



**Content-Based Video Retrieval: Three Example Systems from TRECVID**

Journal:	<i>International Journal of Imaging Systems and Technology</i>
Manuscript ID:	IMA-07-119.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Smeaton, Alan; Dublin City University, Centre for Digital Video Processing Wilkins, Peter; Dublin City University, Centre for Digital Video Processing Worring, Marcel; University of Amsterdam, Intelligent Systems Lab de Rooij, Ork; University of Amsterdam, Intelligent Systems Lab Chua, Tat-Seng; National University of Singapore, School of Computing Luan, Huanbo; Chinese Academy of Sciences, 4Institute of Computing Technology
Keywords:	video retrieval



## Content-Based Video Retrieval: Three Example Systems from TRECVID

Alan F. Smeaton<sup>1</sup>, Peter Wilkins<sup>1</sup>, Marcel Worring<sup>2</sup>,  
Ork de Rooij<sup>2</sup>, Tat-Seng Chua<sup>3</sup> and Huanbo Luan<sup>4</sup>

<sup>1</sup>Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University, Glasnevin, Dublin 9, Ireland.

<sup>2</sup>Intelligent Systems Lab Amsterdam, University of Amsterdam, Kruislaan 403, 1098 Amsterdam, The Netherlands.

<sup>3</sup>School of Computing, National University of Singapore.

<sup>4</sup>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

1  
2  
3 be able to browse it and search it in order to find clips which are of interest or of value to  
4 us.  
5  
6

7 The dominant approach to navigating digital video in large-scale practical applications is  
8 to use video metadata, either automatically determined or manually assigned. Automatic  
9 metadata includes date and time, and provides limited usefulness when the archives are  
10 large. Typically, video is annotated with descriptions which may include title, actor(s),  
11 storyline, perhaps even a dialogue script. Publicly available systems such as Open Video<sup>1</sup>  
12 or the Internet Movie Database<sup>2</sup> are examples of systems using such metadata only, and  
13 which have been in widespread use for large closed archives for some time.  
14  
15

16 Because we can now easily capture and store video and upload it to the internet for  
17 sharing we have many systems where the description of shared video is augmented by  
18 user-assigned terms or tags. The Internet Movie Archive<sup>3</sup> and the popular YouTube  
19 system<sup>4</sup> are examples of systems where content description is determined mostly by end  
20 users directly. This can take the form of user-assigned tags or keywords, or can be user  
21 reviews of the video, and all of these are used to provide video retrieval.  
22  
23

24 While content description from user annotation offers useful navigation possibilities, it is  
25 still one step removed from being able to search actual video content directly. Effective  
26 use of user annotation relies on manual effort and consistent annotation, and this is not  
27 scalable for developing good quality content-based access to large quantities of video. In  
28 this paper we concentrate on the application of direct content access to video where user  
29 queries are matched directly against video content. We present three example systems  
30 which demonstrate differing approaches to content-based video search, each developed in  
31 the context of a large scale worldwide benchmarking activity where dozens of video  
32 indexing and retrieval systems are benchmarked on the same video dataset.  
33  
34  
35

36 The rest of this paper is organised as follows. In the next section we summarise the  
37 benefits of a common evaluation carried out across a number of research groups and in  
38 particular, the annual TRECVID activity. This section is included as background and is  
39 followed by an overview of each of three representative video retrieval systems  
40 developed by Dublin City University/K-Space, by the University of  
41 Amsterdam/MediaMill, and by the National University of Singapore, respectively. These  
42 three systems are chosen from over 20 systems for interactive video search developed for  
43 the 2007 TRECVID search task. Each have different approaches to the task of content  
44 retrieval from video. The similarities and differences between these systems, as well as a  
45 report on their respective performance in the TRECVID evaluation, are presented and  
46 discussed in section 4, followed by some overall conclusions.  
47  
48  
49  
50  
51  
52  
53

---

54 <sup>1</sup> [www.open-video.org](http://www.open-video.org)

55 <sup>2</sup> [www.imdb.org](http://www.imdb.org)

56 <sup>3</sup> [www.archive.org](http://www.archive.org)

57 <sup>4</sup> [www.youtube.com](http://www.youtube.com)  
58  
59  
60

## 2. TRECVideo: A Benchmarking Evaluation Campaign for Video Retrieval

Evaluation and common benchmarking is 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)<sup>5</sup> 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 TRECVideo 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. 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 had a great variety of subject matter. The volume of video data used varied each year, with 160 hours of MPEG-1 video used in 2006 for example.

