The Community and the Crowd: Multimedia Benchmark Dataset Development

Martha Larson
Delft University of Technology, The Netherlands

Mohammad Soleymani
Imperial College London, UK

Maria Eskevich
Dublin City University, Ireland

Pavel Serdyukov
Yandex, Russia

Roeland Ordelman
University of Twente, The Netherlands

Gareth Jones
Dublin City University, Ireland

Benchmarking has helped significantly advance the state of the art in multimedia technologies. The key has been the concept of a benchmark task, which consists of a problem definition, a data collection, ground truth, and an evaluation metric. Over the years, techniques for designing tasks and developing datasets to evaluate multimedia indexing and retrieval algorithms have become highly sophisticated. However, we are currently witnessing rapid, further evolution in benchmarking that is driven by the increased ease with which we can collect and manipulate large amounts of multimedia data.

At the same time, two other important developments have made contributions:

- the benchmarking community, the group of researchers who work on common tasks and share the load of creating multimedia data collections, and
- the maturation of techniques for effectively exploiting crowdsourcing to generate ground truth for large multimedia data collections.

This article discusses how these two developments support new methods of identifying promising multimedia use cases and developing the corresponding tasks and datasets. We argue that a diverse benchmarking community makes it possible both to encourage researchers to focus on specific, highly promising tasks and to promote the creativity and risk-taking needed to advance the state of the art in multimedia technologies. With a specific example, we show that crowdsourcing is a viable method for developing multimedia ground truth.

Our insights are based on experiences with the MediaEval (www.multimediaeval.org) multimedia benchmark, a relatively young, but rapidly growing benchmarking initiative that focuses on human and social aspects of multimedia. MediaEval strives to emphasize the ''multi'' in multimedia, including the use of speech, audio, tags, users, context, and visual content. Because of this emphasis, MediaEval attracts a diverse group of researchers with a range of perspectives on multimedia research. MediaEval works by exploiting this diversity to drive innovation in task design and data-collection development. Here, we take the concrete example of Blip.tv10000, a large dataset of Internet video that was collected, divided, and annotated for use in multiple MediaEval benchmarking tasks. Two tasks that use this data collection, the Genre Tagging and Rich Speech Retrieval tasks, provide an interesting contrast with respect to their motivating use cases and the method applied for ground truth generation.

The MediaEval Multimedia Benchmark leveraged community cooperation and crowdsourcing to develop a large Internet video dataset for its Genre Tagging and Rich Speech Retrieval tasks.

Design Considerations

Generally, consensus exists in the multimedia community regarding important elements to

1070-986X/12/$31.00 © 2012 IEEE

Published by the IEEE Computer Society

consider when developing large video collections. For example, the Internet Archive Creative Commons (IACC.1) dataset used by Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVid), (http://trecvid.nist.gov) involved setting out design requirements for large-scale data sets: the collection should be large and widely used, and it should support real-world use cases.[1] These requirements are uncontroversial and widely accepted but need further interpretation before they can be applied to data-collection development in practice. Using the MediaEval benchmark as an example, we discuss particular considerations that have emerged in our work.

Data collections need to be large to ensure that researchers are working on the large-scale problem and are targeting scalable solutions. However, in reality, a data collection will always be a subset of the totality of data available on the Internet.[1] The social emphasis of MediaEval makes it important to take certain global aspects of the sample into consideration. In particular, sparsely sampled data can have profound effects on the overall properties of a social graph.[2] An online multimedia community's social graph comprises not only explicit links such as friendships, but also implicit links such as comments. When social aspects of multimedia are involved, a large data collection is not useful for research unless it has been designed with generalizability in mind.

Human search behavior should be also considered when deciding on the size and nature of the data collection necessary to tackle certain problems. For example, in the case of multimedia search, users often make a decision about where to search even before they begin. This pre-search decision is easy, well informed, and so natural for users that it often goes unnoticed. For example, users looking for family videos go to their social video-sharing sites. Users looking for user-contributed content capturing a recent event go to a site with real-time information, such as video-sharing and microblogging websites. In short, the user will often make a significant contribution to narrowing the search domain before actually beginning to search. This narrowing changes the multimedia search problem because it shifts the relationship between the relevant and nonrelevant multimedia items. The nature of the features that best discriminate between multimedia items is changed by this shift and the characteristics of algorithms that will perform optimally are also potentially affected. Data-collection design must account for the specific usage settings and not blindly make the assumption that bigger is better.

