Report from the Third International Workshop on Computer Vision meets Databases — CVDB 2007

Laurent Amsaleg
IRISA–CNRS
Laurent.Amsaleg@irisa.fr

Björn Þór Jónsson
Reykjavík University
bjorn@ru.is

Vincent Oria
New Jersey Institute of Technology
vincent.oria@njit.edu

This report summarizes the presentations and discussions of the Third International Workshop on Computer Vision meets Databases, or CVDB 2007, which was held in Beijing, China, on June 10, 2007. The workshop was co-located with the 2007 ACM SIGMOD/PODS conferences and attended by twenty-five participants.

1 Workshop Series Scope

The goal of the CVDB workshop series is to foster interdisciplinary work between the areas of computer vision and databases. We have observed that few researchers in the computer vision community are adopting any of the indexing schemes designed by database researchers. Furthermore, while new and exciting techniques are being developed by computer vision researchers, database researchers are often unaware of such work.

The reason is that, unfortunately, there has been a great gap between the computer vision and database communities. The goal of the CVDB workshop series is to bridge this gap. The idea is to provide database researchers with a snapshot of what computer vision people are dealing with and vice-versa, with the aim of defining research directions that can benefit both communities. There is great expertise on both sides, and the CVDB 2007 workshop was aimed at sharing it by means of keynote speeches, technical presentations, and panel discussions.

2 Workshop Program

We assembled an international program committee of 27 experts from the computer vision and database communities. As was reportedly the case with many other workshops co-located with SIGMOD/PODS 2007, fewer papers were submitted than in previous years. Thus the program committee had to review only nine submitted papers, and in the end, four papers were selected for presentation and publication.

Additionally, we hand-picked two keynote speakers to present their views of the research directions and contributions of the computer vision and database communities. Finally, we assembled a panel to focus on the current and future roles of content-based multimedia retrieval.

After a short introduction, the day started with the first keynote speech on large-scale multimedia retrieval, followed by a technical session with the four papers. After lunch, the second keynote speech, on modeling events with media evidences, was followed by panel discussions.

For details of the papers, tutorials, and panel, please visit the workshop web-site, which will remain open at cvdb07.irisa.fr. The CVDB 2007 proceedings appear in the ACM Digital Library. In the following, however, a summary of the main points of the workshop is presented.

2.1 Keynote I: Large-Scale Retrieval

The first keynote speech, titled "Challenges of and Remedies for Large-scale Multimedia Information Retrieval" was delivered by Edward Chang, director of research at Google, Beijing. According to Edward Chang, with the rapid growth of image and video data, it is increasingly crucial to provide tools that can assist with effective organization and search. Despite advances in several areas, challenges remain for the deployment of a Web-scale multimedia search engine. His presentation described three major challenges and their potential remedies.

The first challenge is image and video annotation, which he claimed is necessary since most users prefer keyword-based search over content-based search. Manual annotation can be subjective and error-prone, whereas machine annotation cannot effectively discover all the desired information. Recent efforts, such as the ESP game, have moved towards fusing human and computer intelligence for improved annotation accuracy.

The second challenge is that measuring similarity, in particular perceptual similarity, is difficult in many cases. For instance, image features can vary based on size and quality of the images. Work on feature constancy can potentially remedy this challenge.

The third challenge that hinders the deployment of a large-scale system is scalability itself. A multimedia

search engine must be able to scale well with respect to both data dimensionality and data quantity. Recent advances in large-scale statistical learning, indexing, and searching were presented.

The major conclusion was that while there are significant challenges, they have been partly addressed and there is continued work on remedies. Furthermore, companies such as Google have been building computing infrastructures that will allow research into scalability, as well as tackling the other challenges at a large scale.

2.2 Technical Papers

The technical paper session consisted of four presentations; it was chaired by Vincent Oria.

First, in [1], Kwietnewski et al. presented the design of a multimedia database application for representing and reasoning about crime scene data. In such a system, a variety of data must be stored at a variety of resolutions, yet grounded in the underlying spatial representation. Second, in [2], Xue et al. presented a description scheme for video content, with support for ontology-based semantic indexing and retrieval. This description scheme integrates domain-specific ontologies and MPEG-7 content description and enhances the semantic interoperability of multimedia. These two papers were jointly awarded a "best student paper" award.

Third, in [3], Ide et al. presented work on name identification of people in news by face matching. Faces are identified using face detection technology and names are identified through closed caption texts; together these evidences allow much improved classification of persons in news. Finally, in [4], Harðarson and Jónsson presented their vision of a personal image browser, which combines OLAP and game-playing technology into a seamless browsing and searching experience.

2.3 Keynote II: Event Modeling

The second keynote speech, titled "Modeling Events with Media Evidences", was delivered by Amarnath Gupta, director of the Advanced Query Processing Laboratory at the San Diego Supercomputer Center. According to Amarnath Gupta, media data such as images and videos often play the role of snapshot evidences of some real-world phenomena. The images and videos themselves are then not the focus, but rather serve some higher purpose.

While images and videos can be assets by themselves, they typically serve more as a documentation of some event or information content. In such applications, it is important to correlate the content of the media data with the states, state-transitions and state aggregates that characterize the events. These applications are further complicated by the fact that events are multi-granular and multi-aspect entities and a single media object might represent more than one granularity of events and a part of or multiple aspects of an event.

The presentation described some interesting and open questions that are raised about modeling events with media evidences. The problem was explored and some initial steps toward a solution were described.

