International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Symbiosis Between the TRECVid Benchmark and Video Libraries at the Netherlands Institute for Sound and Vision

Johan Oomen · Paul Over · Wessel Kraaij · Alan F. Smeaton

Published in International Journal on Digital Libraries: Volume 13, Issue 2 (2013), Page 91-104.
The original publication is available at http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00799-012-0102-3

Abstract Audiovisual archives are investing in large-scale digitisation efforts of their analogue holdings and, in parallel, ingesting an ever-increasing amount of born-digital files in their digital storage facilities. Digitisation opens up new access paradigms and boosted re-use of audiovisual content. Query-log analyses show the shortcomings of manual annotation, therefore archives are complementing these annotations by developing novel search engines that automatically extract information from both audio and the visual tracks. Over the past few years, the TRECVid benchmark has developed a novel relationship with the Netherlands Institute of Sound and Vision (NISV) which goes beyond the NISV just providing data and use cases to TRECVid. Prototype and demonstrator systems developed as part of TRECVid are set to become a key driver in improving the quality of search engines at the NISV and will ultimately help other audiovisual archives to offer more efficient and more fine-grained access to their collections. This paper reports the experiences of NISV in leveraging the activities of the TRECVid benchmark.

Keywords TRECVid benchmark · audiovisual archives · video retrieval · digital libraries

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Johan Oomen
Netherlands Institute for Sound and Vision,
Amsterdam, The Netherlands.
E-mail: joomen@beeldengeluid.nl

Paul Over
National Institute of Standards and Technology,
Gaithersburg, Md. 20899-8940, USA.
E-mail: over@nist.gov

Wessel Kraaij
TNO and the Institute for Computing and Information Sciences,
Radboud University Nijmegen, The Netherlands.
E-mail: wessel.kraaij@tno.nl

Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies,
Dublin City University, Glasnevin, Dublin 9, Ireland
E-mail: alan.smeaton@dcu.ie

1 Introduction

The use of multimedia on the Internet is large and growing at an extraordinary rate. By 2015, one-million minutes of video content will cross the Internet every second [32]. It would take more than five years for one person to watch this amount of video. Content is created both by professionals and, increasingly, by everyday users. Besides the newly created material, a large body of existing, analog material is being migrated to digital files and managed by digital libraries. UNESCO estimates world audiovisual holdings at 200 million hours [29]. As many archives have a mission to disseminate their collections to a wide audience, more and more of this material will become available online.

Advances in multimedia information retrieval make it possible to offer fine-grained access to content on the

shot or fragment level rather than the entire video or program level, and gear access toward the specific needs of user groups. These systems can support exploration of digital libraries that go beyond just retrieval [13]. For instance, it's now possible to examine distributed collections in a digital library across time, space, genre and other dimensions such as color, origin, and so on, and these offer exciting possibilities for leveraging maximum benefit from the collection. Of course the usefulness of such novel ways of exploring distributed collections depends totally on the existence of valid use cases, both existing use cases and use cases not yet in place, in order to fully exercise and then exploit these new forms of exploration.

Audiovisual archives are investing in large-scale digitisation efforts of their analogue holdings and, in parallel, ingesting an ever-increasing amount of born-digital files in their digital deposits. Digitisation opens up new access paradigms and boosts re-use of audiovisual content. Query-log analyses show the shortcomings of manual annotations. This was evidenced in recent research that indicates that *"more and more user groups require and demand access to video fragments rather than entire programs — video fragments accounted for 66% of purchases in one recent study of a broadcast archive. Fine-grained manual annotation of video fragments is prohibitive, as the work involved is inevitably tedious, incomplete, and costly"* [10].

Increasing the use of the collections while managing the cost of collection creation (curation and annotation) requires research designed to better understand user requirements, to hasten the development of better search functionality for external users, and to help reduce cataloguing costs, among other things. To reach its goal of being *"the best archive in the digital domain"*, the Netherlands Institute of Sound and Vision (NISV) employs various sorts of research strategies into, amongst other topics, Multimedia Information Retrieval, or MIR. They include totally in-house studies, collaboration between internal and external experts, work vended out, and support of broad research community efforts. Over the past few years, the TRECVid benchmark has developed a novel relationship with the Netherlands Institute of Sound and Vision (NISV) which goes beyond the NISV just providing data and use cases to TRECVid. Prototype and demonstrator systems developed as part of TRECVid are set to become a key driver in improving the quality of search engines at the NISV and will ultimately help other audiovisual archives to offer more efficient and more fine-grained access to their collections. Other community efforts that are raising the bar of MIR include MediaEval, Pascal, MIREX, ImageCLEF and supportive actions such as PROMISE and CHORUSplus.

This paper is a case study which examines the details of Sound and Vision's relationship with the TRECVid workshop series with special emphasis on TRECVid's use of Sound and Vision video data from 2007 onwards. Section 2 outlines the general requirements for an audiovisual archive that operates at national scale and the necessity to collaborate with the academic and research community. Section 3 elaborates on the activities of TRECVid, specifically during the years 2007 to 2010 when content from Sound and Vision was used. Section 4 provides the summary discussion of what TRECVid researchers have learned from the collaboration, and where these findings contribute to Sound and Vision's research needs.

2 Users and User Requirements of a Media Archive

Sound and Vision is a typical, large-scale audiovisual archive, managing an ever-growing collection that currently comprises more than 750,000 hours of AV material. Dutch Public Broadcasting is one of the major sources of content but it is not the only source. Currently, born-digital material is ingested straight from the broadcast production environment directly into the Sound and Vision archive. Similar to many other audiovisual archives, Sound and Vision is engaged in large-scale digitisation efforts, migrating collections from analog carriers into digital format. At the present time, Sound and Vision's digital holdings comprise 6 petabytes of content and it is expected that this volume will grow to 15 petabytes by 2015 [31]. As the investments in digital libraries can only be justified if the hosted material is successfully accessed and re-used, offering seamless access routes to the content they hold is of crucial importance for archives. Reliable and scalable (automatic) annotation, indexing, and search are thus prerequisites for providing meaningful and efficient access routes to the increasing body of content.

Sound and Vision offers its services to diverse user groups. As Sound and Vision is the business archive for Dutch Public Broadcasters, broadcast professionals (documentary makers, journalists, and news editors) are traditionally an important user group. This user type mainly looks to re-use material in new broadcast productions. A second user group includes students and scholars in the humanities and social sciences who aim to use audiovisual archives as a source for diverse types of inquiry. Strongly connected to this user group are educators who use the audiovisual archive to search for relevant footage that they can use to support a specific

course or lecture. Finally, there is an increasing population of home users who access and explore the archive for personal entertainment or a learning experience.

