TRECVID: A Showcase for Video Indexing and Retrieval Systems

Alan F. Smeaton[1], Peter Wilkins[1], Marcel Worring[2],
Ork de Rooij[2], Tat-Seng Chua[3] and Huanbo Luan[4]

[1]Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University, Glasnevin, Dublin 9, Ireland.

[2]Intelligent Systems Lab Amsterdam, University of Amsterdam, Kruislaan 403, 1098 Amsterdam, The Netherlands.

[3]School of Computing, National University of Singapore.

[4]Institute of Computing Technology, Chinese Academy of Sciences, China.

Abstract

The growth in available online video material over the internet is generally combined with user-assigned tags or content description, which is the mechanism by which we then access such video. However, user-assigned tags have limitations for retrieval and often we want access where the content of the video itself is directly matched against a user's query rather than against some manually assigned surrogate tag. Content-based video retrieval techniques are not yet scalable enough to allow interactive searching on internet-scale, but the techniques are proving robust and effective for smaller collections. In this paper we show 3 exemplar systems which demonstrate the state of the art in interactive, content-based retrieval of video shots, and these three are just three of the more than 20 systems developed for the 2007 iteration of the annual TRECVid benchmarking activity. The contribution of our paper is to show that retrieving from video using content-based methods is now viable, that it works, and that there are many systems which now do this, such as the three outlined herein. These systems, and others can provide effective search on hundreds of hours of video content and are samples of the kind of content-based search functionality we can expect to see on larger video archives when issues of scale are addressed.

1. Video Search as a MMIR Application

Of all the media to which we now have relatively easy access, video, in digital form, is the one which has the steepest growth curve. Digital TV and set-top boxes, personal video recorders, DVDs and more recently internet video such as through YouTube or FabChannel, all contribute to placing enormous video archives at our disposal, if only we could navigate them effectively. Most of the technical issues associated with the video lifecycle are now solved to all practical intents and purposes. We can easily capture and store video, we can compress it, transmit it, and we can easily render it on fixed or mobile platforms. What remains our greatest technical challenge is being able to navigate it, to be able to browse it and search it in order to find clips which are of interest or of value to us.

The dominant approach to navigating digital video in large-scale practical applications is to use video metadata, either automatically determined or manually assigned. Automatic metadata includes date, time, and provides limited usefulness when the archives are large, though may be of use in smaller archives, such as personally-recorded video clips. Manual assignment of content description has always been hugely important and has been the bedrock on which navigation through video archives in all kinds of video libraries, has been based. Typically, at the time a video is included in a library, the video is annotated with content description which may include a title, actor(s), storyline, perhaps even a dialogue script. Publicly available systems such as Open Video[1] or the Internet Movie Database[2] are examples of systems using indexing and search based on metadata only, which have been in widespread use in large closed archives for some time.

We can now easily capture and store video and upload it to the internet for sharing. The proliferation of Web 2.0 applications has led to many systems where the description of shared video is augmented by user-assigned terms or tags. The Internet Movie Archive[3] and the popular YouTube system[4] are examples of systems where content description is determined partly by end users directly. This can take the form of user-assigned tags or keywords, or can be user reviews of the video. All of these can be used as part of the content representation of that video, and be used in retrieval.

While content description from user annotation offers useful navigation possibilities, it is still one step removed from being able to search actual video content directly. Effective use of user annotation relies on manual effort and we also depend on consistent annotation, and this is not scalable for developing fine-grained content-based access to large quantities of video. In this paper we concentrate on a multimedia information retrieval application which is direct content access to video where user queries are matched directly against the video content. In particular we present three example systems developed independently, which demonstrate differing approaches to video search. These were developed in the context of a large scale worldwide benchmarking activity where dozens of video indexing and retrieval systems are benchmarked on the same video dataset using the same search topics or queries and during the same time period.