Data collections need to be long lived and widely used to guarantee that the research community exploits the full benefits of using a specific dataset to track the development of the state of the art. The ability to compare algorithm performance with respect to the same task across sites and over time is a widely cited benefit of benchmarking. A key consideration is creating datasets that are freely distributable, which has led to the increased use of Creative Commons content.

Nevertheless, the ease with which data can be collected from the Internet is a double-edged sword. Researchers enjoy the advantage of collecting large datasets, but the multimedia research community as a whole potentially suffers because it is easy for individual groups to strike out on their own and create their own datasets, limiting the possibilities for cross-site comparison of algorithms. MediaEval attempts to bring a balance between the ease of dataset development and the need to focus research efforts on a finite number of key datasets. It provides a forum within which groups of researchers can come together and concentrate their efforts on creating shared resources.

Finally, data collections need to support tasks that are based on real-world use scenarios and address real users' needs. There are two divergent views on the appropriate relationship between multimedia data collections and the underlying use cases. The first, the *technology-driven view*, is a general effort to create a reasonably representative dataset that will support the development of basic technologies. Once these technologies have been developed, users will decide what to do with them. The IACC.1 data collection follows this philosophy.[1] Exact use cases for the collection are unspecified because the idea is that this activity should be left to the users. In the second approach, the *usage-driven view*, datasets support the development of algorithms that address tasks already carried out by users in another form. The purpose of the algorithms is then to make these tasks easier, automatic.

These two views constitute a larger problem that faces the creation of large-scale multimedia collections. Users cannot use or even imagine

multimedia technologies that do not yet exist. On the other hand, it is difficult to know which technologies to develop without knowing what users are looking for. MediaEval attempts to break out of this cycle by using both views of the connection between tasks and use cases to develop tasks.

Benchmarking Methodologies and Standards

MediaEval adopts the benchmarking practices developed within the text-based information retrieval (IR) community—in particular by TREC and the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org). Text-based IR has a long history of benchmarking, and its impact has been shown to be significant for both textual[3] and multimedia content.[4] Here, we emphasize two important aspects of benchmarking practice that have been inherited from IR and are practiced by multimedia benchmarks including TRECVid[5] and ImageCLEF (www.imageclef.org). First, researchers have widely adopted the Cranfield paradigm for benchmarking practice.[6] This paradigm views task definition (including the set of query topics) and the corresponding ground truth (the target to be achieved on the task) to be an abstract representation of users and their information needs. This representation will, by nature, always be imperfect, so benchmark evaluation must always be considered a complement to and not a substitute for user trials carried out with an operational system in the field. Second, experience has proven the effectiveness of a yearly benchmarking cycle, which has the advantage of coordinating the efforts of research sites to converge at a particular time, such as during a workshop that brings researchers together. The cycle also imposes a natural rhythm for controlling the release of new or updated datasets. Both of these aspects serve to concentrate researchers' efforts.

MediaEval was established in 2008 as VideoCLEF, a track within CLEF. In 2010, it became an independent benchmark and adopted the name MediaEval. Participation in MediaEval is open to any research group or coalition of research groups. MediaEval builds on and extends existing benchmarking practice. MediaEval's yearly cycle involves selecting tasks, releasing a development set together with the corresponding ground truth, releasing a test set, receiving and evaluating the solutions (runs) submitted by the participating teams and publishing the results.

MediaEval adheres to a set of principles that guide the benchmarking process, but the main rules of the benchmark are not imposed from the top down. Rather, they emerge naturally as participants follow the yearly benchmarking cycle and promote certain community practices. Because of MediaEval's social emphasis, most tasks admit the collection of additional data sources from the Internet, especially from multimedia sharing websites. The use of additional data is carefully coordinated to avoid mixing the development and the test set. Furthermore, up to half of the official runs submitted by a team are *required runs*, meaning that the task organizers specify certain combinations of modalities and resources that the runs must necessarily use. For example, teams might be required to submit one run using audio channel only. This practice lets us focus on certain aspects of tasks and make certain that participants' submissions are comparable.