These diverse user groups have a broad range of search needs. Queries can be on the level of what the programme is about, what can be seen in the shots, or both; they can be targeted towards broad categories of topics or genres, a specific programme or a single shot. Some users know exactly what they are looking for, while others have only a vague idea. The needs of television professionals relate to the genre and developmental stage of the programmes they make. A journalist who searches for a shot to illustrate an item in tomorrow's news bulletin may only have time to quickly scan the descriptions of a few programmes for a usable shot, while a documentary maker may have time to view multiple complete programmes before selecting a shot with the right content, atmosphere and aesthetic qualities. Years of experience at the customer service department of Sound and Vision have led to the following broad categorization of user queries from this diverse group of user types [8]:

1. Known item queries:
   - "The item about health care in the NOS news broadcast of the 15th of June 2008",
   - "The documentary by Henk de By about the Dutch painter Mondrian"
2. Subject queries:
   - General areas of interest : "all programmes about the Dutch economy"
   - Recognised areas of interest : "housing problems of Spanish immigrants in Amsterdam during the sixties"
3. Sequences, shots and quotes:
   - Specific: "shots of George Bush announcing war with Iraq",
   - General: "shots of sunsets"; "shots of Newfoundland".

2.1 The iMMix System and Its Cataloguing Rules

Information about clients and their needs is essential for setting the requirements of a video retrieval system. Sound and Vision has developed the iMMix multimedia catalogue system to preserve and manage the ever-growing collection of archive material. iMMix is today's entry point in Sound and Vision's archives, allowing users to view or re-use the material in an environment catering to their needs.

Figure 1 offers a high level overview of the iMMix infrastructure. Material is ingested into the digital storage facility from two sources: 1) current television production and 2) large digitisation efforts focused on the analogue (legacy) collection. The Digital Archive, a state-of-the-art storage capacity (top half of the figure), is managed by the iMMix Catalogue System. As material is digital, it becomes fairly straightforward to create various front ends each specially designed for different user groups. For example, the front end for broadcast professionals offers the possibility to order material. The front end for education features more concise metadata including keywords that match keywords in the textbooks students use.

Because of the temporal nature of audiovisual content, the catalogue description has to function as a substitute of the program itself [3]. A text description of shots and scenes, preferably time aligned, assists a user to quickly grasp the contents of a program without having to view it, even if played back at fast-forward speed. It needs to be noted here that the interaction design and capabilities of the graphical user interfaces plays an important role in supporting the search and exploration process. Integrating keyframes in the interface with search results, for instance, considerably speeds up the retrieval process. Additionally there is the requirement for content re-use, meaning that the description should facilitate easy re-use of parts. Both these demands lead to a cataloguing approach where an audiovisual product is seen as an aggregation of separate parts or elements, essentially a collection of clips. Catalogue entries therefore, are clip-based, item-focused and it is this approach where audiovisual archive practice clearly diverges from the way books and archival documents are usually made accessible, which is not on a chapter level, not on paragraph level, but on the level of an entire book.

The IMMIX system follows the IFLA-FRBR model [33] as a basis for its object-oriented data structure that models various audiovisual resources as well as online archive functionalities, within a professional broadcast production environment. The metadata model is open and flexible and thus can be extended, whenever necessary. The metadata model defines the way metadata should be structured. It is roughly divided into four stages: Work, Expression, Manifestation and Copy. Those four stages represent different layers in the model and Sound and Vision has extended the model so it is better suited for use in the audiovisual domain [14].

One of the other unique characteristics of this area is the semantic richness of audiovisual content. This implies that the same catalogue description has to deal with several different levels of meaning, and consequently, include different viewpoints. For example, for a given video artifact we should consider

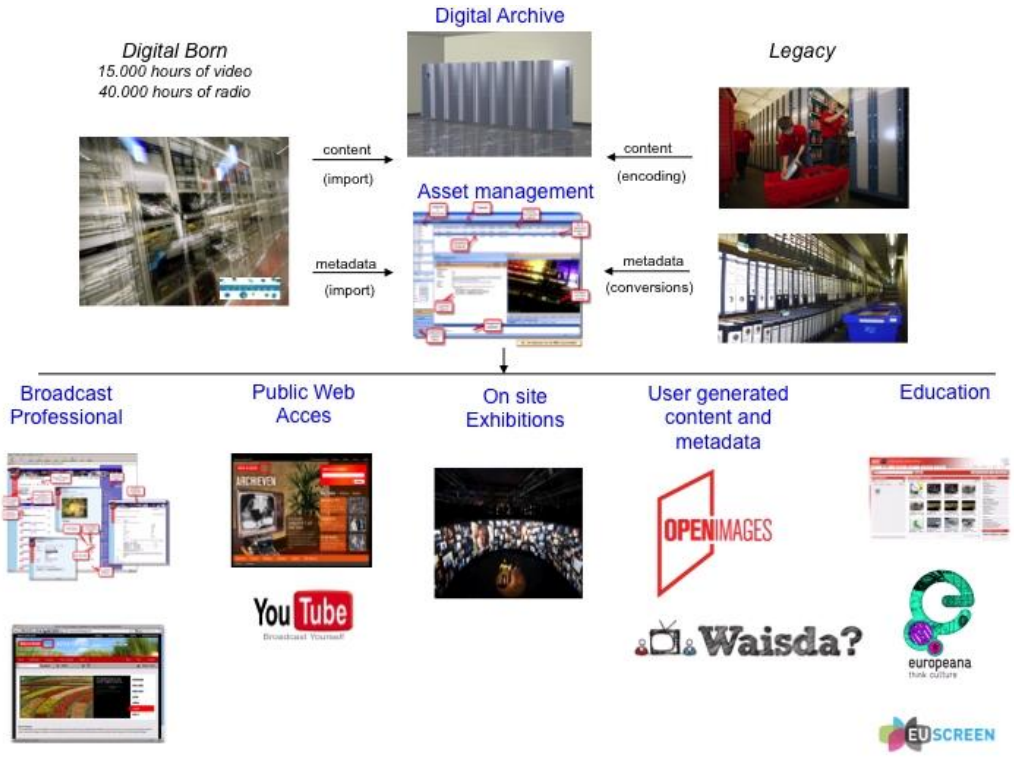

Fig. 1 The iMMix infrastructure at Sound and Vision.

1. Information content: the who, what, when, how and why
2. Audiovisual content: what is to be seen and heard
3. Stock shots: shots that can be re-used in a different context

Sound and Vision stores all metadata in an Oracle database. The cataloguers, responsible for creating catalogue entries, work with a tailor-made metadata editor that follows the iMMix metadata model. All data in the Oracle database are indexed by an enterprise search engine from Autonomy.

2.2 User Studies

To increase our knowledge of user needs, the MuNCH (Multimedia aNalysis for Cultural Heritage)[1] project has helped Sound and Vision design a query logging module that registers user behavior at the website of Sound and Vision. The website provides access to the archives by means of a search interface shown in Figure 2 where users can search by keyword or select a thesaurus term specifying broadcaster, subject, genre, name, etc. Programmes in the search results list can be viewed and ordered online.

