John R. Smith
IBM

VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems

Cees G. M. Snoek, Marcel Worring, Ork de Rooij, and Koen E. A. van de Sande
University of Amsterdam

Rong Yan
IBM T.J. Watson Research Center

Alexander G. Hauptmann
Carnegie Mellon University

Interactive prototypes are often the best way to convince an audience of a new multimedia technology's possible impact. Because of its dynamic audiovisual nature, a multimedia application demonstration communicates applied science more effectively than a static description in a journal publication would. Ideally, a multimedia demonstrator grasps the audience's attention by presenting effective results, advanced multimodal interfaces, novel means of user interaction, or combinations thereof. An interactive demo offers researchers a means to engage their audience in a way that they could never achieve with a written manuscript alone.

All major multimedia conferences have adopted demo sessions where researchers, equipped with their laptop-installed system, show their demo to the conference attendees on a one-on-one basis. While effective, these demo sessions lack a common denominator. Individual systems provide solutions for different tasks on different data sets, and the conference attendees have to become familiar with the peculiarities of each individual demo. Moreover, a one-on-one demo session makes it infeasible to present a system to the entire conference audience. In addition, due to their lack of focus, the impact of demo sessions on the audience is suboptimal.

Editor's Note

Video search is an experience for the senses. As a result, traditional information retrieval metrics can't fully measure the quality of a video search system. To provide a more interactive assessment of today's video search engines, the authors have organized the VideOlympics as a real-time evaluation showcase where systems compete to answer specific video searches in front of a live audience. At VideOlympics, seeing and hearing is believing.

—John R. Smith

Establishing focus

A good way to establish focus is letting several demo systems solve the same problem on a similar data set. A notable example of such a focused effort takes place at the demo session at the National Institute for Standards and Technology's (NIST's) Trecvid workshop (see the "Trecvid Interactive Video Retrieval Task" sidebar).[1] Trecvid promotes progress in video retrieval by providing a large video collection, common retrieval tasks, uniform evaluation procedures, and a forum for researchers interested in comparing their results. In the months preceding the workshop, researchers work on a common retrieval problem. Results are submitted offline and then presented to the benchmark participants during the workshop.

In the Trecvid demo session, participants showcase their video-retrieval systems. Because the audience is knowledgeable in the problem area and the task at hand, the result is a lively demo session where the audience gains deep insight in various video-retrieval systems. Unfortunately, the audience at this event is limited to researchers who participate in the evaluation campaign. Of course, the video-retrieval systems are shown at regular demo sessions, but they're never exposed simultaneously to an audience. A common data set and a common task aid in establishing a focused demo session, but these aren't sufficient to engage an uninformed audience.

Involving the audience

Apart from having systems compete simultaneously on a common task, and having all systems solve the tasks on the same data set, audience involvement can be achieved by communicating overall results in real time. This method allows for on-the-spot performance comparison of different prototype

1070-986X/08/$25.00 © 2008 IEEE

Published by the IEEE Computer Society

Trecvid Interactive Video Retrieval Task

In 2001 the American National Institute for Standards and Technology (NIST) extended its successful Text Retrieval Conference (TREC) series[1] with a track focusing on automatic segmentation, indexing, and content-based retrieval of digital video. With a steady increase in both the size of the video archive analyzed—from 11 hours in 2001 up to 400 hours in the current cycle running until 2009—and the international participants—from 12 in 2001 up to 54 in 2007—this track became an independent evaluation workshop known as Trecvid in 2003.[2]

Trecvid promotes progress in the field of video retrieval by providing a large video collection, uniform evaluation procedures, and a forum for researchers interested in comparing their results. Already, the benchmark is making a huge impact on the multimedia community, resulting in a large number of video-retrieval systems and publications that report on the experiments performed within Trecvid.

One of Trecvid's core tasks is interactive video search,[3,4] which retrieves from a video archive, presegmented into *n* unique shots, the best possible answer set in response to a visual information need. On the basis of this need, a user starts an interactive search session with a video retrieval engine. After inspection on the results obtained, a user can rephrase queries; aiming at retrieval of more and more accurate results.

