Large-Scale Multimedia Data Collections

Benoit Huet
EURECOM

Tat-Seng Chua
National University of Singapore

Alexander Hauptmann
Carnegie Mellon University

The widespread adoption of smartphones equipped with high-quality image-capturing capabilities coupled with the prevalent use of social networks have resulted in an explosive growth of social media content. People now routinely capture the scenes around them and instantly share the multimedia content with their friends over a variety of social networks. The social network functions also ensure that much of this content comes with some form of social annotations.

This environment sets the stage for advances in large-scale media research. It is now easy to gather a huge database of images with social annotations. A resulting challenge then is the design of appropriate tasks to motivate advanced research in new directions. This involves both fundamental studies of use cases and task design methodologies as well as the development of new datasets to cater to different needs.

Given the large quantities of data, acquiring ground truth becomes a tedious and challenging endeavor. The current trend is toward leveraging crowdsourcing platforms such as Amazon Mechanical Turk to help gather ground truth. However, these methods could result in incomplete and/or unreliable "ground truth." In addition, given large-scale datasets, a variety of research, both intended and emerging, will be carried out on these datasets. Thus, questions regarding the reusability and repurposing of datasets to different research needs will arise.

Finally, in this rapidly changing environment, how can we anticipate and plan for new directions in large-scale multimedia research?

This special issue hopes to address these challenges. The call for this special issue was launched after the successful conclusion of the ACM Workshop on Very-Large-Scale Multimedia Corpus, Mining, and Retrieval at the ACM Multimedia 2011 Conference. The articles in this issue cover identification of use cases and task design, dataset development, and basic research over existing datasets. (For more information, visit http://youtu.be/0OtXnqe67Ns to view a video of us discussing the importance of large-scale multimedia collections.)

Special Issue Articles

Motivated by the increased ease with which users can collect and manipulate large amounts of multimedia data and the maturation of techniques for effectively exploiting crowdsourcing, "The Community and the Crowd: Multimedia Benchmark Dataset Development" by Martha Larson, Mohammad Soleymani, Maria Eskevich, Pavel Serdyukov, Roeland Ordelman, and Gareth Jones examines the fundamental tasks of identifying use cases and design tasks. Their goal is to help drive the state of the art forward as well as the generation of ground truth. They discuss how to affordably generate high-quality ground truth multimedia data collection, especially for video, using crowdsourcing. They conclude with a discussion on the Internet video collection developed for the MediaEval benchmarking initiative.

Also on this topic, "Building Reliable and Reusable Test Collections for Image Retrieval: The Wikipedia Task at ImageCLEF" by Theodora Tsikrika, Jana Kludas, and Adrian Popescu discusses the creation of image test collections for the Wikipedia task at ImageCLEF. Based on their experience with organizing evaluations over four years, from 2008 to 2011, they discuss the issues in constructing image collections, topic development, and ground truth creation. After analyzing the reliability and reusability of the resulting test collections, they provide best practice guidelines for building such test collections.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon, in their article "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," present a new

1070-986X/12/$31.00 © 2012 IEEE

Published by the IEEE Computer Society

temporal 2D facial expression database named Acted Facial Expressions in the Wild (AFEW) and its static subset. The database consists of 1,426 video sequences covering 330 subjects extracted from 54 movies. It differs from current databases that tend to be smaller and contain facial expressions captured in controlled environments in relatively static poses. The authors labeled their video clips with one of six basic expressions and captured varied facial expressions; natural head poses and movements; occlusions; subjects from multiple races, both genders, and diverse ages; and multiple subjects in a scene. To facilitate ground truth collection, the article presents a semiautomatic recommender system that suggests only the video clips with a high probability of containing subjects showing meaningful facial expressions. It further proposes experimentation protocols for comparing research results. Overall, this article provides a valuable contribution toward the creation of a facial expression database in real-world environment for facial expression research.

In ''Threefold Dataset for Activity and Workflow Recognition in Complex Industrial Environments,'' Athanasios Voulodimos, Dimitrios Kosmopoulos, Georgios Vasileiou, Emmanuel Sardis, Vasileios Anagnostopoulos, Constantinos Lalos, Anastasios Doulamis, and Theodora Varvarigou also describe a dataset construction project. The authors present a specialized workflow recognition dataset that targets behavioral-recognition research in real-life industry environment. When capturing this large-scale multicamera dataset, the authors used four cameras to depict workers executing industrial workflows of different tasks, such as carrying parts and welding them together in a visually complex industrial environment. The recorded videos include severe occlusions with frequent illumination changes. The dataset is divided into three parts, and the ground truth incorporates activity labeling and a set of holistic features for scene representation. This is the first dataset that covers real workflows in a natural setting, providing a challenging dataset for behavioral- and workflow-recognition research.

The final article, ''Indexing Large Online Multimedia Repositories Using Semantic Expansion and Visual Analysis'' by Xavier Sevillano, Tomas Piatrik, Krishna Chandramouli, Qianni Zhang, and Ebroul Izquierdo, describes the use of existing large-scale image datasets to perform the task of predicting general tags of images from the associated textual metadata and visual features. It focuses specifically on the geotagging task that relies on extracted named entities by exploiting the complementary textual resources such as Wikipedia and WorldNet using a natural language processing framework. The authors conducted two experiments based on the MediaEval 2010 and MediaEval 2011 datasets.

The development of quality large-scale datasets is pivotal to motivating and accelerating research in the desired and new directions.

Future Research

Several factors have thus far provided a solid foundation for large-scale media research: the availability of various kinds large media content, affordable large computing and cloud resources to store and process them, and an eager crowd to perform annotations if guided well for ground truth generation. The development of quality large-scale datasets is pivotal to motivating and accelerating research in the desired and new directions.

In the future, numerous challenges and open issues must still be tackled. First, we must design test datasets to cater to current and emerging research, especially those that involve users and social interactions. Also, rapidly changing technologies and content necessitate the design of an evolving dataset.

Given these initial requirements, establishing ground truth for large-scale and evolving datasets through crowdsourcing becomes important. In particular, how can we institute a working, evolving dataset in which users can continually contribute new data, use cases, and ground truth? **Thus**, an interesting line of new research will be looking at ways to unify different individual efforts to create a

unified Web-scale repository for experimental evaluation.

Researchers must also look to design specific tools with optimal interfaces, allowing easier labeling and class assignment of multimedia data. Lastly, it will be important to develop semi-automated techniques in conjunction with crowdsourcing for dataset annotation and ground truth generation.

This area is rich in challenges and ready for practical research. We expect new research topics and resources to continuously emerge. **MM**

Benoit Huet is an assistant professor in the Multimedia Information Processing Group at EURECOM. His research interests include computer vision, content-based retrieval, multimedia data mining and indexing, and pattern recognition. Huet has a DPhil in computer science from the University of York, UK. Contact him at Benoit.Huet@eurecom.fr.

Tat-Seng Chua is a KITHCT Chair Professor in the School of Computing at the National University of Singapore. His research interests include multimedia information retrieval and live media search of vast user-generated content. Chua has a PhD from the University of Leeds. Contact him at chuats@comp.nus.edu.sg.

Alexander Hauptmann is a senior systems scientist in the School of Computer Science at Carnegie Mellon University. His research interests include speech recognition, speech synthesis, speech interfaces, and language. Hauptmann has a PhD in computer science from Carnegie Mellon University. Contact him at alex+@cs.cmu.edu.