Cees Snoek
University of Amsterdam

Where Is the User in Multimedia Retrieval?

Marcel Worring
University of Amsterdam, The Netherlands

Paul Sajda
Columbia University

Simone Santini
Universidad Autonoma de Madrid, Spain

David A. Shamma
Yahoo! Research

Alan F. Smeaton
Dublin City University, Ireland

Qiang Yang
Huawei Noah's Ark Lab, Hong Kong

Editor's Note

This article summarizes a recent panel discussion at the ACM International Conference on Multimedia Retrieval, where a case was made for making the interacting user a first-class citizen again in multimedia retrieval research.

Compared to information retrieval or computer vision, multimedia retrieval is a relatively young discipline. Many people mark the 1992 Visual Information Management Workshop[1] as the beginning of the field. It was there that researchers recognized the need to consider multimedia data, in particular visual information, as a new type of item that could appear in a digital collection. Although the number of items at that time was still small, typically in the thousands, it was orders of magnitude larger than the tens of images that computer vision research was addressing. From that first workshop, we have this important quote: ''Computer vision researchers will have to identify features required for interactive image understanding, rather than their discipline's current emphasis on automatic techniques, and develop techniques to compute these features in interactive environments.'' The information retrieval field agreed that new techniques were necessary to cope with the specifics of visual data. The notion of visual words, now so popular in visual retrieval, was unheard of at that time. Thus, a new research area was born.

The IBM Query by Image Content (QBIC) system was released not long after the 1992 workshop. QBIC was an early example of a query-by-pictorial-example system, where the user selected example images or specified the required images. New possibilities for museum collections and medical imaging arose, but techniques were not yet mature enough to have much impact. Various research efforts started to improve features, especially in terms of their invariance to various conditions. This early period in content-based retrieval was mainly successful in bridging the sensory gap.[3]

At the end of that period (around 2000), two new conference series started. The ACM Conference on Multimedia Information Retrieval (MIR), which began at the University of Illinois, originally focused on computer vision applications and held its 11th meeting in 2010. The first ACM Conference on Information and Video Retrieval (CIVR) in 2001 had a strong connection to library sciences, which hosts a community of archivists who label data at insertion time and search for images on request. Until 2007, these conferences always included a few nontechnical papers every year that looked at the physical retrieval techniques (such as labeling). The first VideOlympics,[4] where interactive systems were demonstrated in front of a live audience of scientists and media librarians, was also held in 2007 at the Sound and Vision Archive in The Netherlands.

After 2007, these conferences shifted their focus toward the computational side of the problem, with a stronger emphasis on industrial applications. We can largely attribute this shift to the important role of TRECVID.[2] Suddenly, there was easy access to datasets and clearly defined tasks and metrics. Hence, the field had a common goal to pursue and a benchmark to use. Interactive tasks were defined for TRECVID, but the concept-detection task especially flourished. It gave a boost not only in the multimedia retrieval field, but the computer vision community started to embrace the topic as well.

In the early days, we could do what we wanted because we were alone and could easily cater to the task at hand. Then Internet came

1070-986X/12/$31.00 © 2012 IEEE

Published by the IEEE Computer Society

along, and everything changed. Multimedia data has penetrated all of society. Suddenly everything is shared, nothing is owned, and there's no privacy. Such widespread sharing has created new user needs and opportunities, and now the infrastructure necessary to develop new applications is within reach. It might be time to rethink what we are doing in the field.

In 2010, the MIR and CIVR conferences merged into the ACM International Conference on Multimedia Retrieval (ICMR) series. At the 2012 second annual ICMR in Hong Kong, a panel took place with Marcel Worring as the moderator and the other authors of this article as the panelists (see Figure 1). In this panel session, and with active participation of the audience, we explored this intriguing question: Where is the user in multimedia retrieval? The lively panel session also discussed other questions central to the field's development: Do the Internet and the pervasive presence of multimedia lead to new types of users and new avenues for multimedia retrieval? Are we catering sufficiently to the new tasks that come with it?

Figure 1. The panelists, moderator, and scribe of the ICMR 2012 panel. From left to right, Daan Vreeswijk (scribe), Marcel Worring (moderator), Paul Sajda, David Shamma, Alan Smeaton, Simone Santini, and Qiang Yang.

Figure 2. Wordle, created using Worldle.net, based on the titles of the ICMR 2012 papers. This shows that few papers focused on user-centered research.

The Dominance of MAP

Taking a broad perspective, multimedia retrieval can be decomposed into two major problems:

- What does the user want?
- How do we answer a user query?

The original motivation for QBIC and the work of librarians focused on the first perspective, looking for methods to provide optimal user query support. But things have changed rapidly. Now almost all multimedia retrieval methods are data-driven, and many solutions are developed without having an explicitly defined real-world problem to solve. Performance measures focus on the quality of how we answer a query. For example, Figure 2 shows a Wordle, created using Worldle.net, using the title of ICMR 2012 papers.