To limit the amount of user interaction and to measure search system efficiency, all individual search topics are bounded by a 15-minute time limit. Each year, Trecvid provides about 24 topics. For each individual topic, participants can submit a maximum of 1,000 final results, ranked according to the highest possibility of topic presence. At Trecvid, human assessors from NIST inspect the results and compute a common performance metric known as *average precision*.[1] The results are reported at the workshop in the form of a score sheet. Figure A illustrates this process.

Trecvid's emphasis on retrieval performance is not without criticism.[5] Because there is a human in the loop, interactive retrieval results depend on the user interface and searcher expertise. However, these important factors are not evaluated directly in the benchmark.

References

1. E.M. Voorhees and D.K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press, 2005.
2. A. Smeaton, P. Over, and W. Kraaij, ''Evaluation Campaigns and Trecvid,'' *Proc. ACM Int'l Workshop Multimedia Information Retrieval*, ACM Press, 2006, pp. 321-330.
3. A. Hauptmann and M. Christel, ''Successful Approaches in the TREC Video Retrieval Evaluations,'' *Proc. ACM Multimedia*, ACM Press, 2004, pp. 668-675.
4. C.G.M. Snoek et al., ''A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval,'' *IEEE Trans. Multimedia*, vol. 9, no. 2, 2007, pp. 280-292.
5. T.-S. Chua, ''Towards the Next Plateau—Innovative Multimedia Research Beyond Trecvid,'' *Proc. ACM Multimedia*, ACM Press, 2007, p. 1054.

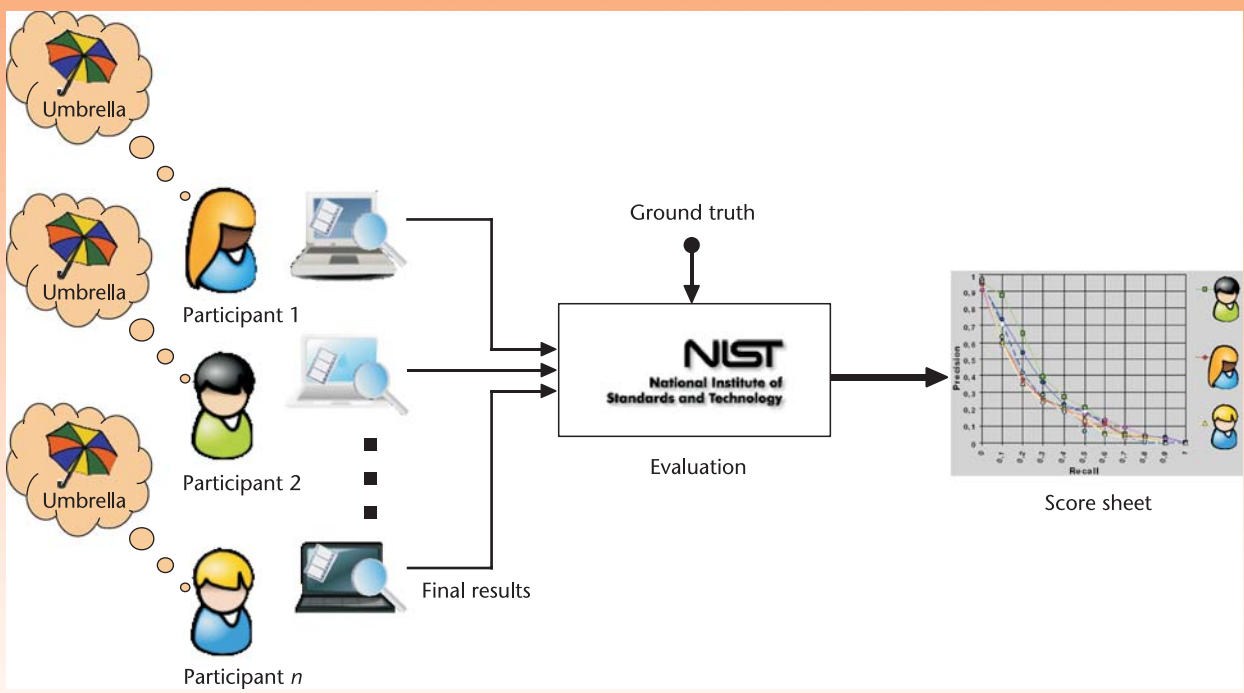

Figure A. Trecvid's interactive search procedure.

systems as well as evaluation of interfaces and ease of use. Hence, the demo session audience receives a good perspective on the possibilities and limitations of current multimedia systems.