The rest of this paper is organised as follows. In the next section we introduce the benefits of a common evaluation carried our across a number of research groups and in particular an introduction to the annual TRECVID activity is included. This is followed by an overview of three representative video retrieval systems from Dublin City University/K-Space, the University of Amsterdam/MediaMill, and the National University of Singapore, respectively. These three systems each have different approaches to the task of content retrieval from video and each has taken part in the

[1] www.open-video.org
[2] www.imdb.org
[3] www.archive.org
[4] www.youtube.com

interactive search task in TRECVID in 2007. The similarities and differences between these systems are presented and discussed in section 4, along with some overall conclusions.

2. TRECVID: A Benchmarking Evaluation Campaign for Video Retrieval

Evaluation and benchmarking has always been important in information retrieval. Since the earliest work of pioneers such as Cleverdon and Salton, measuring the performance of an indexing approach or of a new document ranking algorithm, on a test collection of documents, queries and relevance assessments, has repeatedly been the standard way in which the field of document retrieval has progressed. In the early 1990s, the introduction of the TREC workshops further increased the importance of common datasets, queries and relevance assessments for text-based information retrieval and further emphasized how empirical approaches to information retrieval evaluation are an integral aspect of this research field.

Evaluation and common benchmarking is also important in many kinds of image and vision processing. The development of video compression algorithms, for example, has always taken place in the context of shared and common datasets on which compression proposals can be compared directly. Currently there are several example evaluations for content-based tasks on video including ETISEO (Evaluation du Traitement et de l'Interprétation de Séquences Vidéo)[5] which targeted vision techniques for video surveillance applications involving pedestrians and/or vehicles, PETS (Performance Evaluation of Tracking & Surveillance) (Lazarevic-McManus *et al.*, 2006) which targets object detection and tracking for multi-view/multi-camera surveillance and ARGOS (Joly *et al.*, 2007) which targeted shot boundary detection, camera motion detection, person identification, video OCR and story boundary detection on broadcast TV news, scientific documentaries and surveillance video.

In terms of video retrieval the largest collaborative benchmarking activity for content-based activities is the series of TRECVID workshops, running annually since 2001 (Smeaton *et al.*, 2006). This has involved worldwide participation with over 50 research teams taking part each year in a variety of content-based "tasks" including shot boundary detection, concept or semantic feature detection, automatic summarization as well as content-based video retrieval. From 2003 to 2006 inclusive the data used in the search task was broadcast TV news and for 2005 and 2005 this was multi-lingual video, taken from several broadcasters spread across three languages, English, Chinese and Arabic. In 2007 the data used consisted of educational, cultural, youth-oriented programming, news magazines, historical footage video taken from the Dutch Sound and Vision archive and primarily in Dutch (Over *et al*., 2007*)*. This data contained no TV commercials, no repeated stock news footage, and had a great variety of subject matter. The volume of

[5] www.silogic.fr/etiseo

video data used varied each year, with 160 hours of MPEG-1 video used in 2006 for example.

The interactive search task presented to participants involved applying whatever video analysis and indexing tools a participant had to the search data and building their search system around that data. Participants were also able to take advantage of a variety of data donations made by the research community to the task and these included (for the 2007 TRECVID cycle alone) a master shot segmentation formatted as MPEG-7 for easy interchange, automatic speech recognition output and translation of that into English, low-level features derived from each shot, outputs from 374 semantic feature detectors applied to 2007 data and trained on 2005 data from Columbia University, applied to 2007 data and trained on 2006 test data from City University of Hong Kong, and two sets of manual annotations for 36 semantic features as the result of large-scale collaborative annotation activities (Quénot, 2007), (Jiang *et al.*, 2007).

The definition of the search task required each participating group to submit the results of running each of 24 topics or statements of information need against the search data. The shot lists returned by each participant were pooled together to some depth, duplicates were removed and shots were manually assessed for relevance by the TRECVID organizers. Once this ground truth of relevant shots for each topic was determined, the organizers were then able to compute the absolute performance figures for the submitted runs in terms of precision and recall as measured against the manually assessed pooled ground truth.