2.4 Panel: Content-Based Retrieval

Last on the agenda was a panel discussion under the heading "Is <type='panel' content='content-based retrieval'> really content-based retrieval?" The panel was moderated by Laurent Amsaleg, and consisted of the two keynote speakers, as well as Wei-Ying Ma, principal researcher at Microsoft Research Asia, and Shin'ichi Satoh, professor at the National Institute of Informatics (NII), Japan.

For years, multimedia researchers have been focusing largely on content-based retrieval. Content-based access to multimedia, however, has never really caught on and the multimedia community has not seen much use of its results in the real world. On the other hand, recent trends appear to be changing the multimedia scene very significantly, and some enormous and extremely popular multimedia repositories already exist, such as Flickr, YouTube and DailyMotion. It is interesting to note, however, that almost all multimedia applications arousing interest today are solely relying on human-defined tags, and in fact have no real facilities for content-based access. Multimedia researchers can now gain access to large data sets, with real usage profiles and key needs. But many questions arise, such as: Does this new multimedia scene increase or decrease the need for content-based access? Do we really believe tags are sufficient for our needs? Can tags ever capture all the information inside multimedia documents such as TV broadcasts, video footage, news, etc.? Do we need this information? And so on.

According to Ed Chang, a key problem is that content-based retrieval has not found any "killer applications". It appears that content-based description cannot achieve accuracy above a certain level, and many seemingly relatively simple tasks, such as automatic video surveillance, have proven to be much harder than anticipated. Many issues, however, have been well addressed in content-based retrieval; for example, scalability has been addressed and good approximate indexing techniques exist.

According to Amarnath Gupta, a key problem is that the need for content-based retrieval has been very ill-defined; for most applications very simple segmentation suffices. At the same time, there are many applications where tags address real needs. And, when needed, tags can be created, either by a company or through a collective effort. While tags may not answer all needs, it is not clear that any other method would perform better.

According to Wei-Ying Ma, content-based retrieval has not been a fruitful area to work in for a long time. While there are some relevant applications, such as content-based copy detection and visual earth applications, he has preferred working on text-based methods for multimedia. He believes, however, that content-based methods may become useful for helping to obtain the tags required for the text-based methods, and leverage or enhance other applications, in particular in cooperation with human efforts.

According to Shin'ichi Satoh, using content is indispensable in this new multimedia scene, due to the explosion of data to be accessed. This applies to both content-based access and analysis. There are applications where tags are not sufficient, and the more information used, the better the application.

Significant discussion was raised on this last point of using more varied information to improve application performance. Wei-Ying Ma believes that we may be ready to tackle the image understanding problem, by using many sources of information, such as data and annotations, as well as the significant existing computing infrastructure. Ed Chang, however, was not optimistic, as feature constancy is very hard and image processing is orders of magnitude more expensive than text/tags applications. Amarnath Gupta believes that more information may indeed improve segmentation and annotation, to name some applications, but that it is unlikely to improve actual understanding of the media.

A discussion was raised on the gap that is appearing between industry and academia. Industry now has access to data, queries, and other application information that academia has no chance of obtaining. Furthermore, some companies have built significant computing infrastructures, which academia has no chance of competing with. There was general agreement that this situation is an issue and that the gap is likely to grow in the future, as the application information is indeed a source of competitive advantage and privacy is also an issue. There was also agreement that industry needs academia as a source of students and solutions, and that there are in fact many companies which do not have access to this data and infrastructures either. The method that academia has been using to gain access to this information, and should continue to use, is to send students to the internship programs at these large companies. Often, they come back with a very interesting academic problem, which may turn into research results.

Several other issues were raised in the discussion, such as some potential applications and business models. Finally, Laurent Amsaleg thanked all the participants for very a fruitful and entertaining panel discussion, and closed the workshop.

3 Workshop Conclusions

The goal of the workshop was to bridge the gap between the database and computer vision communities and to define some research directions that can benefit both communities. A first conclusion that can be drawn is that while content-based retrieval has not yielded many strong applications, content-based analysis has been used with success, and may become even more essential in the future as one component of a multi-faceted approach to many applications. A second conclusion is that although this was the third CVDB workshop, progress is slow and most work still addresses either "CV" aspects or "DB" aspects. In fact, it was believed to be necessary to form a recognized conference to entice more young researchers to this area. Based on the discussions during the workshop, there is certainly no shortage of interesting research directions.

4 Acknowledgements

We would like to thank the program committee members, keynote speakers, panelists, authors, local workshop organizers, and attendees, for making CVDB 2007 a successful workshop. We also express our great appreciation for the support from Reykjavík University and Google China.

References

[1] Marcin Kwietnewski, Stephanie Wilson, Anna Topol, Sunbir Gill, Jarek Gryz, Michael Jenkin, Piotr Jasiobedzki, and Ho-Kong Ng. A multimedia database system for 3D crime scene representation and analysis. In *Proceedings of the Third International Workshop on Computer Vision meets Databases*, Beijing, China, June 2007.

[2] Ling Xue, Yuanxin Quyang, Hao Sheng, and Zhang Xiong. Combine MPEG-7 and Semantic Web to enhance the semantic interoperability in multimedia retrieval. In *Proceedings of the Third International Workshop on Computer Vision meets Databases*, Beijing, China, June 2007.

[3] Ichiro Ide, Takashi Ogasawara, Tomokazu Takahashi, and Hiroshi Murase. Name identification of people in news by face matching. In *Proceedings of the Third International Workshop on Computer Vision meets Databases*, Beijing, China, June 2007.

[4] Kári Harðarson and Björn Þór Jónsson. Breaking out of the shoebox: Towards having fun with digital images. In *Proceedings of the Third International Workshop on Computer Vision meets Databases*, Beijing, China, June 2007.