In some ways, the data collections that we develop are independent of the actual benchmark. Once the benchmark has concluded, they remain publicly available. However, our experience with MediaEval has shown that a benchmarking activity centered on data collections lets us fully exploit the potential of the collections and enables new collection development.

The Benchmarking Community

The essence of any benchmark is the core group of researchers that form the benchmark community. The MediaEval community is a growing, but relatively small group—approximately 60 people attended the MediaEval 2011 workshop and the entire community is twice that size. Nevertheless, a key characteristic of the MediaEval community is its diversity. Our goal was to leverage the community and its diversity to develop forward-looking benchmark tasks and their datasets.

The yearly benchmarking cycle begins with collecting proposed tasks for the next year. In addition to a problem definition, data collection, ground truth, and evaluation metric, a task also requires specific characteristics. First, a task must be based on a real-world use scenario. In particular, we encourage researchers to team up with industry or content providers

to define tasks. Second, a task must be associated with a data collection that either exists or can be easily collected and licensed for distribution. Third, it must be possible to generate ground truth for a task within the capacity of the available resources. Fourth, a task must have task champions who are willing to take on the responsibilities of task coordinators. Finally, a task must have five core teams committed to task completion and supporting the coordinators.

The task proposals are gathered into a survey, which is then circulated within the MediaEval community and the multimedia community as a whole. The survey asks about researcher preferences for particular tasks. It also asks about some task design parameters, for example, concerning suitable evaluation metrics or data-collection size. The survey results are then compiled and discussed. The most popular tasks are chosen to run the next year. During the survey process, a task acquires community support as well as a refined definition and core participants or even organizers. This process makes it possible for a single individual—even one who has never before participated in the benchmark—to introduce a highly innovative idea into the benchmarking community and have it evolve into a task with an associated data collection. In the end, we end up with five to six tasks, each with two to four organizers, ideally drawing from both academia and industry.

Since MediaEval was established in 2008, one of the key lessons involves the positive effects of separating official and off-the-record runs. The purpose of a benchmark is, of course, to enable cross-site quantitative comparison of techniques—that is, each task declares a winner, the participant that achieved the highest performance score. This winner is determined using a restricted definition: we only consider the official evaluation metric and the official runs. Task organizers are allowed to complete runs and submit working notes papers, but all their runs are considered off-the-record runs to control for possible advantages. That we explicitly admit off-the-record runs encourages the participants to experiment with risky approaches and gives them the best chance to maximize their scores. By fostering risk taking, we aim to encourage innovation and reduce the danger that the community will converge on a single approach to a task that represents a local optimum.

Exploiting Crowdsourcing

Recently, it has become increasingly easy to acquire large collections of multimedia data by crawling the Internet. The rise of crowdsourcing platforms, and in particular Amazon's Mechanical Turk (MTurk), has revolutionized the creation of ground truth for such datasets at a lower cost while allowing researchers to take into account the combined judgments of a larger, more diverse population of annotators.

Crowdsourcing is the process of outsourcing small tasks on an online platform, where workers carry them out in exchange for a small compensation. MTurk refers to these types as human intelligent tasks (HITs). A requestor can review a worker's performance on a HIT and, depending on the satisfaction, reject or accept the HIT and pay the reward. The biggest challenge to achieving high-quality crowdsourcing is quality control. Workers have incentive to carry out HITs with as little effort as possible in attempt to earn money. Unserious work is harmful in a setting where reliable annotations are needed to create ground truth. The situation is exacerbated if requestors cheat workers directly or encourage shoddy work by accepting all HITs without review. In short, a crowdsourcing platform is a relatively unregulated marketplace and careful HIT design, good quality control, and conscientious treatment of workers is necessary in order to use crowdsourcing to generate ground truth for large-scale multimedia datasets.