In the interactive search variant, participants were allowed to submit a number of runs (up to 6 in 2007) where each topic in each run was limited to the shots deemed to be relevant and found by one person using the participating site's search tool, as found within a 15 minute limit. This simulated the scenario where a searcher has a limited timeframe to find as many shots as he/she can where each shot is relevant to a fixed, unwavering information need. Such a scenario would regularly occur in a newsroom for example, where a production assistant seeks to locate video footage on a news topic to present to a news editor for possible inclusion in a broadcast. Three examples of search topics from 2007 appear below:

- Find shots of hands at a keyboard typing or using a mouse.
- Find shots of a canal, river, or stream with some of both banks visible.
- Find shots of a person talking on a telephone.

Each of the text description of topics in TRECVID is augmented by several illustrative images and/or video clips as exemplars of the information need, corresponding to the scenario where the searcher already has some images/video clips which are relevant to the information need. In the case of the first and last of these topics, the information need is for footage of some event which means that a still keyframe or single image taken from a shot is insufficient for determining an accurate match between shot and topic. This forces participating sites to develop techniques that analyse whole shots rather than keyframes.

The systems described in the next section of this paper are three of the twenty-four groups who participated in the search task in TRECVID 2007 and we describe each of these in turn. The three systems were chosen for their variety rather than their absolute performance characteristics and in order to illustrate the capabilities of contemporary content-based video retrieval systems.

3. Three Sample Video IR Systems

Each participant in the TRECVID search task normally addresses some research question or issue which is of interest to them, and will run more than one variant of their system in order to submit a number of "runs" which are each assessed manually by the TRECVID organisers. For each run we compute retrieval performance figures like precision and recall and these are averaged across the set of topics to give an indicative score of the performance of the system behind each "run". The data used in TRECVID 2007 consisted of approximately 50 hours of post-produced video from the Dutch Sound and Vision archive plus a variety of rich video metadata, automatically derived and donated to the TRECVID effort by participants. Many participating groups in TRECVID use these donations, especially the donations of semantic features, as an important component of their video search systems.

The three systems we will use to illustrate the capabilities of contemporary video retrieval are developed by Dublin City University/K-Space consortium, University of Amsterdam/MediaMill and the National University of Singapore.

3.1 Dublin City Universty/K-Space Interactive Video Retrieval

The team from Dublin City University led a TRECVID 2007 submission on behalf of the K-Space consortium, a large European multi-site grouping with an interest in semantic multimedia information management (Wilins *et al*., 2007). Video processing in this system began by selecting every second I-Frame from the video, and terming these *K-Frames*. For each of these frames, several low-level feature descriptors were extracted based on the MPEG-7 XM, including colour layout, colour moments, homogeneous texture, edge histogram and scalable colour. K-Frames were also segmented into regions using a Recursive Shortest Spanning Tree (RSST) approach (Adamek and O'Connor, 2007), and the same set of MPEG-7 features extracted for each region. Using this shared set of low-level features, several K-Space participants developed several automatic detectors for semantic concepts which determined for each shot whether the feature was present or absent. These included *sports, outdoor, building, mountain, waterscape/waterfront* and *maps* from Institut Éurecom, *face detecton* and *17 classes of audio type* from GET Paris, *building, car* and *waterscape-waterfront* from ITI Thesaloniki, *desert, road, sky, snow, vegetation, explosion/fire* and *mountain* from the National Technological University of Athens, *camera motion* and *number of faces*

visible from Joanneum Research Centre, and finally *maps, sky, weather, US-flag, boat/ship* and *vegetation* from Queen Mary University of London. All of these were then combined in the user interface for the system developed for the interactive search task in TRECVID.