Mean average precision (MAP), most prominently advocated by TRECVID, is a measure that integrates precision and recall. A nice and simple value to benchmark different methodologies, MAP has become the primary target for many researchers. That is, instead of considering whether methods satisfy a specific user task, MAP has become a target of its own. The papers that the community considers ''good'' are those with performance results 1 to 5 percent better in terms of MAP for one of the benchmark tasks. Most commonly, researchers obtain this improvement by making a small variation to the standard pipeline for performing the task. Having better data descriptors as starting point for answering queries is important, but MAP is purely data-driven and only one way of measuring performance. It is certainly weird to use MAP alone when talking about users employing multimedia retrieval to solve their search problems, especially in a connected world with so many ways to use data.

So why is the field so focused on answering queries and, in many cases, only looking at

Users have various tasks to perform, and no one multimedia retrieval model fits all approaches.

improving data descriptors? The major reason we see is that the scientific community rewards reproducible results. We have lost the librarian's perspective because it is so much easier to publish papers about improving a standard task than it is to describe a new insight about user intention or a new interface for browsing results. For example, a new methodology for recommending videos based on their content evaluated on standard YouTube categories, with increased precision and recall, might yield a nice publication. But will users use such an option? We have no clue whether they intend to rely on recommendations in the first place. Still, a paper providing insight on the latter topic would be difficult to publish because it is only descriptive and not easily reproduced.

As a consequence of MAP's dominance, the field has shifted its focus too much toward answering a query. Thus, we are now stuck with a default search model. It is time to go back to understanding queries and answering them on equal terms.

Users and Tasks

Although users sometimes search for multimedia content, it is rarely their ultimate goal. For example, a system might help doctors retrieve images with similar patterns, but such images are only one ingredient doctors use to reach a diagnosis. In addition, users' goals can vary greatly. The medical scenario certainly differs from someone wanting to share holiday pictures or a DJ looking for new material. None of these scenarios fit a simple keyword-ranked results list approach. While searching, users can decide to quit, narrow or broaden their scope, explore new and possibly unrelated directions, or share their data with peers. Can we ever expect one search model to cater to all these types of tasks?

One problem is that users often don't even know what they want from an automatic system. Consider for example a video query system that ''senses'' what you want and just gives you a video to watch. Would you be happy if it's off 20 percent of the time? Probably not. However, if you were choosing yourself, you would likely also be 20 percent wrong but still be happy. So users want more from automated systems than from themselves. Even if the system is 100 percent right, you might still be less satisfied than if you chose yourself and were wrong because there is joy in choosing. Users want a query engine to behave like a concierge; it should give advice when asked and proactively offer suggestions, but never overdo it. Realizing such an information concierge with current multimedia retrieval methodologies, however, is difficult.

Another characteristic of users and their tasks is that they are dynamic. We cannot assume that user needs and characteristics will remain static over time. Thus, we cannot simply provide them with a system once. In fact, by giving them a tool, users will change their behavior. For example, think about the changes that occurred when people started taking pictures with their mobile phones.

So, users have various tasks to perform, and no one multimedia retrieval model fits all approaches. Thus, a better understanding of what users actually want and do when using multimedia retrieval is needed.

Understanding the User

The Web has introduced many new elements into multimedia search and retrieval. We now have access to each other's media, which we can easily use, annotate, and repurpose. As a consequence, we see highly enriched media on which users and their communities have a continuously changing perspective, building upon previous experiences. Yet, the model underlying common multimedia retrieval systems is stateless. When we build a data-driven search system, we don't know what users want, what they do with it, what will be useful, what visual concepts they will use, and so forth. This is a challenge even when catering to professional users, who have specific tasks to perform and who know more or less what they are looking for. Casual users visit social networks for fun, so the problem is more complex and nothing is fixed.

The best way to understand users is to talk to them and carefully listen to how they search in

practice. Let's consider the DJ trying to find information about a new artist. Will he search the Web with technical and objective terms such as ''140 bpm deep house''? No, he will get in touch with his network of people working in the music industry. His multimedia search is a multilayered system, where real content only plays a role at the lower levels. The higher levels are his starting points for search, and these are all related to the data's context instead of the content itself.

There is a known tension between content and context. We have been working on content (pixels, samples) trying to understand it better and mostly ignored context. But what does content tell us about the user? Over the last few years, we have done a great job of narrowing the semantic gap,[3] but only from the data side. Previous research observed that the interpretation of the content depends on the user in a given situation,[3] but recent research has not reflected this. Going forward, we need to sense mind, body, and place while keeping track of the user's earlier system interactions. In this view, content becomes just one of the many information channels at our disposal when answering queries. All the GPS sensors, orientation sensors, sharing activity, and previous use supply additional information channels. For example, we might even use an iPad or mobile phone camera to capture users while they pose a query. In many cases, context might be more decisive than the content.

A new and promising way of sensing users is to look at brain responses when users are accessing multimedia, but correct measurements are a major problem. The simple act of asking people to do something influences their behavior, so we need more unobtrusive ways to measure responses. A general view of user actions might help us measure synchronous responses to stimuli, such as group response to movies. That only gives us information about an average user, however, and we need individualized data. Understanding the user via brain responses has great potential because this could reveal user intention, but this still has a long way to go before it will give us a clear picture of how we should support each and every search task of a user.