[1] MuNCH is part of NWO's CATCH (Continuous Access To Cultural Heritage) program

Fig. 2 Sound and Vision search interface

All actions from users of this website are logged and stored in the query logging database. For example, a record is made of every time a user searches for a keyword, clicks on an item in the result list, renews a search using thesaurus terms, previews a video of the programme, bookmarks a programme, puts a programme in his or her shopping cart, etc. In total, 41 actions are distinguished. In addition, we can track users

over time to record which consecutive actions a user performs during a session. This information on user behavior provides a wealth of data about how users search and what they search for.

One of the dimensions measured by the query logging module is information on frequency of queries. This is valuable information for both Sound and Vision and the video retrieval community, since it allows for the identification of frequently occurring types of queries that merit further attention. Two categories of keywords were found to be used most often: names of people and names of television programmes [10]. Analysis of the user logs links the frequency information to information about the success of a search where success is measured by the number of times an item in the result list is viewed or ordered. By identifying which types of searches give the user a satisfying result, and which do not, we can target our research towards those types that need improvement. The MuNCH project has started to provide detailed analyses of this. This research indicates that users take two and a half times longer to order a fragment of audiovisual material than to order an entire program. This implies that manually reviewing the video material is a cause of the increased length of these sessions" [9]. Also, the query log analysis showed that *"The most frequently occurring keyword searches consisted primarily of program titles. However, these frequently occurring searches accounted for less than 6% of the keyword searches in the archive, leaving a long tail of unaccounted-for queries"* [9]. This work on query log analysis is now being continued within the EU FP7 AXES project[2]. In 2013, a query log module will be built as an extension of the iMMix system, allowing Sound and Vision to take further decisions on improving the annotation process.

2.3 Towards integrating (semi)automatic classification

Not so very long ago – and in many cases still – librarians and cataloguers were in full control of their catalogues. All metadata was manually produced and they had the first and the last say about what was coming in and what was going out. It was the professional who decided what was to be made accessible in the first place and how this should be done. Once an item or a description was entered into the catalogue for the first time, it was there to stay unchanged [3].

As audiovisual analysis and access technology reach levels of maturity that allows implementation in practical archival scenarios, archives are forced to re-evaluate their annotation strategies and access models with respect to their audiovisual content. To support this re-evaluation process, it is crucial to develop hands-on expertise with respect to new approaches in (semi-)automatic content annotation and access via research projects and collaborations with research groups, as well as to understand their opportunities and implications. To this end, Sound and Vision and three universities (University of Amsterdam, VU University and the University of Twente) recently announced a long-term, strategic collaboration regarding the development and implementation of multimedia information retrieval technology. In parallel, the needs of potential user groups with respect to accessing data need to be mapped, not only to select which data need annotations the most, but also to be able to make decisions on annotation levels and levels of automation.

For example, analysis of the Sound and Vision transaction logs revealed that users are not particularly interested in radio material for re-use. Nonetheless, archivists spend valuable time on annotating this type of material. As research pilots showed that automatic speech recognition provides reasonably accurate representations of the spoken word in radio content, re-allocating sparse manual annotation resources from radio to another type of content seems an obvious thing to do. Within the projects Multimedia aNalysis for Cultural Heritage (MuNCH), Vidi-Video[3], and more recently Access to Audiovisual Archives (AXES), extensive analyses of user behavior has been performed, resulting in a number of recommendations which can be seen in Section 4, that are used as a baseline for re-designing the operational services. In the process of exploring the opportunities and implications of new access paradigms, the unique properties of collections held by the archive that might possibly require tailor-made solutions need to be factored in.

The source material (film, television, music, and photographs) has been created over the past century on a variety of carriers and this affects the audible and visual quality range considerably. The collections have been catalogued using different types of cataloguing policies and systems, even by different organisations and organisational practices. Metadata is being converted into a single database, but differences between legacy systems are still visible in the metadata records. Theres a necessity for enriching the existing metadata, for instance by using text-mining to automatically extract keywords. As Sound and Vision also deals with Dutch-language content created over the past century, language-specific technology is required to manage the speech-to-text conversion. Since 2011, speech-to-text has been inte-

[2] `http://www.axes-project.eu/`

[3] `http://www.vidivideo.info/`

grated in the archival workflow for news radio annotation. Some experiments in the area of semantic video retrieval have also been executed in the past years, and it is expected that semantic video retrieval will be be integrated in the coming years. Section 4 will elaborate more on this topic.

Eventually, in the archival workflow, it is the archivist and not the technology expert who will decide whether or not to deploy a certain technology for annotation. Proper education and training of archivists becomes therefore a crucial element for the successful application of the technology. Moreover, by connecting multimedia technology with the human expertise of the professional archivist, being made concrete currently at Sound and Vision within a so-called 'Archivist Support System', some of the constraints with respect to the use of automatic annotation in an archival environment where authority plays such an important role, can, at least partly, be overcome. Consequently, the role of archivists is transformed from creating metadata to managing multiple streams of (semi-)automatically generated annotations and context information and to ensuring quality. Therefore, next to developing experience with current audiovisual access approaches geared toward the peculiarities of the collections in the archive, building up consciousness and educating the archival sector is of significant importance. Collaborative research prototypes that preferably can be evaluated by end-users are highly useful for this purpose. The TRECVid benchmark activity introduced in the next section has proven to be a fruitful breeding place for such research prototypes.

3 TRECVid and its use of the Sound and Vision collection

The TREC Video Retrieval Evaluation (TRECVid) [19] was begun on a very small scale in 2001 to extend the TREC/Cranfield philosophy to the promotion of research in video retrieval — with special emphasis on automatic content-based approaches. TRECVid has built on earlier multimedia research showing that multiple information sources (text, audio, video), each errorful, can yield better results when combined than used alone. Figure 3 depicts the evolution of TRECVid. Following the pattern of TREC, TRECVid provides a common foundation on which researchers develop, test, and improve their prototype systems in a laboratory setting. This foundation, developed by NIST in cooperation with the research community and other stakeholders, includes system tasks based on real user tasks, realistic training and test data, and appropriate measures of system effectiveness and usability. A shared research platform enables informed discussion of approaches and results at an annual workshop at which not only experimental successes but also the inevitable failures can be examined and learned from. As a laboratory exercise with prototype systems, TRECVid results tend to be indicative rather than final judgments. Credible evidence for particular approaches mounts only gradually as algorithms prove themselves repeatedly in various systems and against multiple sets of test data.

After two start-up years, TRECVid began a 4-year cycle in 2003 using broadcast TV news video, first in English and then additionally in Chinese and Arabic, building from 50 to 150 hours of test data each year. Twenty research groups initially, growing to 60 by the fourth year, submitted test system output to NIST for scoring in high-level feature detection (automatic content tagging), and search tasks. Secondary tasks included shot boundary determination, story boundary determination, and camera motion detection.