The DCU/K-Space experiment under investigation in TRECVID was to examine the role of context in the user interface, where context can be described as showing for a given shot, temporally adjacent shots which a retrieval engine may or may not have ranked, i.e. its context. An example of the display of temporal context would be to issue to a retrieval engine, a query of an anchor person from a news broadcast. A temporal context response would be to return to the user, not just matching shots of anchor persons, but also the news story shots that the news anchor was presenting, which would not be visually similar to the initial query but might be relevant to the topic definition.

To examine the role of context, DCU/K-Space designed two user interfaces, known as the `shot based' system, and the `broadcast based' system. Both systems, apart from sharing the same retrieval engine, also shared a common query input panel, topic description panel and saved shot area. The major difference was in the presentation of the results from the underlying retrieval engine.

The `broadcast based' system takes the idea of context to its maximum by ranking not just individual video shots, but entire broadcasts, each consisting of potentially hundreds of shots. This presents an interesting alternative to a shot-only presentation of results and allows a searcher to explore the temporal neighborhood of shots. In Figure 1 we see a horizontal line of shots in rows across the results area. Each row is an entire broadcast, with the best-matching broadcast being the first row. When a user issues a query, the ranked list of broadcasts is presented, and within each broadcast, the row will be centered on the highest matching shot within that broadcast. The coverflow-like interface allows for rapid browsing of shots within a broadcast. Associated with each row is an iconic representation of the broadcast with the offset of the currently displayed keyframes shown as a red vertical bar, and the areas of the broadcast containing highly-scored shots shown in dark gray.

Figure 1 shows the user's multimodal query and includes the text "Find shots of a canal, river or stream with some of both banks visible" which is matched against the machine translation of the automatic speech recognition. Also included are two sample query images which have either been found by the searcher, or form part of the topic definition, and a subset of the available semantic features, in this case *outdoor* only. Query images are matched against the K-frames from each shot using the same low-level features mentioned earlier and each of the modalities (text search and image matching) generates a separate ranking of shots. Using a variation on a query-time weight generation techniques (Wilkins and Smeaton, 2006), the independent result lists are merged at query time with weights being assigned to each retrieval expert which approximate that expert’s likelihood of providing the most relevant responses to the query. The semantic concepts can then be used as filters by the user after a content-based query has been issued and these filters can be set to `positive', `negative' or `off'.

In Figure 2 we can see that the user's query has moved on and s/he has found a total of 6 query images but has disabled the semantic concept feature filtering of *outdoor*.

3.2 University of Amsterdam/MediaMill ForkBrowser

In traditional video retrieval systems users may query video archives by keyword, by example, by concept, by time or by program. Subsequently they browse through the results, and when the results are unsatisfactory the process reiterates. As a consequence of this iterative process a lot of time is spent on query specification. Moreover, when the target search results are not returned by the system in the initial queries a user may run out of query ideas. To alleviate both problems, the MediaMill team of the University of Amsterdam tries to depart from this traditional approach. This is done by providing users with browsers that allow to visualize the entire data set in multiple dimensions. This facilitates interactive exploration. For TRECVID 2007, the focus was specifically on consolidation of proven effective interface components from previous TRECVID editions into a novel browsing environment (Snoek *et al.*, 2007)

The notion of threads is introduced in the ForkBrowser in order to browse through a video data set in multiple directions. A thread is a linked sequence of shots in a specified order, based upon an aspect of their content (de Rooij *et al.*, 2007). Two types of threads are defined: *static threads* which are pre-computed, and *dynamic threads* which are generated on demand during a browse session. The content of a thread is based on a form of similarity between shots in the data set. The MediaMill 2007 video search engine offers the following threads and similarities.

- Visual threads: based on similarity in visual content,
- Time threads: based on temporal similarity between shots,
- Query result threads: based on similarity between shots and a user posed query,
- History threads: based on shots the user has already visited during this search

Each method yields a separate ranking of the data through which the user can browse.