The ESP image labeling game first demonstrated the benefits of crowdsourcing for multimedia.[7] In 2005, Amazon made MTurk public, opening it up to the multimedia community. The first work concentrated on labeling images. Researchers showed that crowdsourcing was effective for multimedia annotation and in particular that it can be used to generate annotations comparable in quality to expert annotations.[8] This early work demonstrated the ability of a majority vote to reduce the noisy judgments. Since then, researchers have used more elaborate mechanisms of quality control, such as recruiting workers and issuing a qualification or having workers validate other workers' work. Recently, literature has begun to appear that offers helpful guidance in the effective deployment of crowdsourcing for evaluation.[9]

Video annotation is notoriously time consuming because, unlike images or text, annotators

must view an entire video. Crowdsourcing thus has enormous potential to aid in the annotation necessary to develop large-scale multimedia data collections. The difficulty of creating a HIT that uses video is twofold. First, video is technically more challenging than displaying text and images and is sensitive to the workers' browser configuration and bandwidth connections. Developing video HITs requires relatively more technical sophistication and more time devoted to communicate with workers who have technical issues. Second, with video there is a large temptation for unserious workers to only partially watch a video.[10,11] Researchers have proposed various methods to maintain quality control with video, including a two-step recruitment process aimed at finding workers willing to carry out a high-commitment task.[11] Within MediaEval, crowdsourcing is used for ground truth creation not only because it is cost effective, but because the workers' behavior can give us insight into how humans react to specific multimedia content, particularly giving us a window on what is interesting and important to human viewers. MediaEval 2010 used crowdsourcing to create a dataset of self-reported boredom assessments for a set of travelogue videos.[11] In MediaEval 2011, crowdsourcing was used to identify segments of video that contain particularly interesting statements.

The Blip10000 Dataset

The Blip10000 data collection was created within a research effort dedicated to improving the combination of multimedia content analysis, user-contributed information such as tags and ratings, and the structure of social communities.[12] As such, the motivations for dataset creation were largely technology driven. However, to ensure that it would support the definition of real-world use cases, attention was paid to choosing a data source that would represent the activity of a naturally existing multimedia community. During the dataset construction process, we attempted to preserve as many characteristics as possible of the data's context in the wild.

The Blip.tv content is created by users who have gone beyond the point-and-shoot video capture methods common on platforms such as YouTube. Blip.tv contributors demonstrate at least basic proficiency in filmmaking. Such content is generally referred to as semiprofessional user-generated (SPUG) content, which tends to be scripted or well thought out. In general, it is aimed specifically at communicating a message or opinion or at entertainment. Blip.tv users publish video content in a series format that follows the same pattern as television—adhering to one particular format or topic, publishing at regular intervals, and targeting a broad audience. Covering a range of topics and styles, Blip.tv is representative of the general SPUG phenomenon. Technically, Blip.tv was a good choice for data collection because the licensing information is available with videos, letting us create a dataset consisting of exclusively Creative Commons licensed material. The goal for the Blip10000 collection was to collect 10,000 videos from Blip.tv. The set actually includes 14,837 videos for a total of 3,500 hours. The collection also includes the video metadata (titles, descriptions, uploader ID, and tags assigned by the uploaders).

Many users tweet about Blip.tv videos, so Twitter is also a good source of information about the social network that contextualizes the Blip.tv content. Technically, Twitter was a good choice because we were able to easily establish the connection between Blip.tv and tweets via the real-time search engine Topsy (http://topsy.com). We collected videos from any show on Blip.tv for which we knew from Topsy that at least one episode from that show had been mentioned in a tweet. Then, we searched Topsy again to collect all users that had mentioned any of these videos in their tweets. This set of users was taken as level 0 in our set. Up to 3,200 tweets were collected for each level 0. The users with which level 0 users communicate were taken as level 1 users, and their profiles were also collected. The collection process stopped after we had collected the users with which level 1 users communicate (level 2 users).

A subset of the Blip10000 data referred to as ME10WWW has been used to develop three tasks within MediaEval: the 2010 Wild Wild Web (WWW) Tagging,[13] 2011 Genre Tagging, and Rich Speech Retrieval tasks. ME10WWW contains 1,974 episodes (247 development and 1,727 test), consisting of approximately 350 hours of data. The episodes in ME10WWW were chosen from 460 different shows. A show with less than four episodes was not considered for inclusion in the set. ME10WWW was released with related resources

that participants can use to approach tasks. Shot boundaries and keyframes were extracted by the Technical University of Berlin.[14] Automatic speech recognition (ASR) transcripts[15] were generously provided by the Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI, www.limsi.fr) and Vocapia Research (www.vocapia.com).