The interactive search task involved applying whatever video analysis and indexing tools each 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 metadata donations made by the research community to the task and these included (for the 2007 TRECVideo cycle alone) a master shot segmentation formatted as MPEG-7, 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. Each of the text description of topics 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.

---

<sup>5</sup> [www.silogic.fr/etiseo](http://www.silogic.fr/etiseo)

1  
2  
3 The shot lists returned by each participant for each topic were pooled together to some  
4 depth, duplicates were removed and shots were manually assessed for relevance by the  
5 TRECVID organizers. Once this ground truth of relevant shots for each topic was  
6 determined, the organizers were then able to compute the absolute performance figures  
7 for the submitted runs in terms of precision and recall as measured against the manually  
8 assessed pooled ground truth.  
9  
10

11 In the interactive search variant, participants were allowed to submit a number of runs  
12 (up to 6 in 2007) where each topic in each run was limited to the shots deemed to be  
13 relevant and found by one person using the participating site's search tool, as found  
14 within a 15 minute limit. This simulated the scenario where a searcher has a limited  
15 timeframe to find as many shots as he/she can where each shot is relevant to a fixed,  
16 unwavering information need. Such a scenario would regularly occur in a newsroom for  
17 example, where a production assistant seeks to locate video footage on a news topic to  
18 present to a news editor for possible inclusion in a broadcast.  
19  
20

21 The systems described in the next section of this paper are from three of the twenty-four  
22 research groups who participated in the interactive search task in TRECVID 2007 and we  
23 describe each system in turn. The three systems were chosen for their variety rather than  
24 their absolute performance characteristics in order to illustrate the capabilities of  
25 contemporary content-based video retrieval systems.  
26  
27  
28  
29  
30  
31

### 32 **3. Three Sample Video IR Systems**

33  
34 Each participant in the TRECVID search task normally addresses some research question  
35 or issue which is of interest to them, and may run more than one variant of their system in  
36 order to submit a number of "runs" which are each assessed manually by the TRECVID  
37 organisers. For each run we can compute retrieval performance figures like precision and  
38 recall and can average these across the set of topics to give an indicative score of the  
39 performance of the system behind each "run". We now describe three systems to  
40 illustrate the capabilities of content-based video retrieval within TRECVID.  
41  
42  
43  
44

#### 45 **3.1 Dublin City University/K-Space Interactive Video Retrieval**

46  
47 The team from Dublin City University led a TRECVID 2007 submission on behalf of the  
48 K-Space consortium, a large European multi-site grouping with an interest in semantic  
49 multimedia information management (Wilkins *et al.*, 2007). Video processing in this  
50 system used every second I-Frame, terming a *K-Frame*, and extracting several low-level  
51 feature descriptors based on the MPEG-7 XM, including colour layout, colour moments,  
52 homogeneous texture, edge histogram and scalable colour. K-Frames were also  
53 segmented into regions using a Recursive Shortest Spanning Tree (RSST) approach  
54 (Adamek and O'Connor, 2007), and the same set of MPEG-7 features extracted for each  
55 region. Several K-Space participants developed several automatic detectors for semantic  
56  
57  
58  
59  
60

1  
2  
3 concepts for each K-frame, including *sports, outdoor, building, mountain,*  
4 *waterscape/waterfront, maps, face detection, 17 classes of audio type, building, car,*  
5 *waterscape-waterfront, desert, road, sky, snow, vegetation, explosion/fire, mountain,*  
6 *camera motion, number of faces visible, weather, US-flag, boat/ship and vegetation.*  
7  
8 These were then combined in the user interface for the system.  
9

10  
11 The DCU/K-Space experiment under investigation in TRECVID was to examine the role  
12 of context in the user interface, where context can be described as showing temporally  
13 adjacent shots. To examine the usefulness of such context, DCU/K-Space designed two  
14 user interfaces, the 'shot based' system, and the 'broadcast based' system. Both systems,  
15 apart from sharing the same retrieval engine, also shared a common query input panel,  
16 topic description panel and saved shot area. The major difference was in the presentation  
17 of the results from the underlying retrieval engine.  
18