In 2007 TRECVid began a 3-year cycle testing systems on automatic and interactive search as well as high-level feature detection against television programming provided by the Netherlands Institute for Sound and Vision. Participation grew to nearly 80 teams with nearly 400 contributing researchers and from 50 hours of test data to nearly 280 in 2009. The shot boundary determination task against the Sound and Vision video was included as were a rushes summarization task against BBC video and an event detection task against airport surveillance video. Starting in 2010, TRECVid started using the HAVIC (Heterogeneous Audio Visual Internet Collection) Corpus [26] which consists of several thousands of hours of unconstrained user-generated multimedia content, in tasks in the surveillance detection area. A bibliometric study of TRECVid's scholarly impact from 2003 through 2009 [27] found 310 workshop papers and 2,073 peer-reviewed articles which had used TRECVid data in some way. A 2010 RTI International economic impact study of TREC/TRECVid [16] found that, for every US dollar that NIST and its partners invested, at least 3.35 to 5.07 USD in benefits accrued to Information Retrieval researchers.

Starting in 2010, TRECVid began using 400 hours of video contributed by many different individuals and groups to the Internet Archive (www.archive.org) — 200 hours for system development and 200 for testing. For the first time in TRECVid many of the videos were accompanied by donor-created metadata such as descriptions and keywords. The video was used for a new known-item search task and an expanded follow-on to the high-level feature detection task called semantic indexing, which tested 130 individual features. In addition, an experimental instance search task against the

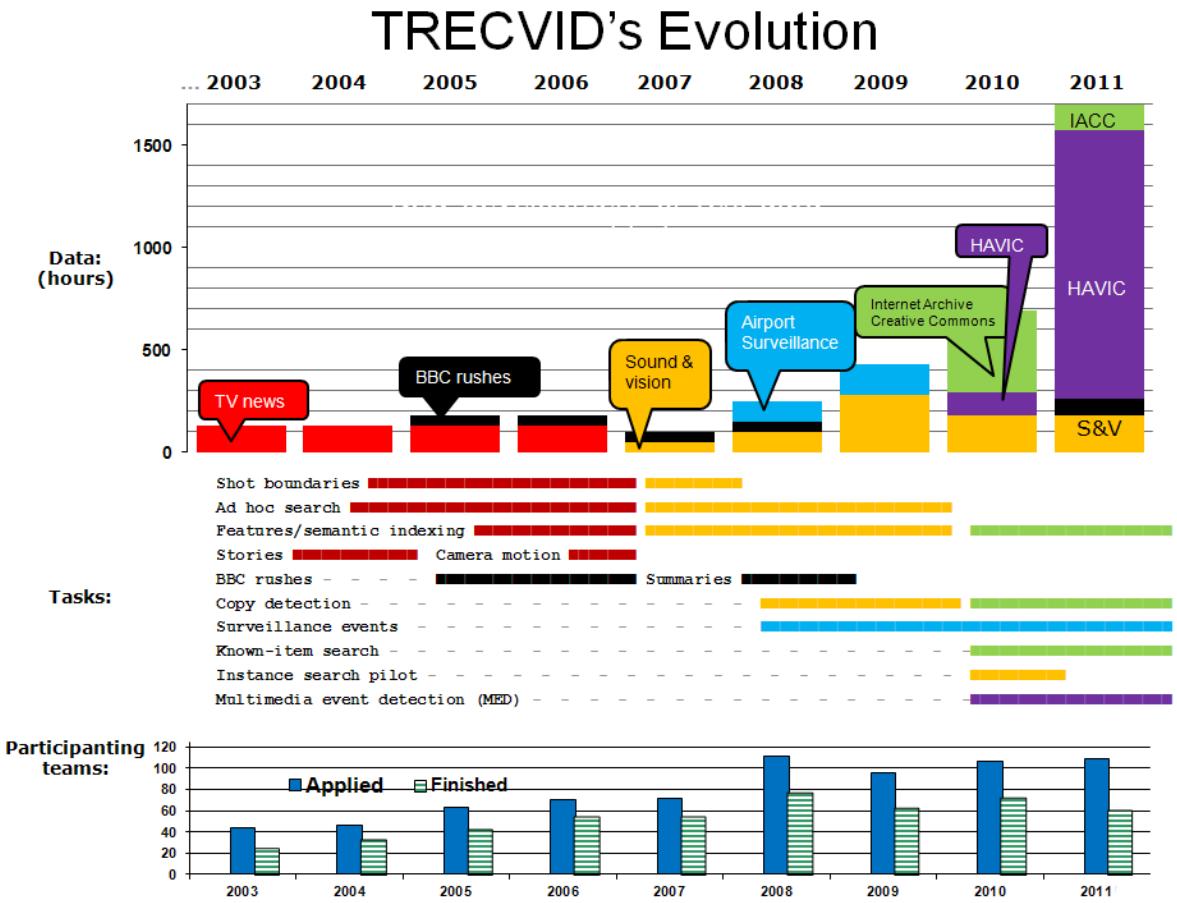

Fig. 3 Evolution of TRECVid reflecting source of video data, tasks and numbers of participating research groups.

Sound and Vision video was tested. In 2011 the known-item search and semantic indexing tasks continued, using a new 200 hours of Internet Archive video. The instance search task used BBC rushes.

3.1 User scenarios driving search task design at TRECVid

TRECVid tasks aim to be abstractions (simplifications) of real high-level tasks (or their component tasks) suitable for laboratory testing. The measures for system effectiveness are designed, among other things, to reflect significant real user priorities. User priorities are collected by the TRECVid organizers through interactions with various stakeholders; owners of large video repositories (Sound and Vision being one example), industry players, humanities researchers and so on. Relating the test tasks to real ones makes it more likely that test results can be related to real world concerns. Starting from real use cases helps motivate various decisions required in designing an evaluation framework. In what follows we discuss three TRECVid end user search tasks: ad hoc search, known-item search, and instance search as well as two component tasks: shot boundary determination and high-level feature detection.

3.1.1 Ad hoc search

From the beginning, the TRECVid ad hoc search task has been based on the use case of someone querying a large archive for video with the intent of reusing it. The video in the archive can be thoroughly analyzed and indexed but basically only once. The specifics of the queries come as surprises to the systems. This scenario does not emphasize unfocused browsing and does not include searching for information, i.e. question answering. There were two reasons for the choice of this scenario. First, it is commercially important in the production of new video and key to some uses within the intelligence community where objects, people, events in the background can be as important as those in the foreground. Second, this sort of search is well-aligned with TRECVid's focus on encouraging relatively greater exploitation of the visual content in videos not just re-

liance on use of text from speech or manually created textual metadata that reflect the creator's main intent.

The few studies of queries against large visual archives [1,5], etc. and analyses within the library community [17] have indicated that search requests could be targeted at people, objects, locations, and events both generic and named, in various combinations, and qualified by needs for particular production effects (shot types, camera motion, etc.). As a result, TRECVid ad hoc search topics – statements of need for video – were constructed by NIST to follow the model described above. By design they present functionality still not available in operational search engines. Subsequent user studies for example those by Halvey and Keane [6] and by Tjondronegoro *et al.* [28] are mostly limited by analysis of searches using existing text-based video search engines. Studies such as [7] and [15] are notable exceptions in that they try to explore user needs beyond the limitations of current applications. Searchers (experts and non-experts) will use more than text queries if available: concepts, visual similarity, temporal browsing, positive and negative relevance feedback. This can be seen clearly in the activities of the VideOlympics[4] and also in work by Christel [2] and by de Rooij *et al.* [4].

