Report from the MMM 2019 Special Session on Multimedia Analytics: Perspectives, Techniques and Applications (MAPTA 2019)

Report by Björn Þór Jónsson, Cathal Gurrin, and Benoit Huet.

Björn Þór Jónsson (http://www.itu.dk/people/bjth/) is an associate professor in the Computer Science Department at the IT University of Copenhagen, Denmark. His research focuses on interactivity and scalability of multimedia analytics and multimedia retrieval applications.

Cathal Gurrin (http://www.computing.dcu.ie/~cgurrin/) is an associate professor at the School of Computing, at Dublin City University, Ireland and a principal co-investigator at the Insight Centre for Data Analytics. His research focuses on multimedia information retrieval and personal data analytics, with special emphasis on lifelogging applications.

Benoit Huet (http://www.eurecom.fr/en/people/huet-benoit) is an assistant professor in the Data Science Department at EURECOM, France. His research focuses on multimedia content analysis and understanding with a particular emphasis on various media types (such as archives, broadcast, and user generated content).

This report summarizes the presentations and discussions of a special session on “Multimedia Analytics: Perspectives, Techniques and Applications” (http://www.itu.dk/people/bjth/MAPTA/) held during the 25th International Conference on MultiMedia Modeling (MMM 2019), in Thessaloniki, Greece on January 9, 2019. The special session consisted of five brief technical presentations, followed by a panel discussion with questions from the audience, moderated by Benoit Huet. The goal of this report is to record the conclusions of the special session, in the hope that it may serve members of our community who are interested in Multimedia Analytics.

Figure 1: Daniel Seebacer presents the first paper at MAPTA 2019, with moderator Benoit Huet observing.

Presentations

First, Daniel Seebacher presented “From Movement to Events: Improving Soccer Match Annotations” [2]. Daniel presented a novel method for semi-automatic definition and detection of events in soccer matches based on player and ball movements. They enabled analysts to visually define and search for complex, hierarchical events, and showed that the required annotation time for complete matches by using our system can be reduced to a few seconds while achieving a similar level of performance to manual annotation.

Second, Lyndon Nixon presented “Multimodal Video Annotation for Retrieval and Discovery of Newsworthy Video in a News Verification Scenario” [2]. Lyndon described a novel system that combines advanced technologies for social-media-based story detection, story-based video retrieval and concept-based video fragment labeling. He also described a professional analytics dashboard and showed, through a case-based study with journalists, that the journalists were well supported in their content discovery work.

Third, Björn Þór Jónsson presented “Integration of Exploration and Search: A Case Study of the M^3 Model” [3]. He presented an initial case study of integrating exploration and search within a single multidimensional media browser, extending the explorative M^3 model to allow searching within an exploration context and exploring within a search context. Following a performance study, he then proposed some research directions for scalability of multimedia analytics.

Fourth, Werner Bailer presented “Face Swapping for Solving Collateral Privacy Issues in Multimedia Analytics” [4]. Werner proposed to solve some privacy issues in multimedia analytics by replacing faces in images by using a generative adversarial network to generate new face

images. He then demonstrated that face swapping does not impact the performance of visual descriptor matching and extraction.

Finally, Seán Quinn presented “The Impact of Training Data Bias on Automatic Generation of Video Captions” [5]. He presented work done to evaluate the impact of bias in training data on the quality of caption generation. Their preliminary findings showed that pruning training data to make it more homogeneous, or diverse, does improve performance, especially compared to random pruning.

Figure 2: The special session was well attended and attendees engaged in a lively discussion with the panel.

Discussions

Following the presentations, Benoit Huet initiated a panel discussion between the speakers and the audience. The first question asked whether generating avatars instead of faces was an option for solving privacy issues. The conclusion of the discussion was that doing so might achieve some of the privacy goals. If the avatars were generated consistently, however, some of the privacy might be lost for people appearing often and in identifiable locations. Furthermore, using avatars might result in changes in the statistical properties of the media signals, which could lead to unexpected effects, while Werner’s approach avoids such statistical changes.

A discussion then started about the importance of interfaces for analytics: when users have little idea where to start and what to look for, creating useful interfaces might be a significant challenge. It was pointed out that many current systems, including systems competing in the Video Browser Showdown (VBS), often start with some kind of query and thus fail when the

user is uncertain about how to start. When the user does not have a specific query in mind, showing information about metadata types and distributions could help significantly, which is indeed a key role of dashboards in analytics applications. Relevance feedback was also mentioned as a potential solution to the problem, but overall the “cold-start” problem remains a challenge for some analytics applications when users are unsure of their intentions or how best to formulate a query.