The combination of the time thread with any other thread resulted in the CrossBrowser which proved effective for the TRECVID interactive search tasks in 2004 and 2005 when a single thread, for example a single concept detector query, is sufficient for the user to find shots which satisfy the topic or information need (Snoek *et al.*, 2007), (Snoek *et al.*, 2006). For topics that require a combination of threads, the RotorBrowser was introduced in 2006 (de Rooij *et al.*, 2007), (Snoek *et al.*, 2006). This browser allows a user to integrate query results with time, visual similarity, semantic similarity and various other shot-based similarity metrics. While effective, this visualization proved overwhelming for non-expert users. To leverage the benefits of having multiple query methods while simultaneously allowing the user to maintain an overview of their results, a new interface was introduced in TRECVID 2007 which combines query by keyword, query by

example, query using 572 semantic concepts, query by time and by program, all combined into a framework which we call the ForkBrowser.

The ForkBrowser visualizes results by displaying keyframes based on the shape of a fork. The contents of the tines of the fork depend on the shot at the top of the stem. The center tine shows unseen query results, the leftmost and rightmost tines show the time thread, and the two tines in between show user-assignable threads. For the TRECVID 2007 benchmark two variants of visual similarity threads are displayed. The stem of the fork displays the history thread. All browse directions, each tine and the stem, are accessible by keyboard and mouse for quick navigation. Every displayed key frame is taken from a single video shot, and the video shot can also be played on demand by rapidly displaying up to 16 frames in sequence from the originating shot. This helps in rapidly answering queries containing explicit reference to motion or to events. Figure 3 depicts the ForkBrowser while searching for "boats moving past". The horizontal tine shows shots from the time thread of the program "Klokhuis", the diagonal directions depict two visual threads to provide the user with similar shots from waterscapes which s/he can browse.

For the MediaMill TRECVID 2007 experiments two interactive runs were submitted. One with an expert user using the CrossBrowser, another with an expert user using the ForkBrowser. Experimental results showed that the ForkBrowser allowed the expert user to achieve nearly the same performance while using significantly less interaction steps.

3.3 NUS-ICT/VisionGo

VisionGo is an interactive video retrieval system developed jointly by the National University of Singapore (NUS) and the Institute of Computing Technology, Chinese Academy of Sciences (ICT). The system is designed to maximize the effectiveness of human annotators through the use of an intuitive User Interface (UI), options for multiple feedback strategies and motion icons. In performing an interactive search, we first utilize results from an automated search for user feedback. The automated search performs analysis of the user's multimodal query, followed by multimodal fusion to retrieve a ranked list of shots. Here, the multimodal fusion uses a combination of text derived from ASR (automatic speech recognition), high-level features (HLFs) automatically detected in shots, and a combination of low-level visual features and motion (Chua *et al.*, 2007). The user then makes use of an intuitive retrieval interface with a variety of relevance feedback options to refine their search results. In addition, we introduce motion-icons, which allow users to see a dynamic series of keyframes instead of a single keyframe during relevance assessment. The results show that this approach can help in providing more effective retrieval.

To maximize the user's interaction efforts, the intuitive UI is designed for fast keystroke actions with quick previews of previous and subsequent sets of shots in the ranked list of shots. A sample interactive UI is shown in Figure 4. The UI is inspired by high throughput interactive game interfaces, which are mainly keystroke based. The UI displays three images at a time in a central active row, with the previous and next rows in view. Each

image corresponds to a single retrieved shot, without any *context* such as previous or following shots. We experimented with various configurations of display and discovered that user reaction time is the quickest when annotating or marking as relevant or non-relevant, three images at a time. The user will determine the images' relevance to the query and annotate the positive ones by hitting pre-a defined set of keys on the keyboard. The system captures the user's input and automatically refreshes itself to display the next row of new keyframes in the ranked list. For fast throughput, we designed a number of shot-cut keystrokes for quick overall actions. In the event that no image is relevant to the query, the user can hit the "Space" key to skip a row. In addition, the "Space" key can also be pressed and held for "fast forward". Alternatively, the "Backspace" key can be used to undo changes and backtrack when the user needs to perform corrections. In experiments at the National University of Singapore, the UI enabled a normal user to annotate up to 3,500 shots based on motion icons or 5,000 shots based on static icons, in only 15 minutes.