Genre Tagging Task

The Genre Tagging task involves assigning videos labels from 26 fixed genre categories: art, autos and vehicles, business, citizen journalism, comedy, conferences and other events, default category, documentary, educational, food and drink, gaming, health, literature, movies and television, music and entertainment, personal or autobiographical, politics, religion, school and education, sports, technology, the environment, the mainstream media, travel, video blogging, and Web development and websites. Participants used speech, audio, visual content, or metadata features and approached the task as either a classification or IR task, following the techniques used for the WWW Tagging Task.[16] The use case was chosen via the usage-driven view. Blip.tv uses a set of genre categories to organize user videos, and we assumed that if this set is used by Blip.tv it must be relevant and useful for users. Thus, we adopted the Blip.tv specified genre categories as the ground truth. Although the ground truth was categories rather than user-contributed tags, we used the name Genre Tagging to signal the relationship with the 2010 WWW Tagging Task. Using the genre category labels available on Blip.tv made the generation of the ground truth for this task affordable and freed up resources in the community to devote more effort to algorithm development.

The usage-driven view is an appropriate choice in this case because there is not a general consensus in the multimedia community on what constitutes a genre. Much work involving visual features defines genre in terms of stylistic categories that tend to be better reflected in the visual channel—for example, the list animations, documentaries, movies, music, sports, news, and commercials[17] can be considered typical of this view of genre. Google proposed genre classification as an ACM Multimedia Grand Challenge task in 2009 and 2010. The task description suggested using the DMOZ Open Directory Project (www.dmoz.org), with top categories that are mainly topical (such as ''health'') or reflect intended audiences (such as ''kids and teens'').

The nature of genre is an ongoing discussion within the multimedia community and the complexity of the question suggests that it is not easy, or perhaps even useful, to strike a compromise. Furthermore, adopting a specific notion of genre from an unpartisan external source helps to settle concerns that the choice of task might be biased to a particular way of approaching that task—for example, visual feature versus spoken content based. Instead, we simply designed algorithms to recreate a classification already in use on the Web. This decision was well received by the community, and the Genre Tagging task was the most popular task in MediaEval 2011. The 10 teams that completed the task approached it with a balanced mix of techniques and features. Results that the algorithms achieved on the dataset are reported in full in the *MediaEval 2011 Working Notes Proceedings*.[13]

Rich Speech Retrieval Task

The Rich Speech Retrieval task involves finding the jump-in points in the test collection that are relevant to the set of test queries. It is considered a known-item task because it models the information needs of a user who is trying to refind a previously viewed video segment. Each query is considered to have only a single relevant result. This task considers richer information about the video than merely the spoken content. In particular, the queries express information needs using three dimensions: the topical content, the speech act, and the visual content. The task uses five different functions of speech, represented as speech acts: apology, definition, opinion, promise, and warning. The Rich Speech Retrieval task takes the technology-driven view on use case development. No existing system can provide users with results based on speech acts, but we conjecture that if the technology existed that users would use it. We base our motivation on the assumption that the reasons that a user would search for a certain utterance are related to the reasons that a speaker would produce that utterance in the first place. The similarity between the units we use and the speech-act-like units that have been used in dialogue act modeling[18] provides further support for the idea that

people would find it useful if we could carry out retrieval on the basis of speech acts.

To carry out the rich retrieval tasks, it is necessary to develop a system that responds to a user query by retrieving a ranked list of jump-in points. The ground truth for the development set includes queries and the corresponding reference jump-in points for each one and is intended to be used to tune the system. The ground truth creation includes identifying appropriate pairs of a query and jump-in point for use in the task (50 for test purposes and 30 for development purposes). The queries were provided to participants in long and short forms. Here is an example apology: ''How does Peter Busch, staff member of the Morning Swim Show, save face after the faux pas he made during his interview with Terry Denison?'' (long form) and ''Peter Busch president chairman Denison morning Swim Show'' (short form). We defined the task with the expectation that participants would approach it using features derived from speech, audio, visual content, or metadata. The *MediaEval 2011 Working Notes Proceedings* reports the results achieved on this task.[13]

The query set was created for the task with MTurk using a HIT that asked workers for two contributions. First, they located portions of the videos that they would be interested in sharing online, and second, they contributed a comment on what the video portion was about (long form) and a query that would allow them to refind the chosen portions (short form). The first step in the HIT development was to create a pilot HIT, have it be carried out by a sizeable number of workers (we used 55), and then review the results. The pilot let us revise the form of the HIT to make it easy to understand for the workers as well as interesting to work on to discourage cheating. The initial pilot contained too many technical terms, such as transcript, quote, and category, that we identified as confusing. We revised the HIT by describing our task and referencing video sharing, a familiar style of working with videos online.