In this article, we describe our experience in organizing the VideOlympics showcase, a demo session held at the 2007 ACM International Conference on Image and Video Retrieval that facilitates real-time evaluation of video search engines. We highlight the real-time evaluation infrastructure, the VideOlympics showcase implementation, and show how the demo session involved the audience for maximum impact.

VideOlympics showcase

There are many evaluation campaigns for video retrieval,[1] all requiring substantial effort in conceiving novel algorithms for content-based analysis, processing of the multimedia data, and building the search engine. To prevent potential participants from being discouraged by the amount of work involved, our goal for the VideOlympics was that participation should have a minimal impact on existing video search engines. This is why we built the VideOlympics on a popular existing evaluation campaign: Trecvid's interactive search task.

In similar spirit as the Trecvid benchmark, our major aim for the VideOlympics is to promote research in video-retrieval research. An additional goal is giving the audience a perspective on the possibilities and limitations of state-of-the-art systems. Where traditional evaluation campaigns like Trecvid focus primarily on the effectiveness of collected retrieval results, the VideOlympics take into account the influence of interaction mechanisms and the advanced visualizations in the interface. To prevent the exclusion of interesting ideas due to a participant's fear of losing—for example, because companies don't want to be associated with an inferior product, or because smaller research teams might feel they can't compete against the resources of the larger groups—we don't strictly score performance. Moreover, the VideOlympics should be fun for the participants and the conference audience. For all these reasons, we make sure that the VideOlympics only has winners.

To make the transition from a regular evaluation campaign to its real-time equivalent, participants must communicate their results simultaneously. For this purpose a client–server architecture seems an appropriate choice. In this architecture, each participating system in the evaluation forms a client that communicates independently to an evaluation server. The evaluation server processes incoming results, prioritizes them using a time stamp, compares them to the ground truth, and updates a score related to the task evaluated. This has the added advantage of instantly communicating overall results to the audience. Figure 1 visualizes the resulting infrastructure.

Video data

To avoid the problem of securing and distributing a substantially large video data set, we adopt the test set of the Trecvid 2006 benchmark. The Trecvid 2006 test set contains about 160 hours of Arabic, Chinese, and English broadcast news, recorded in November 2005. The Fraunhofer Institute provided a camera shot segmentation for this video archive, yielding a total of 79,484 unique shots. For each individual shot, Dublin City University extracted keyframes. In addition to these visual analysis results, a US government contractor made available speech recognition and English machine translations. Each participating video-retrieval system used the common video data in combination with the common shot segmentation. Participants could choose what kind of additional analysis to include in their system.

Participating video search engines

A requirement for participation was that all video search engines should work on a laptop, with all relevant Trecvid 2006 test data on board. To ensure transparency to the audience, participants weren't allowed to use external information sources, such as online information extraction. Table 1 gives an overview of the nine teams that participated in the VideOlympics. Some systems emphasize advanced multimedia analysis techniques, such as near-duplicate detection and multimodal fusion, typically in combination with a modern Web browser or a traditional storyboard interface. Other systems emphasize advanced visualizations of the result set. Almost all participating systems exploit large lexicons of visual concept detectors for retrieval. Taken

