Interactive Searching and Browsing of Video Archives: Using Text and Using Image Matching

Alan F. Smeaton, Cathal Gurrin, and Hyowon Lee

Centre for Digital Video Processing and Adaptive Information Cluster,
Dublin City University, Glasnevin, Dublin 9, IRELAND.

1 Introduction

Over the last number of decades much research work has been done in the general area of video and audio analysis. Initially the applications driving this included capturing video in digital form and then being able to store, transmit and render it, which involved a large effort to develop compression and encoding standards. The technology needed to do all this is now easily available and cheap, with applications of digital video processing now commonplace, ranging from CCTV (Closed Circuit TV) for security, to home capture of broadcast TV on home DVRs for personal viewing.

One consequence of the development in technology for creating, storing and distributing digital video is that there has been a huge increase in the volume of digital video, and this in turn has created a need for techniques to allow effective *management* of this video, and by that we mean content management. In the BBC, for example, the archives department receives approximately 500,000 queries per year and has over 350,000 hours of content in its library[1]. Having huge archives of video information is hardly any benefit if we have no effective means of being able to locate video clips which are of relevance to whatever our information needs may be.

In this chapter we report our work on developing two specific retrieval and browsing tools for digital video information. Both of these are based on an analysis of the captured video for the purpose of automatically structuring into shots or higher level semantic units like TV news stories. Some also include analysis of the video for the automatic detection of features such as the presence or absence of faces. Both include some elements of searching, where a user specifies a query or information need, and browsing, where a user is allowed to browse through sets of retrieved video shots. We support the

[1] Evans, J. The future of video indexing in BBC, at TRECVid Workshop 2003, Gaithersburg, MD, 18 November 2003.

presentation of these tools with illustrations of actual video retrieval systems developed and working on hundreds of hours of video content.

2 Managing Video Archives

The techniques we use in the tools described in this chapter represent just some of the available approaches to managing digital video information. In the first instance, the easiest, and most useful way to organise video archives is to use raw metadata, created at the time the video is created, to index and provide subsequent access to the video. In the case of CCTV, for example, to access video at a given point, security personnel employ a combination of which camera, and what date and time, and this is usually sufficient to allow users to retrieve the video clips they are looking for. If a more refined or accurate content-based search is required then the raw metadata will not be enough and many archive libraries will annotate video content by hand. This can take up to 8 or 10 times real-time (i.e. 8 to 10 hours to hand-annotate 1 hour of original video) and is thus clearly very expensive but is used extensively in TV archives worldwide. To ensure some consistency across annotators and across time, they typically each use an ontology of only some thousands of terms which creates a structured relationship among the pre-defined set of index terms. Figure 1 shows an example of such an annotation system, used as part of TRECVid (see section 4.1) for annotating broadcast news video [16], where the video shot currently being annotated is being assigned the "tags" *standing, outdoors, trees, greenery, water body, waterfall, microphone, female speech and female face*. These annotations can subsequently be used in searching or browsing.

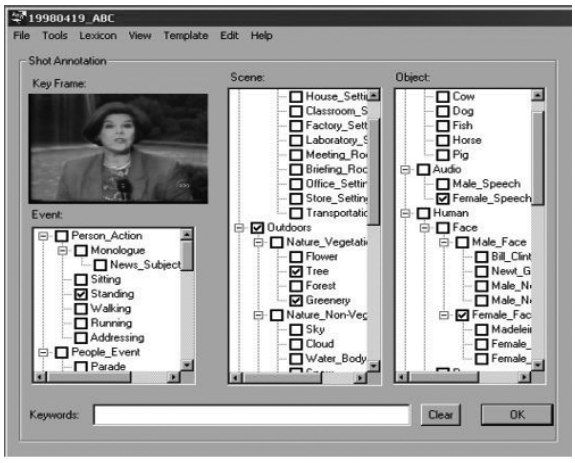

Fig. 1. Sample Manual Annotation of Video

Manual annotation is, naturally, very expensive and used only where there is a specialist need for high-quality searching, but it is not scalable to huge video archives. In the case of a user accessing an archive of broadcast TV or movies on a home DVR to find the exact clip in the movie "Minority Report" where Tom Cruise uses a video browsing system with a gesture-based interface, then manual annotation of the video content will not be available. The movie name or date/time of transmission and recording, and some clue as to how far into the movie that scene occurs will probably be enough only to locate the region of the movie where the scene occurs. The user is probably then going to need to *browse* through the video to locate the exact scene. Consider also the case of a user travelling on business and accessing an archive of tonight's TV news from their local TV station via a web interface to a video archive system. The user doesn't want to play the full 30 minutes of news but will want to *browse* through the stories, skipping those not of interest based on a story skip, possibly of video keyframes or of dialogue, and playing video clips of those stories of interest. Conventional VCR-type controls like play, pause, fast-forward and rewind can be used here, as can more intelligent approaches such as pause detection and removal and variable speed fast forward [21].