The final formulation of the main question in the HIT read as follows:

> Imagine that you are watching videos on YouTube. When you come across something interesting you might want to share it on Facebook, Twitter, or your favorite social network. Now please watch this video and search for an interesting video segment that you would like to share with others because it is (an apology, a definition, an opinion, a promise, a warning).

Figure 1 shows a screenshot depicting the top portion of the HIT (see http://blip.tv/morning-swim-show/friday-september-12-2008-1260838). We found it critical to include examples of each of the speech act categories with the HIT. We found that workers would only watch approximately the first three minutes of the videos looking for suitable portions of video. We addressed this issue by cutting the video into segments to ensure that not all queries were identified at the beginning of the video.

Initially we started with a HIT reward of \$0.11, but we discovered during the pilot that the task was difficult and difficulty varied unpredictably from video to video. In reaction, we raised the HIT reward to \$0.19 and included a final question in which the workers could identify their own bonus (between \$0 to \$0.21), depending on how difficult they found the HIT. We found that workers appreciated our trust, and they appeared to do relatively more careful work on our HIT. The final set of 80 queries was taken from the set of 531 suggestions contributed by 311 workers. Of these suggestions, 54 percent were accepted as successful completions of the HIT (workers received) and of the accepted HITs 72 percent were chosen for inclusion in the query set.

Conclusion

The MediaEval Multimedia Benchmark exemplifies how community cooperation and crowdsourcing can be leveraged to push forward the state of the art in multimedia research. As crowdsourcing techniques are applied to develop additional large-scale multimedia data sets in the future, it will be important the continue to emphasize good task design, quality control, and conscientious treatment of workers. Relevance to real-world use scenarios and usefulness for multiple research purposes will deserve careful consideration. When it comes to choosing specific use cases, multimedia dataset development faces the challenge of maintaining the balance between the technology- and usage-driven views. The MediaEval community exploits consensus on task selection as a way to prevent wild goose chases after tasks that are completely infeasible or have no basis in practical

