Report from the First International Workshop on Computer Vision meets Databases — CVDB 2004

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 First International Workshop on Computer Vision meets Databases, or CVDB 2004, which was held in Paris, France, on June 13, 2004. The workshop was co-located with the 2004 ACM SIGMOD/PODS conferences and was attended by forty-two participants from all over the world.

1 Workshop Scope

For a long time, the computer vision community has been working on content-based multimedia retrieval. Researchers from that community aim at defining better content-based descriptors and extracting them from images. The descriptors obtained are often represented as points in multi-dimensional spaces and some metrics are used during similarity retrieval. Their focus is on increasing the recognition power of their schemes and they usually evaluate their strength using data sets that fit in main memory because they try to avoid the secondary storage management burden.

Facilitating the management of very large amounts of data and removing this disk burden has long been a strong motivation for the database community. This is particularly crucial for multimedia databases whose sizes grow very fast. As such, researchers in databases have proposed many smart multidimensional indexing schemes with some elegant algorithms to compute nearest-neighbor and top-n queries.

Yet, it is surprising to see that only few works in the computer vision community have adopted any of these indexing schemes. A common reason evoked is that the description schemes that database researchers use are way too simplistic. Therefore, it is hard for computer vision researchers to foresee how indexes could behave when used with a modern and powerful description scheme. Additional reasons given include the assumptions on the distribution of data, the ability to only retrieve the single nearest neighbor of query points, and the use of approximate search schemes that give little clue as to the quality of the returned results.

The goal of this workshop was to bridge this gap between the two communities. The idea was to provide database researchers with a snapshot of what computer vision people are dealing with and vice-versa, with the aim of defining some research directions that can benefit both communities. There is great expertise on both sides, and this workshop was aimed at sharing it by means of tutorials and presentations. In addition, we provided a panel for exchanging ideas with professional image users and providers.

2 Workshop Program

We assembled an international program committee of 31 experts from the computer vision and database communities. The program committee had to review 25 submitted papers. In the end, eight papers were selected for presentation and publication. Additionally, we hand-picked two tutorialists to present their views of the research directions and contributions of the computer vision and database communities, respectively. Finally, we assembled a panel to focus on the applications of image databases in the near and distant future. We would like to thank the program committee members, tutorialists, and panelists, as well as the authors of all papers, both the accepted and rejected ones.

For details of the papers, tutorials, and panel, including slides from all presentations, please visit the workshop web-site, which will remain open at cvdb04.irisa.fr. The CVDB 2004 proceedings will appear in the ACM Digital Library. Five papers have also been selected for publication in a special issue of the Multimedia Tools and Applications journal.

After a short introduction, the day started with a technical session of four papers, followed by the two tutorials. After lunch, a second technical session of four papers took place, followed by two hours of panel discussions. In the following, a summary of the main points of each of these is presented.

2.1 Computer Vision Tutorial

The computer vision tutorial "Image + Database $\neq$ Image Database" was presented by Roger Mohr, professor of Computer Science at the Institut National Polytechnique de Grenoble, France.

According to Roger Mohr, computer vision researchers have made significant progress with low-level description schemes and many meaningful applications are operational today, although many issues are still open. Many of these successful description schemes are based on some form of local descriptors, where a combination of many individual descriptors together describes the whole image. For these schemes, however, describing millions of images may result in billions of image descriptors. This large amount of data leads to a research challenge for the database community, namely to provide (approximate) search methods that are efficient in high dimensional spaces and can cope with erroneous data (outliers).

On the other hand, little progress has been made on high-level description schemes that increase the abstraction level and return more semantics from the image contents. Such schemes are intented to automatically describe images, for example in terms of objects they may represent ("a bicycle" or "grandma in Venice"). Having such high-level semantics would obviously yield many interesting applications, such as classification based on common concepts, rather than visual similarity.

This lack of progress leads directly to Roger Mohr's second research challenge, directed at the computer vision community, which is to deliver useful semantic information from images. From his point of view learning seems today the only way to go in order to increase the level of the descriptions. Learning, however, poses many hard challenges. For example, supervised learning gives great results but is not a realistic solution in the case of large scale image collections since the number of examples that need to be pre-classified becomes very large. Also, providing a fair sample of negative examples is very problematic. Fully unsupervised solutions do not work today, and therefore a middle ground has to be defined. Of course, in order to work with such high-level descriptors at a large scale, efficient data management is needed. In order to solve this second research challenge, the database community research challenge must therefore first be solved.

2.2 Database Tutorial

The database tutorial "Nearest Neighbor Search on Multimedia Indexing Structures" was presented by Thomas Seidl, professor of Computer Science at RWTH Aachen University, Germany.

Thomas Seidl described the prototypical multimedia queries, including similarity range queries and k-nearest neighbor queries. He then presented an overview of the main techniques proposed by the database community to efficiently process k-NN queries in various settings. This included direct k-NN search on various indexes, multi-step k-NN query processing for complex distance functions and methods for high-dimensional spaces.