To allow for more flexibility and to provide a range of options for users to click during relevance feedback, we propose to segregate interactive feedback into three distinct types, namely **recall-driven**, **precision-driven** and **temporal locality-driven feedback**. Each strategy aims at leveraging different aspects of user feedback data. At any time, if the user feels that the search and feedback process is not progressing well, he/she is able to select any feedback strategy button located at the bottom-left corner of the interface to enhance the search performance.

Recall-driven feedback employs general features such as the ASR text tokens and HLFs from relevant shots to perform query expansion. This option has been found to be the most effective in finding many new relevant shots in the initial stage of a search. Given the set of positively annotated shots, this process makes use of text and HLF scores to iteratively adjust the retrieval function. First, we extract highly discriminating text tokens from ASR texts of relevant shots by using the 0.5 feedback formula (G. Salton *et al.*, 1983). Second, the HLF scores for positively labeled shot are used to estimate the new score of HLF with respect to the query. Finally, we compute the new score for shot by fusing the new text and HLF scores.

Precision-driven feedback uses motion, visual and audio features in an SVM-based active learning environment targeting at improving precision. It uses active learning to provide long term improvements to classifiers. Fused with a performance-based adaptive sampling strategy, this process continuously re-ranks instances as the user annotates shots as relevant or non-relevant. The performance-based sampling strategy will adaptively choose instances either most ambiguous or most relevant from the classification output with emphasis on maximizing precision in a minimal time.

Temporal locality-driven feedback essentially returns shots from neighboring shots from the positively labeled set, as it is noted that positive shots tend to cluster near each other within the same story.

Based on these multiple feedback strategies, a user is able to choose the type of feedback that is more suitable based on his/her intuition or experience, in order to maximize

performance. We are currently experimenting and analyzing the effectiveness of using different feedback strategies and user interface options on interactive search.

We observed that many visual-oriented queries tend to be associated with objects in motion in the video. It is therefore necessary to provide some information on motion in the shot icon to facilitate the annotation process. Specifically, instead of displaying an icon with a static keyframe for each video shot, we construct a summarized clip comprising a sequence of progressive keyframes which can show moving picture information. We call this a motion icon or micon. Through the use of micons in previewing shots, the user has a clearer idea of what motion information is in the shot and can identify relevant shots more quickly and with better confidence. For example, for a motion-oriented query on "Find shots of one or more people walking up stairs", if the users were to be presented only with a single frame bounded in red box as shown in Figure 5, this shot would have been judged irrelevant. However, through the use of a micon, the user can identify straight away that this is a relevant shot.

The use of micons may also help in situations when the wrong keyframe was chosen for a shot. For example, for a non-motion oriented query such as: "Find shots of a canal, river, or stream with some of both banks visible". The shot with the keyframe shown in Figure 6 would be deemed not relevant. However, through the use of micons, we can assess that the shot is relevant to the query. The tradeoff in using micons is that the display speed and user reaction speed, is slower.

4. Interactive Video Search and Multimedia Information Retrieval

The three interactive video search systems presented in this paper are both similar and different. The similarities are that each supports a multimodal query from a user – a combination of text, sample images(s) and semantic features – which is implemented by running multiple shot ranking algorithms for each of the modalities and fusing them together at search time. Each supports a preview of a whole shot by presenting sets of keyframes, called *micons* by NUS, to allow a user to determine whether an event of some kind occurs within a shot.

Yet despite these similarities there are huge differences in the interfaces and user interactions among the three systems which have afforded each of them to explore some aspect of the retrieval interaction as an experiment. DCU/K-Space experimented with the effects of local context and within-broadcast impact on retrieval quality; University of Amsterdam/MediaMill experimented with the effects of different threads including a history thread, while National University of Singapore experimented with the effects of different relevance feedback algorithms.