It was pointed out, however, that evaluation of interfaces is both a very difficult research topic and one that is hard to publish. Groups participating in VBS, for example, are generally not large enough to allow for a novice-friendly interface to be designed. As a result, the interfaces are essentially “overfitted” to their developers, which is why the novice session was added to the competition. Interestingly, the “novices” found at MMM are probably quite well prepared compared to the general population, so that fact that even those “novices” perform significantly worse than the system developers does not bode well for the user-friendliness of the current generation of VBS systems.

Following up on this interface discussion, there was a question about the challenges involved in bridging the gap between traditional database approaches, such as SQL queries, and more multimedia-oriented approaches, such as nearest-neighbor search. Given the complexity of teaching SQL to even CS students, giving novice users a query language interface is likely not a viable approach, the gap from query languages to actual usability is simply to wide. Based on some experiences, users want to start quickly and simply; they will not formulate complex queries to start their interaction, but they are happy to subsequently apply multiple filters and thus gradually build up a complex query. Developing an interface that adapts dynamically to the expertise level of the user was proposed as an interesting approach. Users must also trust the system, so giving users insight into why the system is proposing results is important. Overall, it was pointed out that the search-exploration axis proposed for multimedia analytics applications outlines the different types of tasks that need to be solved, but not how to solve them; the question that interface developers then face is how to implement the tasks of the axis.

A question was asked about the translation of the audio tracks of videos to text transcriptions: How far are we from a “perfect” solution? While state-of-the-art approaches report excellent results on benchmarks, applying the same approaches to the VBS collection yielded very unsatisfactory results. So are the datasets producing bias? It was pointed out that many benchmarks datasets contain very simple actions with limited vocabulary, while real videos depict complex scenarios with general descriptions, and on top of that datasets rarely capture cultural differences. Overall, if the training data has problems, they will be reflected in the solutions as well. However, given the progress made in machine learning over the last decade, as well as recent progress in zero-shot and few-shot learning approach, there is reason to believe that approaches may develop that can handle specialized situations with small training sets. With smaller datasets, however, the risk of overfitting and biases is higher, so the ability to detect such problems becomes important.

As time was running short, the participants were asked to give their perspectives. In short summary, their perspectives showed that work is needed at all levels of multimedia analytics systems, from the underlying multimedia understanding algorithms used by the system itself, through the models of interactions with users, to the actual system interfaces that the users collaborate with. Clearly, the research field of multimedia analytics is ripe for more work in the coming years.

Acknowledgments

The session was organized by the authors of the report, in collaboration with Stevan Rudinac and Laurent Amsaleg, who could not attend MMM. The panel format of the special session made the discussion much more lively and interactive than that of a traditional technical session. We would like to thank the presenters and their co-authors for their excellent contributions, as well as the members of the audience who contributed greatly to the success of the session.

References

The following papers were presented at the special session and published in the proceedings of MMM 2019:

[1] Manuel Stein, Daniel Seebacher, Tassilo Karge, Tom Polk, Michael Grossniklaus, and Daniel A. Keim. *From Movement to Events: Improving Soccer Match Annotations.* In International Conference on Multimedia Modeling (pp. 130-142). Springer.

[2] Lyndon Nixon, Evlampios Apostolidis, Foteini Markatopoulou, Ioannis Patras, and Vasileios Mezaris. *Multimodal Video Annotation for Retrieval and Discovery of Newsworthy Video in a News Verification Scenario*. In International Conference on Multimedia Modeling (pp. 143-155). Springer.

[3] Snorri Gíslason, Björn Þór Jónsson, and Laurent Amsaleg. *Integration of Exploration and Search: A Case Study of the M^3 Model*. In International Conference on Multimedia Modeling (pp. 156-168). Springer.

[4] Werner Bailer. *Face Swapping for Solving Collateral Privacy Issues in Multimedia Analytics*. In International Conference on Multimedia Modeling (pp. 169-177). Springer.

[5] Alan F. Smeaton, Yvette Graham, Kevin McGuinness, Noel E. O'Connor, Seán Quinn, and Eric Arazo Sanchez. *The Impact of Training Data Bias on Automatic Generation of Video Captions*. In International Conference on Multimedia Modeling (pp. 178-190). Springer.