What was clear, however, was that these techniques would not be satisfactory to address the first research challenge presented in the computer vision tutorial. This, of course, indicates a major research direction for the CVDB research community.

2.3 Technical Papers

The papers were organized into two sessions. The first session was geared more towards "techniques", while the second session was geared towards "applications".

In the "techniques" session, which was chaired by Shin'ichi Satoh, four papers addressed a wide range of topics from the computer vision and database areas. First, in [1], Cornacchia, van Ballegooij, and de Vries presented a study of how to implement applications involving multi-dimensional data sets on top of an RDBMS. Using several optimizations, they were able to match the performance of an application developed in Matlab. Then in [2], which was arguably the paper that best merged computer vision and database aspects, Lai, Goh and Chang focused on addressing the challenges of two scalability issues for active learning methods to deal with increasing dataset sizes and concept complexity. They presented remedies, explained limitations, and discussed future directions that such research might take. In [3], Singh et al. presented an initial framework for capturing and processing digital media-based information, based on the notion of "events". Their implementation specifically targets the problem of processing, storage, and querying of multimedia information related to indoor group-oriented activities such as meetings. Finally, in [4], Yamane et al. proposed that the similarity of images be evaluated using a measure of distance in a multi-vector feature space based on pseudo-Euclidean space and an oblique basis. Using this similarity measure, some of the loss of discriminability associated with quadratic-form distance measures is resolved.

In the "applications" session, which was chaired by Patrick Gros, four papers addressed a range of topics in the presentation and management of multimedia data. First, in [5], Albanese, Cesarano, and Picariello proposed a system to assist a user in browsing a digital collection by making recommendations. The system combines computer vision techniques and taxonomic classifications to measure the similarity between objects and takes into account previous user behavior. In [6], Bartolini, Ciaccia, and Patella presented another image browsing system, the personalizable image browsing engine (PIBE). The principal features of PIBE include the possibility of locally modifying the browsing structure by means of graphical personalization actions, and of persistently storing such customizations for subsequent browsing ses-

sions. In [7], Gosselin and Cord dealt with content-based image indexing and category retrieval in general databases. They compared seven classification strategies to evaluate the active learning contribution in CBIR. Finally, in [8], Moënne-Loccoz et al. considered the challenges of video document retrieval, which include balancing efficient content modeling and storage against fast access at various levels. They detailed the framework they have built to accommodate their developments in content-based multimedia retrieval.

2.4 Panel

The panel on "Future Applications and Solutions" was coordinated by M. Tamer Özsu, professor of Computer Science at the University of Waterloo. Other panelists were Jean Carrive of INA, Sébastien Gilles of LTU Tech. and Izabela Grasland of Thomson R&D France, as well as the two tutorialists. The goal of the panel was to be a forum for exchanging ideas on the applications of image and video data, and to allow the panelists to clearly describe what kind of tools they would need to facilitate the management of their large volumes of multimedia data.

Tamer Özsu opened the panel. His presentation reiterated some of the challenges mentioned by Roger Mohr in his tutorial. For Tamer Özsu, one of the primary challenges of the future is to obtain meaningful semantics from images and to represent and exploit those semantics in a smart way, both in terms of meaningful applications and appropriate database support.

Overall, for the invited industrial panelists, three main issues were fostering the panel. According to Sébastien Gilles, the first issue is the scale of real life systems dealing with multimedia data. Traditional O/RDBMSs scale very well, but the performance of the plug-ins offering multimedia data management facilities provided by vendors does not scale as well, making them inappropriate for dealing with large multimedia indexing tasks. Another aspect of scale for real systems is the requirement for deployment over a distributed and clustered architecture. In this case, performing (re)indexing or classification tasks, while maintaining the overall quality of service, is challenging. Finally, real system are alive, which means that the data they store evolves and therefore non-static database solutions are needed. Dealing with dynamic data is a twofold problem: first, data might be inserted and/or deleted from the database and the indexing structure must be updated accordingly – most state of the art schemes can not do this – and second, the description of data also evolves through time and, therefore, being able to query a database where images are described according to various description schemes seems mandatory.

Dealing with real data and with data sets of realistic sizes poses another set of issues, which were described by Jean Carrive. The most obvious ones are linked to performance since exploiting data (such as data streamed on TV) must be fast enough to absorb the huge volumes that are broadcast and accurate enough that the data can be later exploited for business purposes. For example, analyzing a real news program presents several challenging tasks for computer vision researchers such as cut detection, motion estimation, face recognition, noise segmentation, etc. Individual solutions already exist, each providing a good analysis, but merging them in a software suite is also very challenging and raises many issues. For example, the total cost of analyzing a media stream must stay below its delivery rate. Also, one has to face the potential contradictions between modules: a module analyzing the soundtrack of a sport event might detect a goal while another module doing motion analysis might output a break in the game.