19  
20 The 'broadcast based' system takes the idea of context to its maximum by ranking not just  
21 individual video shots, but entire TV broadcasts. This presents an alternative to a shot-  
22 only presentation of results and allows a searcher to explore the temporal neighborhood  
23 of shots. In Figure 1 we see a horizontal line of shots in rows across the results area.  
24 Each row is an entire broadcast, with the best-matching broadcast being the first row.  
25 When a user issues a query, the ranked list of broadcasts is presented, and within each  
26 broadcast, the row will be centered on the highest matching shot within that broadcast.  
27 The coverflow-like interface allows for rapid browsing of shots within a broadcast.  
28

29  
30 Figure 1 shows the user's multimodal query and includes the text "Find shots of a canal,  
31 river or stream with some of both banks visible" which is matched against the translation  
32 of the automatic speech recognition. Also included are two sample query images which  
33 have either been found by the searcher, or form part of the topic definition, and a subset  
34 of the available semantic features, in this case *outdoor*. Query images are matched  
35 against the K-frames from each shot using the same low-level features mentioned earlier  
36 and each of the modalities (text search and image matching) generates a separate ranking  
37 of shots. Using a variation on a query-time weight generation techniques (Wilkins and  
38 Smeaton, 2006), the independent result lists are merged at query time with weights being  
39 assigned to each retrieval expert which approximate that expert's likelihood of providing  
40 the most relevant responses to the query. The semantic concepts are then used as filters  
41 by the user after a content-based query has been issued and these filters can be set to  
42 'positive', 'negative' or 'off'. In Figure 2 we can see that the user's query has moved on  
43 and s/he has found a total of 6 query images but has disabled the semantic concept feature  
44 filtering of *outdoor*.  
45  
46  
47  
48  
49

### 50 51 **3.2 University of Amsterdam/MediaMill ForkBrowser**

52  
53 The MediaMill team at the University of Amsterdam departed from the traditional cycle  
54 of query-browse-query by providing users with video browsers that allow to visualize the  
55 entire data set in multiple dimensions, thus facilitating interactive exploration. For  
56 TRECVID 2007, the focus was specifically on consolidation of proven interface  
57  
58  
59  
60



1  
2  
3 components from previous TRECVID editions into a novel browsing environment (Snoek  
4 *et al.*, 2007)  
5  
6

7 The notion of threads was introduced in the ForkBrowser in order to browse through a  
8 video data set in multiple directions. A thread is a linked sequence of shots in a specified  
9 order, based upon an aspect of their content (de Rooij *et al.*, 2007) including *static*  
10 *threads* which are pre-computed, and *dynamic threads* which are generated on demand.  
11 The content of a thread is based on a form of similarity between shots in the data set.  
12  
13

14 The combination of a time thread with any other thread resulted in the CrossBrowser  
15 which proved effective for the TRECVID interactive search tasks in 2004 and 2005 when  
16 a single thread was sufficient for the user to find shots which satisfy a topic or  
17 information need (Snoek *et al.*, 2007), (Snoek *et al.*, 2006). For topics that require a  
18 combination of threads, the RotorBrowser was introduced in 2006 (de Rooij *et al.*, 2007),  
19 (Snoek *et al.*, 2006). This allows a user to integrate query results with time, visual  
20 similarity, semantic similarity and various other shot-based similarity metrics. While  
21 effective, this visualization proved overwhelming for non-expert users. To leverage the  
22 benefits of having multiple query methods while simultaneously allowing the user to  
23 maintain an overview of their results, an interface was introduced in TRECVID 2007  
24 which combines query by keyword, query by example, query using 572 semantic  
25 concepts, query by time and by program, all combined into a framework which is called  
26 the ForkBrowser.  
27  
28  
29