Each ad hoc topic in TRECVid contains a short textual description in English of the needed video, e.g., "Find shots of people shaking hands" or "Find shots of two people sitting at a table". In addition, there are usually several example videos clips and example images. Originally audio-only examples were included but this was soon dropped as relevant audio was usually found with relevant video in the video examples. If all parts of the topic are presented to the test system, one can think of the topic as documenting an intermediate stage of the search in which some of the desired video has been found along with example images from various sources and more, similar video is wanted. TRECVid also tested systems in a "text-only" condition by presenting only the textual part of the topic to the system and a "visual-only" condition in which the text was not presented. Search systems had full access to the video in the test collection but not to any manually created metadata for those videos. At least two dozen topics were created each year for testing.

The TRECVid test collections for ad hoc search were pre-processed to automatically identify shots and in response to each topic, and the test systems had to return a list of up to 1,000 shots ranked according to the weight of evidence that it met the need described by the topic. Several measures of system effectiveness were calculated for each search but the primary single measure was average precision — a metric which combines a measure of whether the system returns *only* the desired video and an estimate of whether it returns *all* the desired video in the collection. Average precision also strongly favors systems that put desirable video high in the ranked list of results — reflecting the searcher's presumed reluctance to work his/her way deep into the results list. The ad hoc search task used broadcast news video from 2003 to 2006 and Sound and Vision videos from 2007 to 2009. Results vary from topic to topic but a top-rated fully automatic search system in 2009 returned on average 5 relevant shots in the top 10 shots returned for each topic. Interactive systems performed significantly better, returning 8 relevant shots among the top 10 returned per topic.

[4] `http://www.videolympics.org`

3.1.2 Known-item search

The known-item task models the situation of a person who had seen a video, remembers a few things about it, believes a given archive contains it, but is not sure how to go directly to it. TRECVid assumed the user/system would start with a text-only query describing whatever was remembered about the content of the desired video. Participating systems could be fully automatic, in which case they returned for each query a list up to 100 videos ranked in descending likelihood of being the target, or they could be interactive, in which case they returned a single video purported to be the target. The known-item search task used Internet Archive videos in 2010 and 2011.

The test queries were created by NIST contractors. Given a video, the query creator watched it, closed his/her eyes to recall salient elements, and then formulated the query text — a list of objects, people, locations, etc. which the creator thought would be likely to uniquely identify the target video, e.g., "Find the video interviewing a man seated in white shirt, grey trousers and speaking about art". The query creator then added a list of words or phrases identifying important visual clues from the query text: objects, people, locations, etc., e.g., "man, white shirt, grey trousers, art". Automatic systems were scored using an average across all 300 queries and a per-query measure that diminished sharply as the depth in the result set increased. Interactive systems were scored based on the number of targets found. Final system effectiveness was measured with an average over all 300 queries. Sixty-seven of the 300 queries were not successfully answered by any system. The top automatic system scored 0.4 out of 1.0, while the best interactive system scored 0.7.

3.1.3 Instance search

In instance search the system's task is to find instances of a specific person, character, object, or location starting from a few image examples of the target, access to the video from which the examples came, and the type of the target (person, character, etc.). Relevant user scenarios could involve seeing a person (well-known or not) during a video search and wanting to find more video of him/her, wanting to find all videos shot in a particular location or looking for all videos containing a particular product or logo. Systems were scored as for an ad hoc search reflecting a strong desire to find the correct answer at the top of the ranked list returned by the search system. The instance search task used Sound and Vision videos in 2010 and BBC rushes video in 2011. In 2010, the first year, some of the 22 topics were harder than others and the best interactive system found on average eight correct shots in the top 10 returned, while automatic systems averaged less than one correct shot in the top 10.

3.1.4 Two component tasks

In addition to evaluating high-level tasks such as search, TRECVid early on identified e.g., shot boundary determination and high-level feature detection as important component tasks and worked to promote progress in these two areas.

Shots provide a more fine-grained unit of retrieval than the whole video and have been used as such in evaluating the ad hoc and instance search tasks. This is only possible because effective fully automatic shot boundary algorithms exist. Shot boundary determination was run in TRECVid from 2001 through 2007 against video that varied in the frequency of shot transition types [18]. Systems were asked to locate by frame all the abrupt transitions from one shot to another (cuts) and the gradual transitions and indicate the type (cut or gradual) of each. On both broadcast news and Sound and Vision data, best systems found better than 90% of the actual cuts (recall) and of the transitions they found, better than 90% were cuts (precision). Finding gradual transitions was more difficult but again best systems achieved about 70% recall and 80% precision.

3.1.5 High-level feature detection

In high-level feature detection, which can also be seen as automatic tagging, systems use training data to create a set of models for a set of features [20]. The models can then be matched against test video to detect the presence or absence of a feature. Search systems can try to use the models in representing and then executing user search queries. A hierarchy of automatically derived features can help bridge the gap between pixels and meaning and can assist search, but problems abound. What is the right set of features for a given application? Given a query, how do you automatically decide which specific features to use?

In TRECVid feature detection the systems are given textual definitions of a set of features to detect and the test data divided into shots. They are to return for each feature a list of up to 2,000 shots ranked by likelihood that the shot contains the feature at some point. In 2009 the Sound and Vision test videos comprised 380 hours, or 412 files making up almost 94,000 shots from 184 unique program titles. Forty-two research groups submitted 222 experimental results, each for 20 features and all were manually judged for correctness. The features were as follows: Classroom, Chair, Infant, Traffic intersection, Doorway, Airplane-flying, Person-playing-a-musical-instrument, Bus, Person-playing-soccer, Cityscape, Person-riding-a-bicycle, Telephone, Person-eating, Demonstration-Or-Protest, Hand, People-dancing, Nighttime, Boat-Ship, Female-human-face-closeup, Singing. For a given system, results vary greatly from feature to feature, but as a point of reference, in 2009 a top-tier system found on average 60% correct shots in the top 30 shots returned. A retrospective study [23] at the University of Amsterdam showed continuing improvement in their MediaMill feature detection system – a doubling of the absolute system score used by TRECVid from 2006 to 2009, even when training data was of a different genre from the test data.

It is worth noting that for all these tasks, right across the spectrum of TRECVid user scenarios which we model, the evaluation metric used for each has been selected independently to suit the task, the data, and the motivating use case. No one metric would meet all the needs of all tasks and so TRECVid has developed with the tasks running fairly independently, though as we will see in the next subsection, sharing datasets and other resources.

3.2 Datasets - characteristics and their implications

Characteristics of training and test data determine which tasks the data can realistically be used with in TRECVid and with how well well the various algorithms will work. Because video data is very difficult to acquire, TRECVid changes it very slowly, using 3 large datasets in the course of 10 years: broadcast news, Sound and Vision programming, and beginning in 2010 consumer-donated video from the Internet Archive. In what follows we take note of some of the data characteristics, particularly of

the Sound and Vision video, that have been significant for TRECVid tasks and systems.