Last, Izabela Grasland highlighted the mismatch between the way computer scientists assess the strength of their solutions and the satisfaction of end users. While response time, number of I/Os, precision, and recall are nice metrics, they poorly match non-professional users' expectations. In addition, interfaces matter much and it is clear that computer vision and database researchers not only have to start working with each other but must also start working with researchers that specifically work on human-computer interaction. Seamless integration of multiple display devices, ways to query and/or browse large collections of images, ways to effectively keep track of images (how can anyone deal with hundreds of folders, each containing thousands of images?), and simultaneously using keywords and visual similarities are challenging issues.

Several other issues were raised in the ensuing one hour discussions, including how database research can feed into computer vision research, the potential differences in the requirements of various alternative application domains (e.g., medical images, hyperspectral images, videos, etc.) and the importance of joint exploitation of multiple media, such as video images, sound and text.

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. The first conclusion that can be drawn from the workshop is that there is great need for this forum for interaction between the computer vision and database communities. In this first workshop of the CVDB series, most papers addressed either mostly "CV" aspects or mostly "DB" aspects. This was to be expected, as the goal of the workshop

is to facilitate the interaction of these two disjoint research communities. We anticipate that in the next CVDB workshop, the papers will be more focused on combining computer vision and database aspects. Based on the discussions during the workshop, there is certainly no shortage of interesting research directions, such as retrieval performance, semantics and learning, new and interesting application domains, and joint exploitation of multiple media.

It seems that the database community has not been working with computer vision researchers to a sufficient extent. As a result, the computer vision community has not accepted the techniques proposed by the database community. Database researchers have to work with computer vision experts in order to know what support these experts need to have, for which descriptors, with which constraints and for which applications; doing this is very important to be sure that the appropriate problems are being addressed. For example, since the state of the art in computer vision has shifted away from color histograms and other global image descriptors, developing efficient search algorithms for more advanced description schemes would be of primary importance for computer vision people. Working with computer vision researchers would also allow the database community access to realistic image collections, both in terms of contents and size, as well as techniques to assess the quality of the retrieval process, which is particularly important when working with approximate search algorithms.

The interaction between the computer vision and database communities, however, must be a two-way street and therefore the computer vision researchers also need to start looking towards the database community. Most of the descriptor schemes developed have never been evaluated against very large datasets because of lack of database support. Therefore, it is not clear in many cases that the efficiency of the retrieval process will scale sufficiently well for large collections, for example due to the complexity of the distance calculations. More importantly, however, it is not clear that the effectiveness of the retrieval will scale either, as the recognition power of the description schemes may dissipate when dealing with ever larger collections. It is clear that database techniques are required to enable computer vision researchers to work with collections of meaningful sizes.

4 CVDB 2005

The atmosphere of the workshop was very cordial and we expect the participants to start interfacing the two research areas. Based on the observed need for a forum for exchanging ideas and results that are at the intersection of the computer vision and database research areas, we have decided to make the CVDB workshop an annual event and will apply again for co-location with SIGMOD/PODS in Baltimore, Maryland in June 2005. We welcome any suggestions for ideas to address in CVDB 2005, as well as offers to participate in the work, for example by joining the program committee. Such suggestions can be sent via e-mail to cvdb@irisa.fr. We look greatly forward to this next edition of CVDB and we hope that it will be a successful event.

Acknowledgements

We would like to express our great appreciation for the support provided by:

- Canon Research Centre France;
- INRIA;
- Reykjavík University; and
- The French Ministry for Education and Research.

References

[1] Roberto Cornacchia, Alex van Ballegooij, and Arjen P. de Vries. A case study on array query optimisation. In *CVDB 2004*, June 2004.

[2] Wei-Cheng Lai, Kingshy Goh, and Edward Y. Chang. On scalabilty of active learning for formulating query concepts. In *CVDB 2004*, June 2004.

[3] Rahul Singh, Zhao Li, Pilho Kim, Derik Pack, and Ramesh Jain. Event-based modeling and processing of digital media. In *CVDB 2004*, June 2004.

[4] Yasuo Yamane, Tadashi Hoshiai, Hiroshi Tsuda, Kaoru Katayama, Manabu Ohta, and Hiroshi Ishikawa. Multi-vector feature space based on pseudo-euclidean space and oblique basis for similarity searches of images. In *CVDB 2004*, June 2004.

[5] Massimiliano Albanese, Carmine Cesarano, and Antonio Picariello. A multimedia data base browsing system. In *CVDB 2004*, June 2004.

[6] Ilaria Bartolini, Paolo Ciaccia, and Marco Patella. The PIBE personalizable image browsing engine. In *CVDB 2004*, June 2004.

[7] Philippe H. Gosselin and Matthieu Cord. A comparison of active classification methods for content-based image retrieval. In *CVDB 2004*, June 2004.

[8] Nicolas Moënne-Loccoz, Bruno Janvier, Stéphane Marchand-Maillet, and Éric Bruno. Managing video collections at large. In *CVDB 2004*, June 2004.