By obtaining more information about a query's context through these various information channels, the better chance we have of providing the right answer. An important step in understanding user behavior is properly integrating all these information channels—clearly a multimedia problem. Furthermore, we should consider how to present all these information channels to the user in a way that best helps answer the query while at the same time helping the system to understand the user. We shouldn't expect, however, that we can do this alone. Other fields, such as cognitive and social science and human-computer interaction will need to contribute; multidisciplinary teams are the only way to really understand both the data and the users and to provide the system support to bring them together.

The field needs more descriptive application- and interaction-oriented papers and, most importantly, the acknowledgement that these papers are useful.

The Research Community

In the multimedia retrieval community, the emphasis has moved toward quantitative results to such an extent that the user has moved into the background. The multimedia retrieval community and end users alike will benefit when this trend is reversed and the user returns as a first-class citizen. But this requires someone to stand up and shout. PhD students cannot change this trend because it is too risky to go for them to forge alternative paths. The problem is that researchers just respond to the stimuli they are given. If quantitative results are what get published, this is what you do as researcher. There is clearly a cultural problem. You're trying to build something that works, but you're required to ask and solve fundamental questions. Hence, the field needs more descriptive application- and interaction-oriented papers and, most importantly, the acknowledgement that these papers are useful.

Thus, the burden lies in the hands of the program chairs and reviewers who ultimately decide what is publishable. It is their task to judge the relative importance of (minor) improvements in MAP for standard tasks versus innovative techniques and ideas.

One step in this direction is to upgrade the role of demos because high-quality demos can form the bridge between users and systems. At ICMR 2012, for example, demo papers counted as papers and were part of the proceedings. Still, the demo papers were only two pages and thus considered inferior to regular contributions. We believe that when demos are truly innovative they should get the credit they deserve. In the long run, we can then expect more and better demo submissions.

Conclusion

What started as a field with an emphasis on optimally serving users' interactive information needs has now become dominated by methods that focus on improving the MAP of a clearly defined task disconnected from its application. With the pervasiveness of the Internet and all the sensors available to derive contextual user information, it is time to bring the data and the user back together. As a field, we must consider understanding the subjective and descriptive nature of users and understanding data as equally interesting research topics that are both worthy of publication. **MM**

Acknowledgments

We thank the ICMR 2012 conference organizers Horace Ip, Yong Rui, and Chong-Wah Ngo for giving us the opportunity to have this panel in Hong Kong. We further thank the audience for their active participation. Last, but certainly not least, we thank Daan Vreeswijk for carefully taking notes during the panel; his work formed the basis for this article.

References

1. R. Jain, ''NSF Workshop on Visual Information Management Systems: Workshop Report,'' *Proc. Storage and Retrieval for Image and Video Databases* (SPIE), 1993, pp. 198–218.
2. C. Thornley et al., ''The Scholarly Impact of TRECVid (2003–2009),'' *J. Am. Soc. Information Science and Technology* (JASIST), vol. 62, no. 4, 2011, pp. 613–627.
3. A.W.M. Smeulders et al., ''Content Based Image Retrieval at the End of the Early Years,'' *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, 2000, pp. 1349–1380.
4. C.G.M. Snoek et al., ''VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems,'' *IEEE Multimedia,* vol. 15, no. 1, 2008, pp. 86–91.

Marcel Worring is an associate professor at the University of Amsterdam. His research interests are in multimedia analytics, bringing together multimedia analysis, interaction, and information visualization. Contact him at m.worring@uva.nl.

Paul Sajda is a professor of biomedical engineering at Columbia University. His research interests include the neural basis of rapid decision making and hybrid brain-machine interfaces for interactive media annotation and retrieval. Contact him at psajda@columbia.edu.

Simone Santini is a professor at the Universidad Autónoma de Madrid, Spain, where he works on context models for user-aware access to multimedia data, diversity and novelty in multimedia information retrieval and formal languages for multimedia and time data retrieval. Contact him at simone.santini@uam.es.

David A. Shamma heads the Human Computer Interaction research group at Yahoo! Labs. He investigates how people interact, engage, and share media experiences both online and in the world. He is also the coeditor of *Arts and Digital Culture* for ACM's Special Interest Group on MultiMedia. Shamma has a PhD from Northwestern University. Contact him at aymans@acm.org.

Alan F. Smeaton is a professor of computing and deputy director of CLARITY: Centre for Sensor Web Technologies at Dublin City University. He is the cofounder of the annual TRECVid evaluation benchmarking campaign and his interests are in multimedia information retrieval. Contact him at alan.smeaton@dcu.ie.

Qiang Yang is the head of Huawei Noah's Ark Research Lab and a professor at the Hong Kong University of Science and Technology. His research interests include data mining and artificial intelligence. Contact him at qyang@cse.ust.hk.