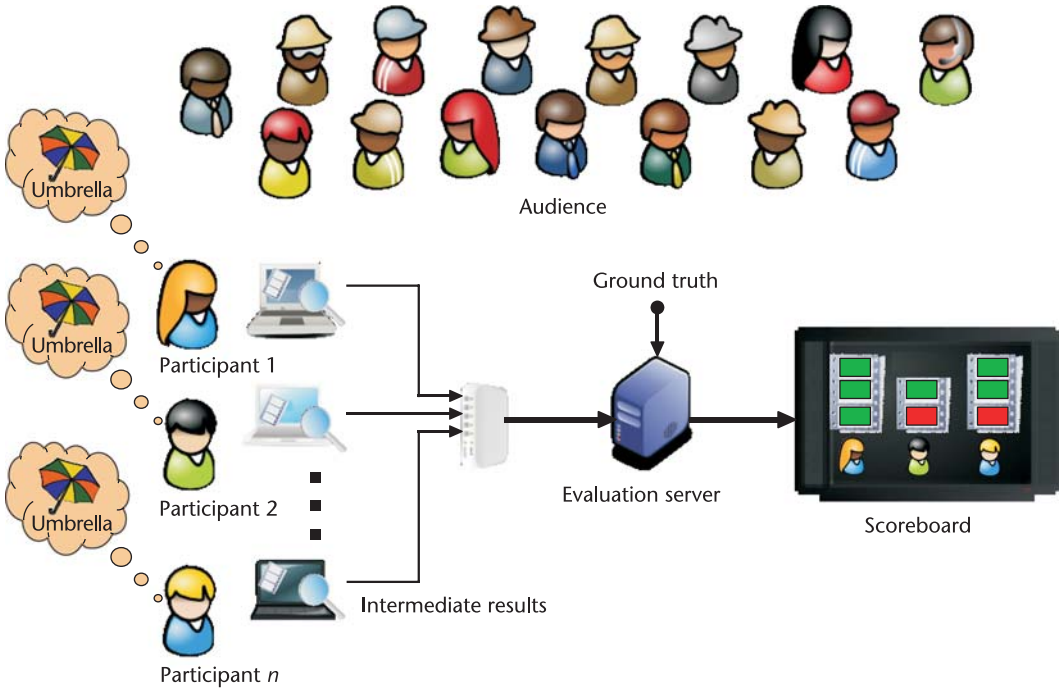

Figure 1. Infrastructure for real-time evaluation of multimedia retrieval systems. Individual participants are interconnected to an evaluation server. The evaluation server receives intermediate results, compares them to a ground truth, and displays results on a scoreboard to the audience in real time.

together the various video search engines represent a wide range of current technology.

Search topics

Apart from the requirement that all search engines operate on the same video archive, participants must solve the same problems. These problems are defined as topics that mimic visual information needs. NIST provided the text-only search topics, similar to those used in previous Trecvid interactive search tasks; namely, find shots of

- a person filling a vehicle with fuel,
- one or more soccer goalposts,
- one or more helicopters in flight,
- an outdoor night scene with people,
- at least one person and at least ten books,
- a bridge over visible water,
- at least one adult person and at least one child,
- US President George W. Bush, Jr. walking;

These topics were unknown to the participants until they were revealed during the VideOlympics showcase.

Table 1. Participating multimedia retrieval systems in the VideOlympics showcase.

Video search engine	Research institute
CuVid	Columbia University, US
MediaMagic	FX Palo Alto Laboratory, US
Marvel	IBM Research, US
Informedia	Carnegie Mellon University, US
ITI	Centre for Research & Technology Hellas, Greece
MediaMill	University of Amsterdam, Netherlands
NII-SEVIS	University of Tokyo, Japan
SmartV	Tsinghua University, China
VisionGo	Chinese Academy of Sciences, China, and National University of Singapore, Singapore