Content-based interactive video search, as represented by the three systems in this paper, is not yet mainstream in terms of usage by a large population of users. Yet there is a need for this kind of functionality, especially as the volume of video available to us grows, and the demands of users have to be met. The techniques needed to realize a widespread

deployment of this, such as an internet-scale deployment, are well under development and are effective on archives of the order of hundreds of hours of content. The three systems presented here are representative of the state of the art and there are several other systems in TRECVID which can do likewise.

References

T. Adamek and N. O'Connor. Using Dempster-Shafer theory to fuse multiple information sources in region-based segmentation. In ICIP 2007 - Proceedings of the 14th IEEE International Conference on Image Processing, 2007.

TS Chua *et al.* TRECVID 2007 Search Tasks by NUS-ICT. In Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.

Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. ACM International Conference on Image and Video Retrieval (CIVR'07), Amsterdam, The Netherlands, 2007.

P. Joly, J. Benois-Pineau, E. Kijak, and G Quénot. The Argos Campaign: Evaluation of Video Analysis Tools. In Proceedings of the International Workshop on Content-Based Multimedia Indexing, 2007. CBMI'07, pages 130-137, 2007.

N. Lazarevic-McManus, J. Renno, J. and G.A. Jones, G. A. 2006. Performance evaluation in visual surveillance using the F-measure. In Proceedings of the 4th ACM international Workshop on Video Surveillance and Sensor Networks (Santa Barbara, California, USA, October 27 - 27, 2006). VSSN '06, 45-52.

P. Over, G. Awad, W. Kraaij and A.F. Smeaton. TRECVID 2007 - An Introduction. In Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.

G.M. Quénot. Active learning for multimedia. In Proceedings of the 15th international Conference on Multimedia (Augsburg, Germany, September 25 - 29, 2007). MULTIMEDIA '07. ACM Press.

O. de Rooij, C.G. Snoek and M. Worring. Query on demand video browsing. In Proceedings of the 15th international Conference on Multimedia (Augsburg, Germany, September 25 - 29, 2007). MULTIMEDIA '07. ACM Press, 811-814.

G. Salton and M.J. McGill. Introduction to Modern Retrieval. McGraw-Hill Book Company, 1983.

A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, 321-330.

C. G. M. Snoek, J. C. van Gemert, Th. Gevers, B. Huurnink, D. C. Koelma, M. Van Liempt, O. De Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H.C. Thean, C. J. Veenman, M. Worring. The MediaMill TRECVID 2006 Semantic Video Search Engine. I~n Proceedings of TRECVID 2006, Gaithersburg, Md., November 2006.

C.G.M. Snoek, M. Worring, D.C. Koelma and AWM Smeulders. A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. IEEE Transactions on Multimedia, 9(2), pages 280-292, 2007.

C.G.M. Snoek, I. Everts, J.C. van Gemert, J.M. Geusebroek, B. Huurnink, D.C. Koelma, M. van Liempt, O. de Rooij, K.E.A. van de Sande, A.W.M. Smeulders, J.R.R. Uijlings and M. Worring. The MediaMill TRECVID 2007 Semantic Video Search Engine. In Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.

P. Wilkins, P. Ferguson, and A. F. Smeaton. Using score distributions for querytime fusion in multimedia retrieval. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, 51-60.

P. Wilkins *et al.* K-Space at TRECVid 2007. In Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.

Figure 1: User interface for Dublin City University K-Space Search System

Figure 2: User interface for Dublin City University / K-Space Search System

Figure 3: User interface for University of Amsterdam's Search System

Figure 4: User interface for National University of Singapore's VisionGo Search System

Figure 5: A sequence of multiple keyframes for shot213_62

Figure 6: A sequence of multiple keyframes for shot149_62