30 The ForkBrowser visualizes results by displaying keyframes based on the shape of a fork.  
31 The contents of the tines of the fork depend on the shot at the top of the stem. The center  
32 tine shows unseen query results, the leftmost and rightmost tines show the time thread,  
33 and the two tines in between show user-assignable threads. For the TRECVID 2007  
34 benchmark two variants of visual similarity threads are displayed. The stem of the fork  
35 displays the history thread. Every displayed key frame is taken from a single video shot,  
36 and the video shot can also be played on demand by rapidly displaying up to 16 frames in  
37 sequence from the originating shot. This helps in rapidly answering queries containing  
38 explicit reference to motion or to events. Figure 3 depicts the ForkBrowser while  
39 searching for “boats moving past”. The horizontal tine shows shots from the time thread  
40 of the program “Klokhuis”, the diagonal directions depict two visual threads to provide  
41 the user with similar shots from waterscapes which s/he can browse.  
42  
43  
44  
45  
46

### 47 3.3 NUS-ICT/VisionGo

48  
49 **VisionGo** is an interactive video retrieval system developed jointly by the National  
50 University of Singapore (NUS) and the Institute of Computing Technology, Chinese  
51 Academy of Sciences (ICT). The system is designed to maximize the effectiveness of  
52 human annotators through the use of an intuitive User Interface (UI), options for multiple  
53 feedback strategies and motion icons. In performing an interactive search, the system first  
54 uses results from an automated search based on the user’s multimodal query, followed by  
55 multimodal fusion to retrieve a ranked list of shots. The fusion uses a combination of text  
56  
57  
58  
59  
60

1  
2  
3 derived from ASR (automatic speech recognition), high-level features (HLFs)  
4 automatically detected in shots, and a combination of low-level visual features and  
5 motion (Chua *et al.*, 2007). The user then makes use of an intuitive retrieval interface  
6 with a variety of relevance feedback options to refine their search results. In addition,  
7 motion-icons are introduced which allow users to see a dynamic series of keyframes  
8 instead of a single keyframe during relevance assessment.  
9  
10

11 To maximize the user's interaction efforts, the intuitive UI is designed for fast keystroke  
12 actions with quick previews of previous and subsequent sets of shots in the ranked list of  
13 shots. A sample interactive UI is shown in Figure 4. The UI is inspired by high throughput  
14 interactive game interfaces, which are mainly keystroke based. The UI displays three  
15 images at a time in a central active row, with the previous and next rows in view. Each  
16 image corresponds to a single retrieved shot, without any *context* such as previous or  
17 following shots.. The user will determine the images' relevance to the query and annotate  
18 the positive ones by hitting pre-a defined set of keys on the keyboard. The system captures  
19 the user's input and automatically refreshes itself to display the next row of new keyframes  
20 in the ranked list. In experiments at the National University of Singapore, the UI enabled  
21 a normal user to annotate up to 3,500 shots based on motion icons or 5,000 shots based  
22 on static icons, in only 15 minutes.  
23  
24  
25  
26

27 To allow for more flexibility and to provide a range of options for users to click during  
28 relevance feedback, interactive feedback is segregated into three distinct types, namely  
29 recall-driven, precision-driven and temporal locality-driven feedback. Each strategy aims  
30 at leveraging different aspects of user feedback data. At any time, if the user feels that the  
31 search and feedback process is not progressing well, he/she is able to select any other  
32 feedback strategy to enhance search performance.  
33  
34

35 Recall-driven feedback employs general features such as the ASR text tokens and HLFs  
36 from relevant shots to perform query expansion. This option has been found to be the  
37 most effective in finding many new relevant shots in the initial stage of a search. Given  
38 the set of positively annotated shots, this process makes use of text and HLF scores to  
39 iteratively adjust the retrieval function. Precision-driven feedback uses motion, visual  
40 and audio features in an SVM-based active learning environment targeting at improving  
41 precision. It uses active learning to provide long term improvements to classifiers. Fused  
42 with a performance-based adaptive sampling strategy, this process continuously re-ranks  
43 instances as the user annotates shots as relevant or non-relevant. Finally, temporal  
44 locality-driven feedback essentially returns shots from neighboring shots from the  
45 positively labeled set, as it is found that positive shots tend to cluster near each other  
46 within the same story. Based on these multiple feedback strategies, a user is able to  
47 choose the type of feedback that is more suitable based on his/her intuition or experience,  
48 in order to maximize performance.  
49  
50  
51  
52

53 Many visual-oriented queries tend to be associated with objects in motion in the video. It  
54 is therefore necessary to provide some information on motion within each shot.  
55 Specifically, we construct a summarized clip comprising a sequence of progressive  
56 keyframes which can show moving picture information. We call this a motion icon or  
57  
58  
59  
60



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 more confidence.