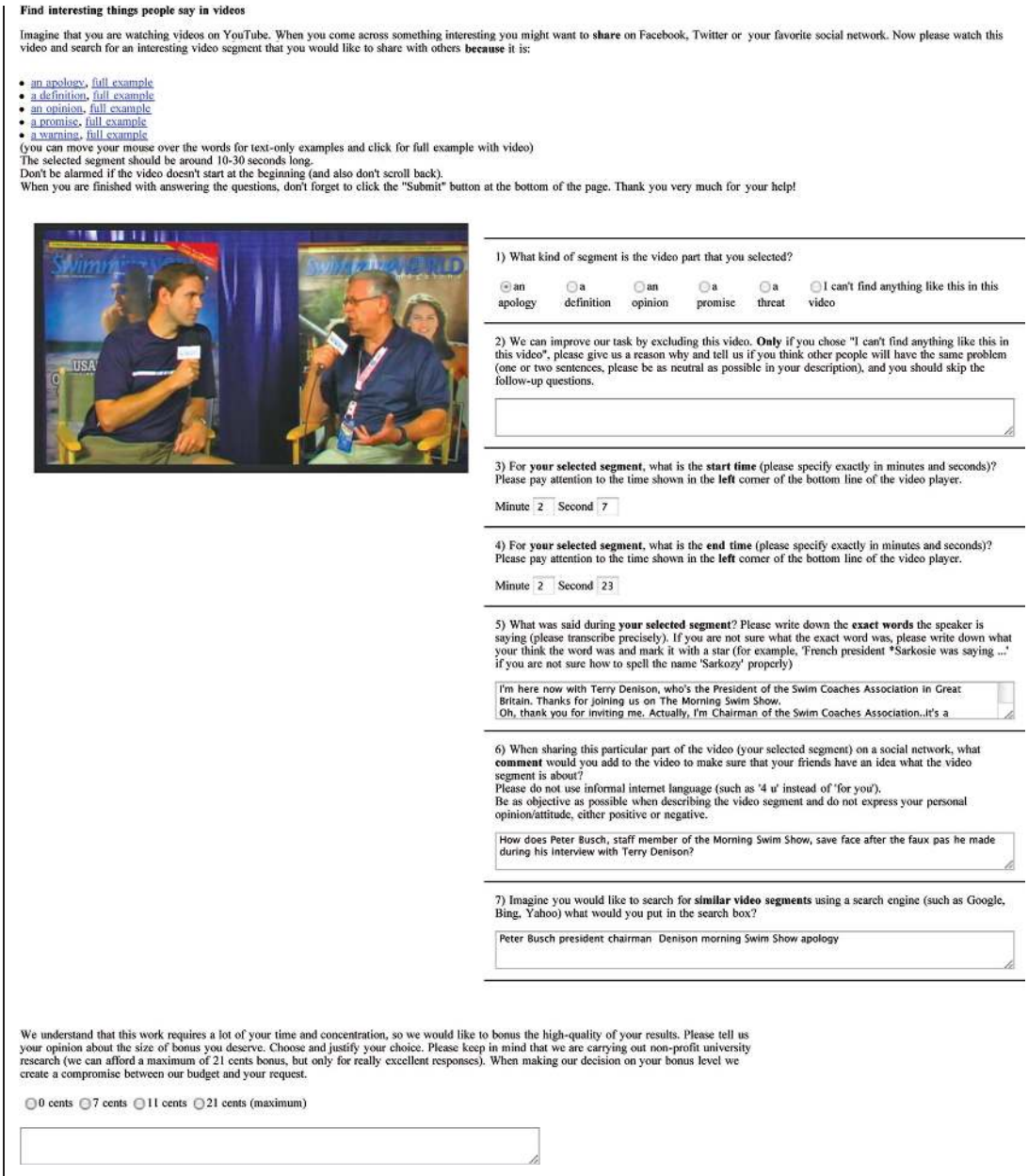

Figure 1. Screenshot of the human intelligent task (HIT) used on the Amazon Mechanical Turk crowdsourcing platform. Crowdsourcing helped the organizers generate the ground truth for the MediaEval 2011 Rich Speech Retrieval task.

application. At the same time, flexibility is high enough to create a forum in which risk-taking and innovation can flourish. **MM**

References

1. P. Over et al., ''Creating a Web-Scale Video Collection for Research,'' *Proc. ACM MM'09 Workshop on Web-Scale Multimedia Corpus* (WSMC), ACM Press, 2009, pp. 25–32.
2. G. Kossinets, ''Effects of Missing Data in Social Networks,'' *Social Networks,* vol. 28, no. 3, 2006, pp. 247–268.
3. B.R. Rowe et al., ''Economic Impact Assessment of NIST's Text Retrieval Conference (TREC) Program: Final Report,'' Nat'l Inst. Standards and Technology, 2010; http://trec.nist.gov/pubs/2010.economic.impact.pdf.
4. C.V. Thornley et al., ''The Scholarly Impact of TRECVid (2003–2009),'' *J. Am Soc. for Information Science and Technology,* vol. 62, no. 4, 2011, pp. 613–627.
5. A.F. Smeaton, P. Over, and W. Kraaij, ''Evaluation Campaigns and TRECVid,'' *Proc. ACM Int'l Conf. Multimedia Information Retrieval,* ACM Press, 2006, pp. 321–330.
6. C.W. Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems,* tech. report, Cranfield Coll. of Aeronautics, 1962; http://dspace.lib.cranfield.ac.uk/handle/1826/836.