In the period between 2003 and 2005, TRECVid used broadcast TV news video from the late 1990s, first only in English and then additionally in Chinese and Arabic. The source broadcasters were NBC, CNN, ABC, MSN, CCTV4, PHOENIX, NTDTV, LBC, and Alhurra.

Sound and Vision provided 400 hours of video to be used for research by participants in TRECVid between 2007 and 2009. The 400 hours were chosen from a subset of the Teleblik collection selected to eliminate cartoon, news programming, and the so-called "talking heads". 300 hours came from frequent programming and 100 from various rarer programs. The subsequent division into test datasets sought only to achieve particular amounts (starting at 50 hours and doubling in each succeeding year) and a similar mixture of program sources in the several test datasets. TRECVid's avoidance of highly constructed collections reflects a learned respect for the complexity and variety of real video and the difficulty of predicting which characteristics of a particular video will make it an easy or hard match for a given topic.

In 2010 TRECVid started using Internet Archive video, which as expected exhibited much greater diversity in content, form, originating device, etc. than was found in professionally produced Sound and Vision or broadcast news video. The collection was created by randomly sampling a large set of videos available under Creative Commons licenses from the Archive and having durations from 10 seconds to 3.5 minutes. We wished to focus on short videos as most US Internet video viewing has involved short videos[5]. The collection was in no other way filtered or constructed in order to ensure realism.

The Sound and Vision videos were similar to the broadcast TV news in a number of ways but also significantly different both to the human observer and to the system algorithms developed by TRECVid participants while working on broadcast news. These differences affected search systems as well as shot boundary determination and feature detection software though the goal of developing video analysis tools which work across domains remains somewhat elusive [24]. Both sorts of video are clearly professionally produced with different but internally consistent program structure and styles. The Sound and Vision videos contained no advertising segments, while the broadcast news did. In general the news video seemed more likely to contain multiple (near) copies of some segments such as file footage of a famous person, object, or event used repeatedly or single source news video shots included in many reports of the same incident. Both the Sound and Vision and the broadcast TV news video contain people recurring as characters and as themselves. Examples include program hosts and sketch characters in the Sound and Vision programming and news anchors, reporters, program hosts, and famous people and celebrities in the news in the broadcast TV news video.

Both broadcast news and Sound and Vision video contained footage shot in a studio and outside, but broadcast news exhibited greater variety in the quality of non-studio video (e.g., from war scenes, disasters, etc). While "talking heads" both in the studio and outside are very frequent in news video, the Sound and Vision programs also features, if to a lesser extent, people talking in interviews, in discussion programs, one-on-one, and in small groups.

Another important characteristic of the Sound and Vision video is the speech. Text from speech via automatic speech recognition (ASR) is a powerful source of information for indexing but its usefulness varies by video genre. However, not everything/one in a video is talked about, or is newsworthy. Audible mentions are often offset in time from visibility. Not all languages have equally good ASR. TRECVid systems tested against broadcast TV news video found text from speech the single most useful basis for search; this was not the case when testing TRECVid systems against the Sound and Vision video. We have no systematic explanation of why this was so but some observations about the data may lead to some interesting hypotheses. Although both sources contain speech in various languages, Dutch predominates in the Sound and Vision video while, English, Arabic, or Chinese does in the TV news video, according to the broadcaster.

Since the TRECVid ad hoc topics contain text in English only, systems faced a cross-language retrieval problem and only automatic solutions were allowed. That is, systems could use automatic translations of the topics into the language of the video collections or convert the video's speech to text and then automatically translate it into the language of the topic text. Unfortunately no objective measurements exist for the quality of the various automatic speech recognition and machine translation.

It should be noted that as the number of gradual transitions between shots in the Sound and Vision video was significantly lower than in the broadcast news video used by TRECVid. This reduced the overall complexity of the shot determination task compared to broadcast news video. The 2007 data's shots were much longer

[5] `www.comscore.com`, 2009 US Online Video Viewing Sets Record

(275.3 frames/shot) on average than those in the broadcast news video from 2006 (157.7 frames/shot).

The Sound and Vision videos lend themselves to instance search because they contain programs with a relatively small set of actors and presenters. For many programs the actors appear and re-appear in various roles, with a small number of costume changes, and in settings (rooms, scenes) also re-appear with some variation. The videos can be seen and analyzed as representing several small worlds to be searched.

TRECVid did not use available professionally created metadata for the Sound and Vision videos since it was trying to promote development of systems not dependent on manually created metadata even though hybrid systems may in the end be the best solution. The broadcast news as used also had no metadata. The Internet Archive did have keywords and descriptions provided by the video donor. Results from the known-item search task in 2010 suggest the best results could largely be attributed to matching query text to video metadata.

4 Summary discussion

In this Section, we discuss what TRECVid and Sound and Vision have learned from each other, and where the TRECvid findings contribute to Sound and Vision's research needs and where not. We address how TRECVid's relevance to Sound and Vision might be increased and what are the current top research questions for Sound and Vision.

TRECVid engages a world-wide set of researchers, yielding a greater variety of approaches at once than work with one team. In the study into the scholarly impact of TRECVid (2003-2009) [27] it was noted that there is a great deal of interchange of intermediate data and other resources among the teams taking part each year. It was also noted that techniques which are introduced one year and shown to work well are then quickly incorporated into the work of other teams in the next year. Thus technology transfer among research teams is quite rapid.

The relationship between TRECVid and Sound and Vision is unique, and symbiotic, and is summarised in Figure 4. It is more than just a case of NISV providing video data and a deeper insight into use cases which was the initial and the "easiest" part of the relationship. As a result, researchers working within the framework of TRECVid have developed prototype and demonstrator systems which fit the NISV use cases and operate on Sound and Vision data. These act as showcases to Sound and Vision of what is possible, a kind of glimpse into the near future for what search systems in video libraries might look like. Following that, Sound and Vision can then choose to engage directly with some of the research organisations in a knowledge transfer which described in the next subsection.

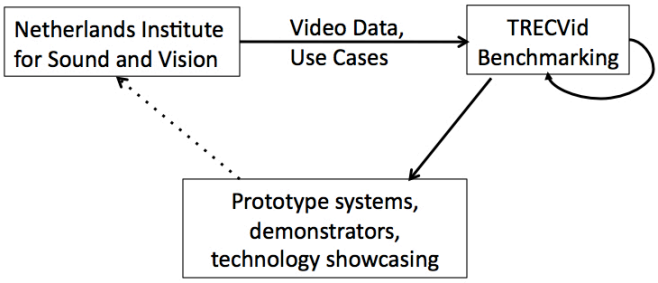

Fig. 4 Diagrammatic relationship between NISV and TRECVid.