#### 4. A Comparison of Performance of the Three Search Systems

Since one of the purposes of the TRECVID benchmarking activity is to compare the performance of various content-based video retrieval approaches, in this section we summarise how the three systems introduced earlier performed in their official TRECVID submissions. It is important to remember that each of the TRECVID participating groups set out to examine some research question and achieved this by comparing performance from among their allowed six runs.

There are many topical research areas within the broad area of content-based video retrieval including (semantic) concept detection, query formulation, algorithms for image/keyframe matching, fusion of individual retrievals (colour, texture, edge based, text based, concept based, etc.), browsing interfaces and the issue of how to effectively incorporate relevance feedback into the user experience. In fact successfully exploiting relevance feedback into a retrieval interface and presenting it as an integral part of the user experience is a major challenge in content-based video retrieval and an issue that the NUS system concentrated upon as its research question. Relevance feedback is a hugely important aspect of the user interaction in video retrieval because video is so rich in terms of content, and any video search system will generally require much interaction with its enduser, through relevance feedback and video browsing, in order for the user to locate relevant shots and so it is an area of much research activity as can be seen in the overview in (Zhou and Huang, 2002) and more recent work, for example in (Tao *et al.*, 2008).

Within the set of up to six runs allowed from each site participating in TRECVID, each group can control the one variable, the endusers, which cannot be controlled in performance comparisons across sites. What this means is that it is acceptable to compare precision and recall figures within a site's runs, but comparisons across sites will have an uncontrolled variable as users will vary in their levels of expertise, experience, or even in their level of motivation for performing the searches.

Notwithstanding the above caveat, there is value for us here in examining the performance of the best runs from each site but only in looking at the absolute values. Figure 7 shows the precision-recall figures for these runs and shows that the three systems are quite comparable – the codes in the legend refer to the official TRECVID run names used to distinguish participants and their submitted runs. What is most noteworthy from the point of view of this paper is the performance of the three systems at the high precision end of the scale, corresponding to the 'early' parts of the user's search. Here we can see that performances for all of the systems are very effective, meaning that each of these systems, and many of the others developed in TRECVID, provide effective tools for helping users locate relevant shots, and these are all based on content-based searching.

## 5. Content-Based Video Retrieval Conclusions

The three interactive video search systems presented in this paper are both similar and different to each other. 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 then fusing their outputs together at search time. Each supports a preview of a whole shot by presenting sets of keyframes, called *micons* by NUS and K-Frames by K-Space, 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 experiences 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. Collectively, however, what the three systems demonstrate is that there are now many systems which can provide effective content-based retrieval of video shots from archives of several hundreds of hours of video content, in a fast, effective and user-friendly manner. TRECVID search systems represent the state-of-the-art in content-based video searching yet this is not mainstream in terms of usage by a large population of users. The techniques needed to realize a widespread deployment of this, such as an internet-scale deployment, are under development and represent one of the largest challenges in this field.

### Acknowledgements

AS and PW wish to acknowledge support from Science Foundation Ireland under grant 03/IN.3/I361.

### 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.

1  
2  
3 Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang. Towards Optimal Bag-of-Features for  
4 Object Categorization and Semantic Video Retrieval. ACM International Conference on  
5 Image and Video Retrieval (CIVR'07), Amsterdam, The Netherlands, 2007.  
6

7  
8 P. Joly, J. Benois-Pineau, E. Kijak, and G Quénot. The Argos Campaign: Evaluation of  
9 Video Analysis Tools. In Proceedings of the International Workshop on Content-Based  
10 Multimedia Indexing, 2007. CBMI'07, pages 130-137, 2007.  
11

12  
13 N. Lazarevic-McManus, J. Renno, J. and G.A. Jones, G. A. 2006. Performance  
14 evaluation in visual surveillance using the F-measure. In Proceedings of the 4th ACM  
15 international Workshop on Video Surveillance and Sensor Networks (Santa Barbara,  
16 California, USA, October 27 - 27, 2006). VSSN '06, 45-52.  
17