However, there is a lot more that can be done to help a user locate desired conent, for example the Físchlár-TV [26] system which is a web-based shared video retrieval system that lets users record, *browse* and playback television programmes online using their web browser. A programme recorded by one user enters a shared repository and can then be viewed by any other user of Físchlár-TV. The total video archive size is about 400 hours of video and operates as a first-in first-out queue which usually results in a programme being available for just over three weeks before being removed from the archive to make way for newly recorded programmes. TV schedules are used to allow a user to record a programme by simply clicking on a hyperlink. By default, all programmes in the archive are sorted by date and time in decreasing order of freshness and are listed in on the left side of the interface (see Figure 2). Selecting a programme title will cause full programme details to be displayed on screen. Each recorded programme is represented by metadata (title, date, time, and a short description) and video keyframes which are extracted automatically from the programme and presented on screen for the user. In this way a user can *browse* through the content of a programme, seeking a desired section and when the desired section is found (for example, that scene in Minority Report) clicking on the keyframe begins playback of video from that point.

While a video retrieval system such as Físchlár-TV is clearly very useful, in allowing a user to quickly browse an entire TV programme by examining a collection of keyframes, the user still needs to know which programmes to browse. However, when presented with a large archive of content, a user will, in many cases, be unsure of what programme they are looking for, for example, a video archive that contains all recordings of a late night chat show, how can a user, without knowing the date of broadcast, find the interview with

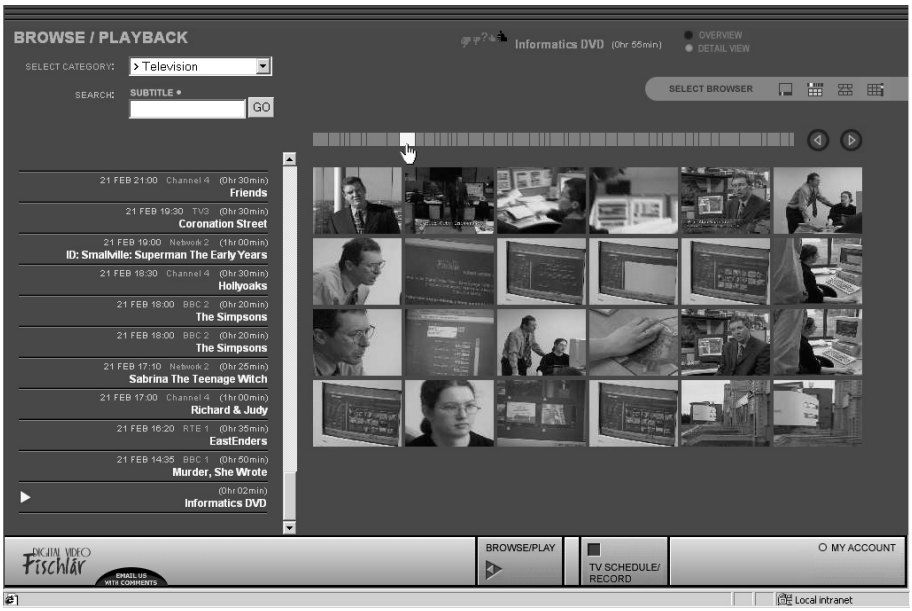

Fig. 2. Browsing a recorded video programme in the Físchlár-TV System

Madonna where she throws a glass of water in the host's face ? In this case there is a need for video *searching* through the actual video content, using some keywords from the dialogue between Madonna and the chat show host [24] or if CCTV video, using some representation of the object corresponding to the CCTV suspect [14].

