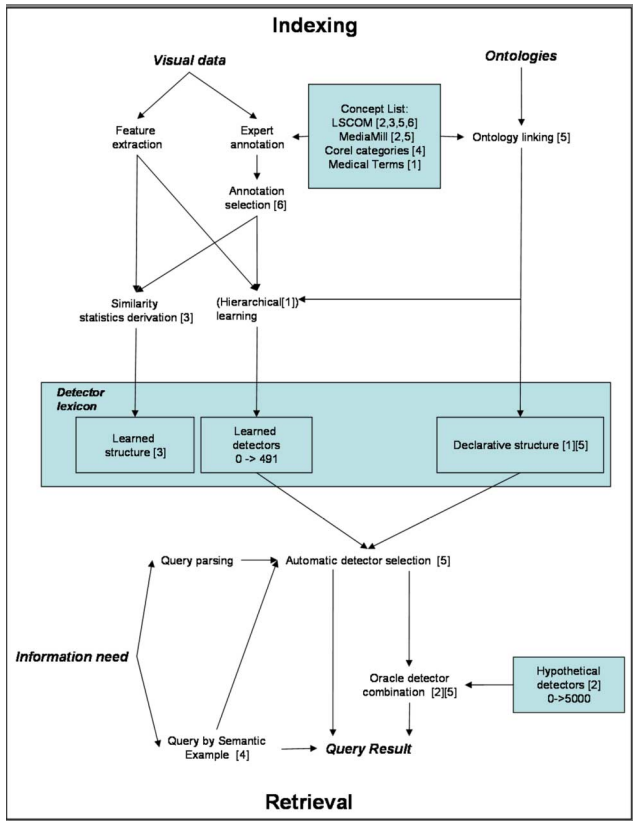

Fig. 1. Overview of the main focus of the different papers in this special issue. The figures gives a blueprint for a future architecture of a system which will bridge the semantic gap.

Digital Object Identifier 10.1109/TMM.2007.898913

When a learned or declarative structure is added to the lexicon, we can use the hierarchical relations between the concepts to improve the detection process. Fan [1] defines a hierarchical learning method which explicitly takes the structure into account.

Learning a large lexicon of concepts is a computation-intensive process. But as the process is off-line and concepts can be learned in parallel, the computational resources in principle are there to get to structured lexicons of 10 000+ concepts, connected to an ontology such as Wordnet. The current bottleneck is the availability of a sufficient number of annotated examples.

III. Image and Video Retrieval

In retrieval the user approaches the visual archives with some information need and expects results which satisfy this need. In early systems "query by visual example" has been a popular query mode. However, as this is a low-level specification the result is often poor. With a large detector-based lexicon alternative methods of querying the archive emerge. Rasiwasia [4] presents a new query paradigm: "query by semantic example". To this end, detectors are first used to analyze semantic characteristics of the visual data and subsequently the results are compared to the stored data.

When visual examples are not available, the most appropriate query mode is just expressing the information need as a textual expression. The task for the system is then to select the most appropriate set of detectors from the lexicon. Snoek compares a number of different strategies based on visual examples, the description of a concept, and the structure of Wordnet using a 491 concept-detector lexicon.

At this point the latter lexicon of learned detectors is the largest one available. Hauptmann [2] explores what lies beyond and uses a lexicon of perfect oracle detectors with noise added to arrive at the conclusion that for the broad domain of news, approximately 5000 concepts with an average precision of 0.1 would be sufficient for practical use.

Both Snoek and Hauptmann consider the use of detector combinations in retrieval the important step forward but resorted to oracle detector combinations to analyze its potential.

IV. Conclusion

Large-scale processing is becoming possible; the computational resources are there. The papers in this special issue show that we are now making concrete steps towards general data-driven methods that can be linked to a significant number of meaningful concepts.

Overviewing this arena, we see the following three issues as the major research challenges for the coming years. Firstly, we need to solve the problem of large-scale annotation. One option is to find ways to use large existing web collections for this purpose. Secondly, methods have to be developed for detector combination. Two of the papers show that oracle-based detector combination can improve retrieval. Thirdly, advances in quality improvement and scalability are still needed. One paper posed 5000 concept detectors with 10% mean average precision as the target. Time will have to show whether this is the right goal. In fact, if detector combination becomes possible (Hauptman's estimate is based on isolated detectors) and more elaborate links between the detector lexica and ontologies are established and used in learning, the target might actually be lower.

Acknowledgment

We would like to thank the Editor-In-Chief, Hong-Jiang Zhang, for allowing us to make this exciting special issue. Thanks also to Associate Editor Mohan Kankanhalli for handling the paper, of which we are co-authors. Last, but not least, thanks to the reviewers for their extensive efforts in providing comments and suggestions for the submitted papers.

Marcel Worring, *Guest Editor*
Department of Computer Science
University of Amsterdam
Amsterdam, The Netherlands
worring@science.uva.nl

Guus Schreiber, *Guest Editor*
Free University Amsterdam
Amsterdam, The Netherlands
guus@cs.uva.nl

References

[1] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology to boost hierachical classifier training for automatic multi-level annotation," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 939–957, Aug. 2007.

[2] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[3] M. Koskela, A. F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 912–922, Aug. 2007.

[4] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.

[5] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.

[6] T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi, "Modelling human judgement of digital imagery for multimedia retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 967–974, Aug. 2007.

Marcel Worring is an Associate Professor of Computer Science at the University of Amsterdam, Amsterdam, The Netherlands. He is leading the MediaMill team, which has participated in all TRECVID editions and has published over 100 scientific publications in journals and proceedings.

Dr. Worring is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the Chair of the IAPR TC12 on Multimedia and Visual Collections.

Guus Schreiber is a Professor of Intelligent Information Systems at the Free University Amsterdam, Amsterdam, The Netherlands. He is also Scientific Director of the IST Network of Excellence "Knowledge Web". He has published some 140 articles and books. In 2000, he published (with MIT Press) a textbook on knowledge engineering and knowledge management, based on the CommonKADS methodology.

Dr. Schreiber is Co-Chairing the W3C Semantic Web Deployment Working Group and is the former Co-Chair of the Web Ontology and the Semantic Web Best Practices groups.