7. L. von Ahn and L. Dabbish, ''Labeling Images with a Computer Game,'' *Proc. ACM SIGCHI Conf. Human factors in Computing Systems* (CHI), ACM Press, 2004, pp. 319–326.
8. S. Nowak and S. Rüger, ''How Reliable Are Annotations via Crowdsourcing: A Study about Inter-annotator Agreement for Multi-label Image Annotation,'' *Proc. ACM Int'l Conf. Multimedia Information Retrieval* (MIR), 2010, pp. 557–566.
9. O. Alonso, D.E. Rose, and B. Stewart, ''Crowdsourcing for relevance evaluation,'' SIGIR Forum, vol. 42, no. 2, 2008, pp. 9–15.
10. J.-I. Biel, O. Aran, and D. Gatica-Perez, ''You Are Known By How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube,'' *Proc. 5th Int'l AAAI Conf. Weblogs and Social Media* (ICWSM), AAAI Press, 2011, pp. 446–449.
11. M. Soleymani and M. Larson, ''Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus,'' *Proc. SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation,* 2010; www.ischool.utexas.edu/~cse2010/CSE2010-Proceedings.pdf.
12. N. Ramzan et al., ''The Participation Payoff: Challenges and Opportunities for Multimedia Access in Networked Communities,'' *Proc. ACM Int'l Conf. Multimedia Information Retrieval* (MIR), ACM Press, 2010, pp. 487–496.
13. M. Larson et al., eds., *MediaEval Benchmark 2011: MediaEval 2011 Benchmark Workshop,* 2011; http://ceur-ws.org/Vol-807.
14. P. Kelm, S. Schmiedeke, and T. Sikora, ''Feature-Based Video Key Frame Extraction for Low Quality Video Sequences,'' *Proc. 10th Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS), IEEE CS Press, 2009, pp. 25–28.
15. L. Lamel and J.-L. Gauvain, ''Speech Processing for Audio Indexing'' *Advances in Natural Language Processing,* LNCS 5221, Springer, 2008, pp. 4–15.
16. M. Larson et al., ''Automatic Tagging and Geotagging in Video Collections and Communities,'' *Proc. 1st ACM Int'l Conf. Multimedia Retrieval* (ICMR), ACM Press, 2011, article no. 51.
17. B. Ionescu et al., ''Content-Based Video Description for Automatic Video Genre Categorization,'' *Proc. 18th Int'l Conf. Advances in Multimedia Modeling* (MMM), K. Schoeffmann et al., eds., LNCS 7131, Springer, 2012, pp. 51-62.
18. A. Stolcke et al., ''Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech,'' *J. Computational Linguistics,* vol. 26, no. 3, 2000, pp. 339–373.

Martha Larson is a senior researcher in the Multimedia Information Retrieval Lab of Delft University of Technology. Her research interests include speech and language technology for multimedia retrieval. Larson has a PhD in theoretical linguistics from Cornell University. Contact her at m.a.larson@tudelft.nl.

Mohammad Soleymani is a Marie Curie fellow with the Intelligent Behavior Understanding Group at Imperial College London. His research interests include affective computing and multimedia information retrieval. Soleymani has a PhD in computer science from the University of Geneva. Contact him at m.soleymani@imperial.ac.uk.

Maria Eskevich is a doctoral student in the School of Computing at Dublin City University, Ireland. Her research interests include spoken document retrieval and evaluation metrics for speech search. Eskevich has a master's degree in theoretical and applied linguistics from Saint Petersburg State University. Contact her at meskevich@computing.dcu.ie.

Pavel Serdyukov is a senior researcher at Yandex. His research interests include enterprise and entity search, query log analysis, and location-specific retrieval and recommendation. Serdyukov has a PhD in computer science from the University of Twente. Contact him at pavser@yandex-team.ru.

Roeland Ordelman is a senior researcher in the Multimedia Retrieval at Human Media Interaction lab of the University of Twente and a manager of research and development at the Netherlands Institute for Sound and Vision. His work focuses on enhancing the exploitability of audiovisual content for various types of user groups using (semi-)automatic annotation approaches. Ordelman has a PhD in computer science from the University of Twente. Contact him at roeland.ordelman@utwente.nl.

Gareth Jones is an investigator with the Center for Next Generation Localisation (CNGL) and a faculty member of the School of Computing at Dublin City University. His research interests include multimedia, multilingual, and personal search. Jones has a PhD in electrical and electronic engineering from the University of Bristol. Contact him at Gareth.Jones@computing.dcu.ie.