¿From this introduction to interactive searching and browsing of video archives we can already see that there are at least three separate ways in which we may want to access digital video information; using raw or annotated *metadata* as a basis for searching, *browsing* through the actual video content and *searching* through the actual video content. Other content-based access tools could include summarisation, automatic gisting and highlight detection but we are not concerned with those in this chapter. Neither are we concerned here with techniques for searching through raw metadata. Instead we concentrate here on techniques for supporting interactive searching and browsing video content based on using text and image matching.

As a result of extensive research in the very recent past there are now robust, scalable and effective techniques for video analysis and video structuring which can turn unstructured video into well-formed and easy to manage video shots. There are also semi-automatic techniques for video object extraction, tracking and classification though these are not yet scalable to large video archives. There are good techniques available for recognising features in video, from simple features such as indoor/outdoor and faces/no faces present, to the more challenging naming of individual faces or naming of buildings and

locations. Many of these techniques have been developed for the purpose of automatic video analysis on large video archives which can, in turn, support searching and browsing. Ideally state-of-the-art systems such as Informedia from CMU [10] or MARVEL from IBM Research [8] would be able to manage tens or hundreds of thousands of hours but as of now they are able to manage hundreds or maybe just thousands of hours of video.

The Informedia System, developed as part of ongoing research at CMU since 1994 has developed and integrated new approaches for automated video indexing, navigation, visualisation, search and retrieval from video archives. In a similar way to the two retrieval systems detailed later in this chapter, the Informedia system brings together various strands of video retrieval research into one large system that provides retrieval facilities over news and documentary broadcasts (from both TV and radio) in a one terabyte video archive. The Informedia system combines speach recognition, image feature extraction and natural language processing technologies to automatically transcribe, segment and index digital video content. Key features of the Informedia system include the extraction of name and location data from the videos, face identification, summary generation, dynamic video linkage, event characterisation and novel visualisation techniques.

MARVEL (Multimedia Analysis and Retrieval System), from IBM research is a prototype multimedia analysis and retrieval system, the aim of which is to automatically annotate multimedia (not just video) data by using machine learning techniques that model semantic concepts. MARVEL organises semantic concept ontologies and automatically assigns concept labels to video data, thereby reducing the need for human annotation of content from 100% to only 1-5% (which is required for the machine-learning processes to operate effectively). MARVEL is migrating from a current ontology of about 100 concepts to a large scale ontology of 1,000 concepts, designed to model broadcast news video. Given the ability of MARVEL to automatically determine concepts occurring in video data, a user search system incorporates this functionality along with text search functionality to produce a powerful multimedia retrieval system.

The video analysis techniques used in such video retrieval systems, and the subsequent video searching and browsing, are the focus of the work reported in this chapter, with the rest of this chapter being organised as follows. In the two sections to follow we provide system descriptions of the Físchlár-News and Físchlár-TRECVid systems developed for accessing an archive of RTÉ broadcast TV news in the case of Físchlár-News and for accessing an archive of CNN and ABC TV news in the case of Físchlár-TRECVid. In section 5 we illustrate how each system supports both searching and browsing in different ways because the information needs which each was designed to address are very different. Despite the very different nature of the underlying information needs, and the resulting systems, we are able to show how the searching and browsing operations in both systems are very tightly intertwined

in each system, which underscores the main point of this chapter which is to stress how equally important search and browse are for video navigation.

3 Físchlár-News: System Description

Físchlár-News is an online archive of broadcast TV news video which makes use of various content-based video indexing techniques to automatically structure TV news video to support searching, browsing and playback of the news video on a conventional web browser. An example usage scenario would be a user who is travelling and wishes to keep up with news events at home, but who does not have the time to view an entire news programme, rather would like to be able to view news stories from both missed news programmes and the entire archive.