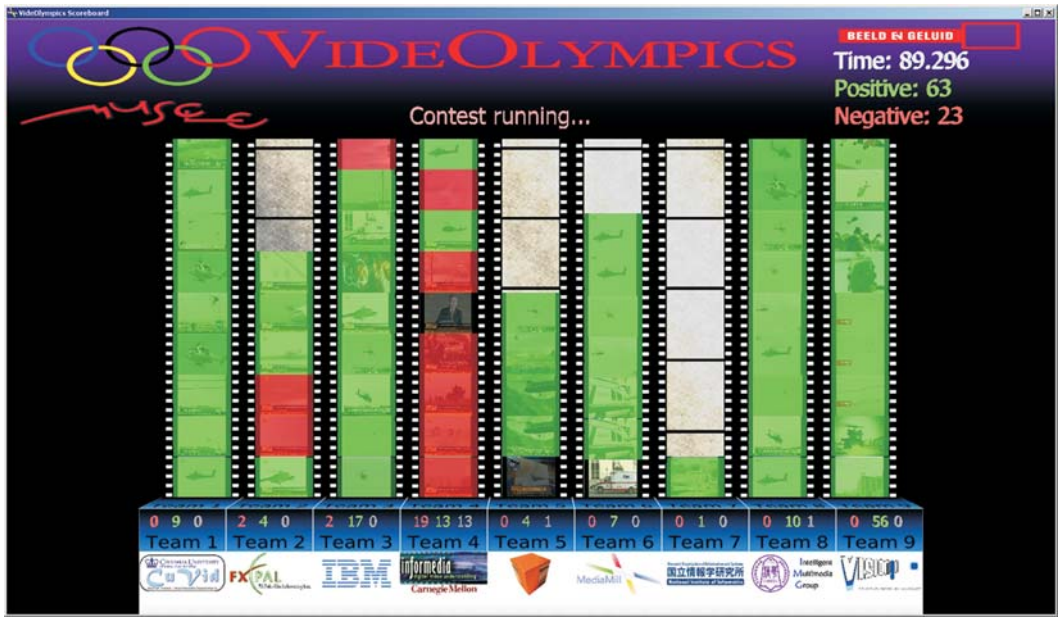

Figure 2. Scoreboard used at the VideOlympics. The audience gets immediate visual feedback on whether retrieved shots are correct (green), wrong (red), or unknown (gray).

Evaluation server

In contrast to Trecvid, where overall results are submitted at the end of the search session, VideOlympics participants submit results immediately to the evaluation server over a 100-Mbit dedicated Ethernet hub. This encouraged quick retrieval of relevant results as well as unique results not found by others. After retrieving a relevant shot, participants submit a unique team identifier in combination with a unique shot identifier using a single HTTP GET call. In the interest of the evaluation, participants couldn't correct submitted shots, nor could they reorder submitted shots in a different sequence. Once a shot was submitted, it was final. To limit the amount of user interaction and to assure the efficiency of search systems, participants had a five-minute limit on all individual search topics.

The evaluation server compared submitted shots with the predetermined ground truth and displayed it on the scoreboard in combination with an overlayed color indicating whether the result was correct. To allow for swift visualization of submitted results, all keyframes from the video data were loaded into memory and displayed as soon as they were retrieved. Figure 2 shows the scoreboard's interface.

Showtime

The VideOlympics was hosted in the futuristic building of *Beeld en Geluid*, the broadcast video archive of the Netherlands. By situating all participants in the building's auditorium, we offered the audience the opportunity to see all the searchers in action. To give the audience maximum visibility, we provided participants with a flat screen that they could connect to their laptop. Three 32-inch LCD television screens displayed the scoreboard to the audience. To engage the audience even further, a ringmaster, Alex Hauptmann, explained the procedure, introduced the topics, and actively commented on the results. Finally, he had the audience vote for their personal favorite video search engine. In the spirit of the Olympic Games, participation was more important than winning, so the results were not meant for publication. Winners received Golden Retriever awards; we also awarded Golden Retrievers to less exact categories, such as

- most impressive interface,
- public favorite, and
- most easy to use by the audience.

Figure 3 depicts a collage of pictures taken during the event. A video trailer of the VideOlympics is available at http://www.videolympics.org.

Figure 3. Visual impression of the first VideOlympics at the Dutch broadcast video archive. Note the involvement of the audience.

Conclusion

The first VideOlympics brings content-based analysis to the archive and allows for many-to-many communication between video search engines and their audience. It was a great success. The VideOlympics provided the excitement of a competition without the associated stress on the participants. For the first time, the audience was able to compare different multimedia retrieval systems on the same tasks and see how they performed with unrehearsed topics. Many audience members felt they understood the technology's capabilities after seeing it in live action and in several system variations. We will have another VideOlympics at the 2008 ACM International Conference on Image and Video Retrieval. While all participants will go home with a Golden Retriever award, the real winner will be the audience. **MM**

Acknowledgments

We thank the searchers and the audience for participation. We thank Johan Oomen and Claartje Kok for the local arrangements at *Beeld en Geluid*. We thank Paul Over from NIST for providing the search topics. Finally, we thank Fabchannel for producing the video trailer. The MultimediaN BSIK project and the Muscle European Network of Excellence sponsor the VideOlympics.

Reference

1. A. Smeaton, P. Over, and W. Kraaij, ''Evaluation Campaigns and Trecvid,'' *Proc. ACM Int'l Workshop Multimedia Information Retrieval*, ACM Press, 2006, pp. 321-330.

Contact author Cees G.M. Snoek at cgmsnoek@science.uva.nl.

Contact editor John R. Smith at jsmith@us.ibm.com.