4.1 Research and Knowledge Transfer

Sound and Vision benefits from TRECVid primarily through its strategic collaboration with the Intelligent Systems Lab Amsterdam at the University of Amsterdam (UvA). UvA has developed the MediaMill semantic video search engine that has been competing as a constant top performer in the TRECVid benchmark for a number of years. The MediaMill system has improved over the past years using the Sound and Vision data that was supplied to TRECVid and therefore likely to be relevant also for other videos in the archive.

Over recent years, UvA and Sound and Vision have been collaborating in a number of research programmes, based on the MediaMill search engine. These programmes have resulted in a number of demonstration applications, notably the Pinkpop Rock Festival multimedia search engine. This search engine facilitates semantic access to archival rock-and-roll concert video. The real-world application of multimedia retrieval technology (see Figure 5) was created in a collaboration between University of Amsterdam, University of Twente, system integrator Videodock and Sound and Vision. It combines content-based video retrieval and speech-to-text technology. For this pilot, Sound and Vision and UvA designed a *pipeline* for video concept detection, covering all steps from training, analyzing, indexing, and integrating the results in a website available on the web. The complete video archive contains 94 concerts covering 32 hours in total. Primarily, the Pinkpop demonstrator was a proof of concept, introducing the possibilities of content-based video search to the general public. The key novelty is a crowdsourcing mechanism, which

relies on online users to improve, extend, and share automatically detected results in video fragments using an advanced timeline-based video player shown in Figure 6.

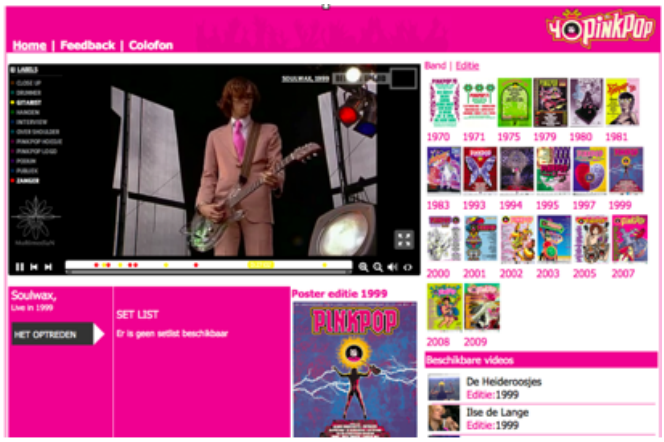

Fig. 5 Pinkpop Rock Festival multimedia search engine home page.

Fig. 6 Pinkpop Rock Festival multimedia search engine in-video search interface. The colored dots in the player indicate where semantic concepts are detected.

The video search engine was available online from December 2009 to February 2010. In the course of three months, almost 10,000 users visited the site and watched concert videos with the in-video browser. The user-feedback mechanism of the in-video browser made it possible to harvest positive and negative user judgements on automatically predicted video fragment labels. 958 users provided feedback on the video fragment labels. We received feedback on a total of 726 different fragments. Further analysis was conducted on the 510 fragments that received at least two judgements. In total these fragments received 3,567 different tags distributed over 62 concerts. The evaluation of these tags provided proof that user feedback can be exploited for incremental learning of visual detectors [22].

Through this project, Sound and Vision has managed to get a clear indication of the subsequent steps necessary to extract semantic concepts. A lot of effort has been invested in providing sufficient training data. The necessity of creating manually labeled visual examples is one of the fundamental obstacles in automatic indexing based on supervised machine learning.

TRECVid is a laboratory testing exercise, the results of which at best are indicative rather than conclusive and are no substitute for testing with real users in a real work environment. The results can be used for initial generation and testing of new approaches. Some of these may warrant further, more expensive testing in contexts progressively closer to Sound and Vision's operational reality. The TRECVid work is been carried forward on two distinct ways: validation in operational systems and active participation in research programmes.

4.1.1 Validation in operational systems

Collaboration with the TRECVid community supports Sound and Vision in formulating requirements to support computer vision as it redefines its enterprise architecture using TOGAF (The Open Group Architecture Framework) method and toolset[6]. These relate to outlining technical requirements to provide an interface to concept detectors, specific requirements to manage time-based metadata in search indexes and so on. To help ensure that research outcomes becomes actualized into a viable Web product or service, Sound and Vision operates a so-called laboratory environment, effectively a clone of the operational iMMix system. This laboratory environment allows Sound and Vision and their to test and benchmark outcomes of scientific research before it hits the mainstream market or is deployed in the operational iMMix system.

4.1.2 Active participation in research programmes

Sound and Vision participates in initiatives that incorporate results of TRECVid (see the tasks outlined in Section 3) in integrated systems. For instance, the four-year EU-funded project project Access to Audiovisual Archives (AXES[7]) develops a number of tools to support users in interacting with audiovisual libraries in new ways. In particular, apart from a search-oriented scheme, the project explores how suggestions for audiovisual content exploration can be generated via a myriad of information trails crossing the archive. The con-

[6] `http://pubs.opengroup.org/architecture/togaf8-doc/arch/`

[7] `http://www.axes-project.eu`

sortium includes a number of universities that are also participants of the TRECVid benchmark. The project consortium is developing tools for content analysis and deploying weakly supervised classification methods. Similar to the Pinkpop search engine, users are engaged in the annotation process and with the support of selection and feedback tools, this will enable the gradual improvement of tagging performance.

A second example is the work executed within the five-year Commit[8]) research program. The national research program is organized around several subprojects, one of which is called Socially Enriched Access to Cultural Media, or SEALINCMedia. Sound and Vision plays a major role in this subproject, which aims at facilitating natural and intuitive access to multimedia content in large, interoperable, linked cultural-media collections. As part of SEALINCMedia, UvA and Sound and Vision are exploring how user-tagged visual data provided by media sharing sites such as Flickr and YouTube can be used to create a large volume of visual examples. The results of this research effort will be evaluated within future editions of TRECVid.

4.2 Bootstrapping TRECVid output in digital video libraries

The VideOlympics mentioned earlier, and the "Video Browser Showdown"[9] at the International Conference on MultiMedia Modeling Conference in 2012, provided a playful platform to compare results from various contributing visual search engines in real time. In this evaluation effort a number of systems are linked to each other and simultaneously execute an interactive search task. It shows systems that are significantly different from existing operational systems and provides the market with input on future interaction design.

Over a number of years, TRECVid researchers, at least partly using Sound and Vision video and some use cases, have produced not only the demonstrator systems mentioned earlier but also a body of evidence supporting various approaches to video search and related video analysis problems. The insights from this body of scientific evidence will prove useful in the development of video library systems deployed at Sound and Vision and some of these insights are as follows:

- the combination of multiple information sources (text, audio, video), each errorful, achieves better results when combined than when used alone [25];
- the effectiveness of a hierarchy of automatically derived features in bridging the gap between pixels and meaning can be useful to assist search [11];
- the accuracy of automatic shot boundary detection and typing, and the feasibility of shot-based access allows for greater precision than clip-based access. This is confirmed by a recent user study [30] involving a number of broadcast professionals;
- the feasibility of automatic tagging based on analysis of a shot's visual and aural content, not just on file names or manually assigned tags, even on large amounts of video of interest, is now available to us [23];
- there is a varying but often significant usefulness of text from speech as a basis for video retrieval [25];
- there is significant impact of the human in the semi-automated search process or in the video tagging loop [25];
- there is a feasibility and acceptance of search (interfaces) driven by more than just keyword input but rather also by content-based approaches such as visual concepts, visual similarity, temporal browsing, positive and negative feedback, etc., presented in a variety of designs [4];
- there is an increase in performance of automatic tagging systems using more than one keyframe per shot to represent the shot and the concomitant need for faster processing [21].