18  
19 P. Over, G. Awad, W. Kraaij and A.F. Smeaton. TRECVID 2007 - An Introduction. In  
20 Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.  
21

22  
23 G.M. Quénot. Active learning for multimedia. In Proceedings of the 15th international  
24 Conference on Multimedia (Augsburg, Germany, September 25 - 29, 2007).  
25 MULTIMEDIA '07. ACM Press.  
26

27  
28 O. de Rooij, C.G. Snoek and M. Worring. Query on demand video browsing. In  
29 Proceedings of the 15th international Conference on Multimedia (Augsburg, Germany,  
30 September 25 - 29, 2007). MULTIMEDIA '07. ACM Press, 811-814.  
31

32  
33 A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In  
34 Proceedings of the 8th ACM International Workshop on Multimedia Information  
35 Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press,  
36 321-330.  
37

38  
39 C. G. M. Snoek, J. C. van Gemert, Th. Gevers, B. Huurnink, D. C. Koelma, M. Van  
40 Liempt, O. De Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H.C.  
41 Thean, C. J. Veenman, M. Worring. The MediaMill TRECVID 2006 Semantic Video  
42 Search Engine. In Proceedings of TRECVID 2006, Gaithersburg, Md., November 2006.  
43

44  
45 C.G.M. Snoek, M. Worring, D.C. Koelma and AWM Smeulders. A Learned Lexicon-  
46 Driven Paradigm for Interactive Video Retrieval. IEEE Transactions on Multimedia,  
47 9(2), pages 280-292, 2007.  
48

49  
50 C.G.M. Snoek, I. Everts, J.C. van Gemert, J.M. Geusebroek, B. Huurnink, D.C. Koelma,  
51 M. van Liempt, O. de Rooij, K.E.A. van de Sande, A.W.M. Smeulders, J.R.R. Uijlings  
52 and M. Worring. The MediaMill TRECVID 2007 Semantic Video Search Engine. In  
53 Proceedings of TRECVID 2007, Gaithersburg, Md., November 2007.  
54

55  
56 D Tao, X Tang and X Li. Which Components are Important for Interactive Image  
57 Searching ? IEEE Transactions on Circuits and Systems for Video Technology, 18(1),  
58 pages 3-11, 2008.  
59  
60

1  
2  
3  
4  
5 P. Wilkins, P. Ferguson, and A. F. Smeaton. Using score distributions for querytime  
6 fusion in multimedia retrieval. In Proceedings of the 8th ACM International Workshop  
7 on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27,  
8 2006). MIR '06. ACM Press, 51-60.  
9

10  
11 P. Wilkins *et al.* K-Space at TRECVID 2007. In Proceedings of TRECVID 2007,  
12 Gaithersburg, Md., November 2007.  
13

14 X. Zhou and T.S. Huang. Relevance Feedback in Content-Based Image Retrieval: Some  
15 Recent Advances. *Information Sciences Applications*, 148(1-4), 2002, 129-137.  
16  
17  
18  
19  
20

21 Figure 1: User interface for Dublin City University K-Space Search System  
22  
23

24 Figure 2: User interface for Dublin City University / K-Space Search System  
25  
26

27  
28 Figure 3: User interface for University of Amsterdam's Search System  
29  
30

31  
32 Figure 4: User interface for National University of Singapore's VisionGo Search System  
33  
34

35 Figure 5: A sequence of multiple keyframes for shot213\_62  
36  
37

38  
39 Figure 6: A sequence of multiple keyframes for shot149\_62  
40  
41

42  
43 Figure 7: Performance figures for three systems from TRECVID  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**Figure 1: User interface for Dublin City University K-Space Search System**  
451x287mm (72 x 72 DPI)

Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

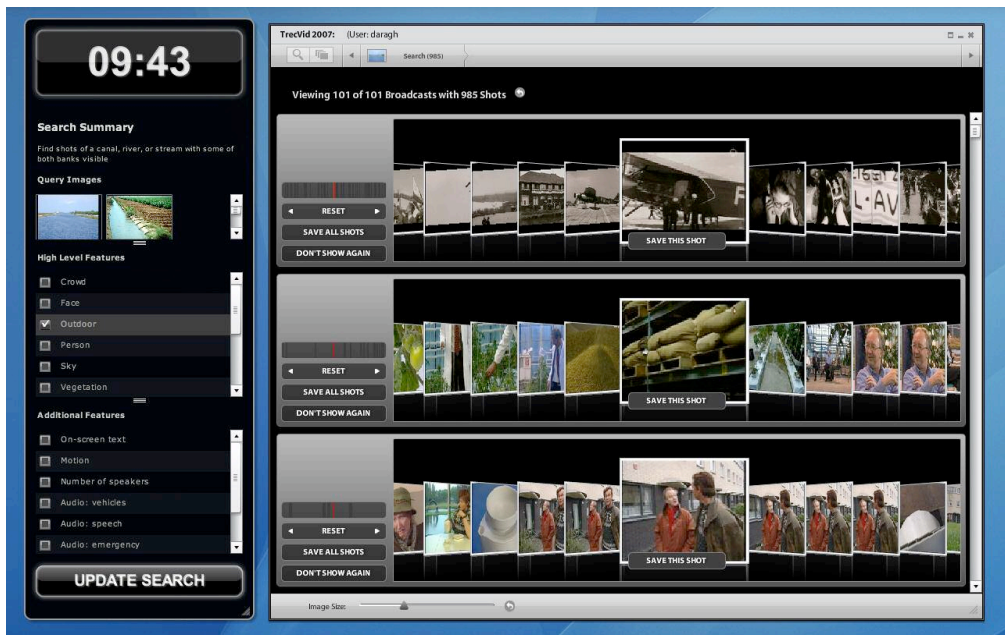
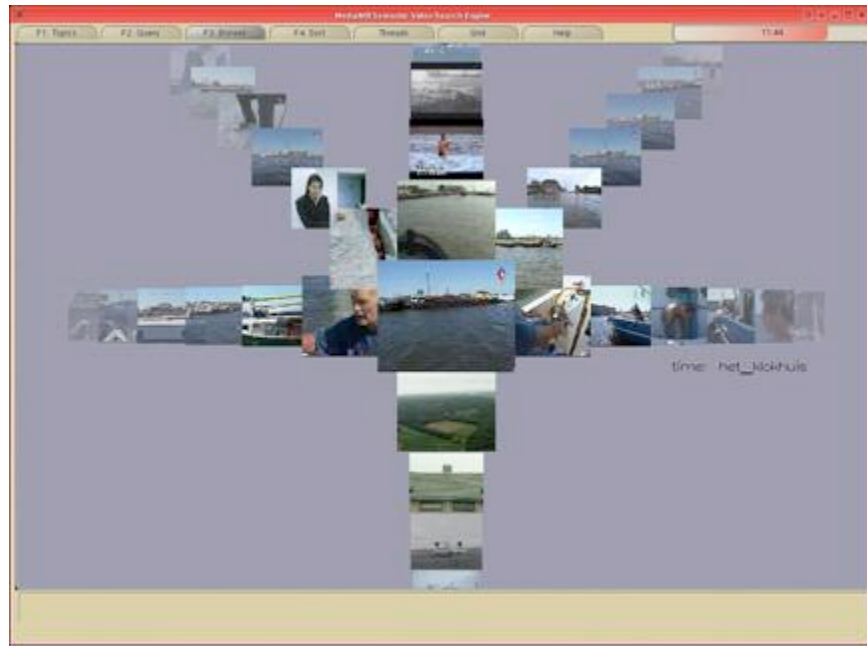


Figure 2: User interface for Dublin City University / K-Space Search System  
452x285mm (72 x 72 DPI)

Review



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**Figure 3: User interface for University of Amsterdam's Search System**  
152x113mm (72 x 72 DPI)

Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Figure 4: User interface for National University of Singapore's VisionGo Search System  
150x108mm (72 x 72 DPI)



**Figure 5: A sequence of multiple keyframes for shot213\_62**  
163x26mm (72 x 72 DPI)

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**Figure 6: A sequence of multiple keyframes for shot149\_62**  
163x26mm (72 x 72 DPI)

For Peer Review



219x147mm (72 x 72 DPI)

Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60