A recent interview with the director of Sound and Vision entitled "Video Search Still a Tough Nut to Crack" [12] states that it is assuring that the active and ever-growing TRECVid community is working closely with end-user stakeholders such as large video digital libraries to further push the state of the art in this area. It is through on-going user-studies and applying the outcomes of research that the archives will be able to maintain their relevance. At the Netherlands Institute for Sound and Vision, the collaboration with the TRECVid community has proved extremely beneficial. The research organisations benefit as well, as they have access to content and usage data and can work on user-driven research topics. The authors hope this model can serve as a blueprint for similar collaborations with other video digital libraries.

Acknowledgements Johan Oomen is supported by the AGORA project of the NWO CATCH programme. Alan Smeaton would like to acknowledge support of Science Foundation Ireland through grant number 07/CE/I1147.

[8] http://www.commit-nl.nl

[9] http://mmm2012.org/vbshowdown/

References

1. L. H. Armitage and P. G. B. Enser. Information Need in the Visual Document Domain: Report on Project RDD/G/235 to the British Library Research and Innovation Centre. School of Information Management, University of Brighton, 1996.
2. M. G. Christel. Establishing the utility of non-text search for news video retrieval with real world users. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 707–716, New York, NY, USA, 2007. ACM.
3. A. De Jong. Users, producers & other tags. Trends and developments in metadata creation, Lecture at the FIAT/IFTA conference, Lisbon, Portugal, Oct 2007.
4. O. de Rooij, C. G. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 485–494, New York, NY, USA, 2008. ACM.
5. P. G. B. Enser and C. J. Sandom. Retrieval of Archival Moving Imagery — CBIR Outside the Frame. In M. S. Lew, N. Sebe, and J. P. Eakins, editors, *Image and Video Retrieval, International Conference, CIVR 2002, London, UK, July 18-19, 2002, Proceedings*, volume 2383 of *Lecture Notes in Computer Science*. Springer, 2002.
6. M. J. Halvey and M. T. Keane. Analysis of online video search and sharing. In *Proceedings of the eighteenth conference on Hypertext and hypermedia*, HT '07, pages 217–226, New York, NY, USA, 2007. ACM.
7. M. Hertzum. Requests for information from a film archive: a case study of multimedia retrieval. *Journal of Documentation*, 59(2):168–186, 2003.
8. L. Hollink, G. Schreiber, B. Huurnink, M. Van Liempt, M. de Rijke, A. Smeulders, J. Oomen, and A. De Jong. A multidisciplinary approach to unlocking television broadcast archives. *Interdisciplinary Science Reviews, 34*, 2(3):253–267, 2009.
9. B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis. *Journal of the American Society for Information Science and Technology*, 61(6):1180–1197, June 2010.
10. B. Huurnink, C. G. M. Snoek, M. de Rijke, and A. W. M. Smeulders. Today's and tomorrow's retrieval practice in the audiovisual archive. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 18–25, New York, NY, USA, 2010. ACM.
11. B. Huurnink, C. G. M. Snoek, M. de Rijke, and A. W. M. Smeulders. Content-based analysis improves audiovisual archive retrieval. *IEEE Transactions on Multimedia*, 14(4):1166–1178, 2012.
12. J. Jackson. Video search still a tough nut to crack. *Computer World*, May 2010.
13. W. McCarty. Beyond retrieval? Computer science and the humanities, November 2007. Keynote lecture for the CATCH Midterm Event, Den Haag, The Netherlands.
14. J. Oomen and H. Smeulders. D2.1 first analysis of metadata in the cultural heritage domain (project deliverable MultiMatch project), 2006.
15. H.-T. Pu. An analysis of failed queries for web image retrieval. *Journal of Information Science*, 34(3):275–289, June 2008.
16. B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic Impact Assessment of NISTs Text REtrieval Conference (TREC) Program. Technical Report Project Number 0211875, RTI International, June 2010.
17. S. Shatford. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging and Classification Quarterly*, 6(3):39–61, 1986.
18. A. F. Smeaton, P. Over, and A. R. Doherty. Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114:411–418, April 2010.
19. A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
20. A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.
21. C. Snoek, M. Worring, J.-M. Geusebroek, D. Koelma, and F. Seinstra. On the surplus value of semantic video analysis beyond the key frame. In *International Conference on Multimedia and Expo, 2005. ICME 2005. IEEE. Amsterdam, The Netherlands*, page 4 pp., july 2005.
22. C. G. Snoek, B. Freiburg, J. Oomen, and R. Ordelman. Crowdsourcing rock n' roll multimedia retrieval. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1535–1538, New York, NY, USA, 2010. ACM.
23. C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 43(6):76–78, June 2010.
24. C. G. M. Snoek, K. E. A. van de Sande, D. C. Koelma, and A. W. M. Smeulders. Any Hope for Cross-Domain Concept Detection in Internet Video. In *The TRECVid Workshop*, National Institute of Standards and Technology, Gaithersburg, Md., USA, November 2010.
25. C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.
26. S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous Audio Visual Internet Collection. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).
27. C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003-2009). *Journal of the American Society for Information Science and Technology*, 62(4):613–627, 2011.
28. D. Tjondronegoro, A. Spink, and B. J. Jansen. A study and comparison of multimedia web searching: 1997-2006. *J. Am. Soc. Inf. Sci. Technol.*, 60(9):1756–1768, Sept. 2009.
29. UNESCO. United Nations Educational, Scientific, and Cultural Organization: International Appeal for the Preservation of the World Audiovisual Heritage, April 2005. Last Accessed January 2012.
30. W. van den Heuvel. Expert search for radio and television: a case study amongst dutch broadcast professionals. In *Proceedings of the 8th international interactive conference on Interactive TV&Video*, EuroITV '10, pages 47–50, New York, NY, USA, 2010. ACM.
31. E. van Velzen. Must Archives Become IT Organisations? Lecture presented at FIAT-IFTA World Conference. Dublin, Ireland, 2010.
32. D. Webster. IP traffic to quadruple by 2015, June 2011. `http://blogs.cisco.com/sp/ip-traffic-to-quadruple-by-2015`.

33. Y. Zhang, I. Subirats, A. Salaba, C. Nicolai, M. Zeng, D. Hillmann, M. Zumer, and D. Neal. FRBR implementation and user research. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–3, 2010.