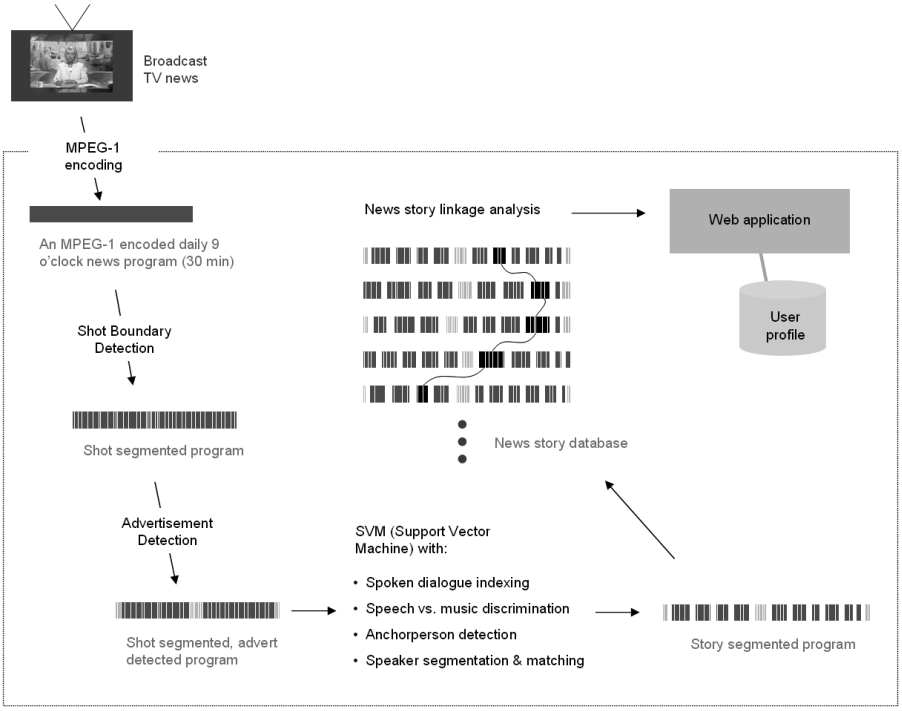

Fig. 3. Automatic Processing of News Video in the Físchlár-News System

The system's automatic processing of video is illustrated in Figure 3. At 9 o'clock every evening, Físchlár-News records the TV news from the Irish national station RTÉ into MPEG-1 (top-left of Figure 3), along with the

closed-caption data (spoken dialogue text supplied by the broadcaster) transmitted at the same time. The encoded MPEG-1 file goes through a series of automatic content-based indexing processes, starting with *shot boundary detection* [2] which segments the news video into individual camera shots, followed by *advertisement detection* [22] which removes TV ads that sometimes appear at the beginning, the end, and in the middle of the broadcast news. The ads-removed parts of the video are then subject to *news story segmentation* [17] which involves a number of content analysis techniques including speech/music discrimination and anchorperson detection, and their output combined using a machine-learning technique to determine more reliable news story boundaries. The system requires these multiple evidence combinations at this stage because accurate news story segmentation is still a major challenge in the video retrieval community with various approaches being tried [4, 13]. The closed-caption signal is also indexed with conventional IR techniques and aligned with the corresponding video data to support text-based searching. The outcome of all this is that a day's broadcast TV news is automatically structured into topical, individual news stories each of which is again segmented into a number of shots. Once this stage is reached, the structured video is stored in the news story database in which all previous days' news stories have been indexed and are available for retrieval (top-right of Figure 3). Currently this database contains over 3 years of news, amounting approximately 8,000 news stories. These news stories are available via a conventional web browser, allowing news story-based searching, browsing and playback.

Figure 4 shows a screen shot of the web interface. On the left side of the screen the monthly calendar allows access by date. When a user clicks on a date, that day's news stories are presented on the right side of the screen, each story with an anchorperson keyframe, and first two lines of closed-captions. In Figure 4 a user is searching for stories related to politician Paul Bremer. The user typed in 'Bremer' in the query box and clicked on the GO button. The search term was matched against the indexed closed-captions and the resultant news stories returned. In Figure 4 five news stories were retrieved as a result and presented on the right side of the screen. The user can simply play any of the retrieved stories by clicking on the 'Play this story' button at the end of each closed-caption summary which will pop up a video player plug-in and start playing the story, or browse more detail of the story by clicking on the title of the story. Figure 5 shows a screen when the user selected a fourth story in the retrieval result from Figure 4 (story dated 23 August 2003) to browse more detail. The user is then presented with a "storyboard" of the story, a full list of keyframes from each camera shot contained in the story with interleaved closed-caption text, providing detail of the story for quick browsing (right side of Figure 5). The shot-level storyboard shows all major visual content of that story in one glance, without requiring a video playback, and is featured in the majority of digital video retrieval systems available today. Clicking on any of these keyframes will pop up the player plug-in and

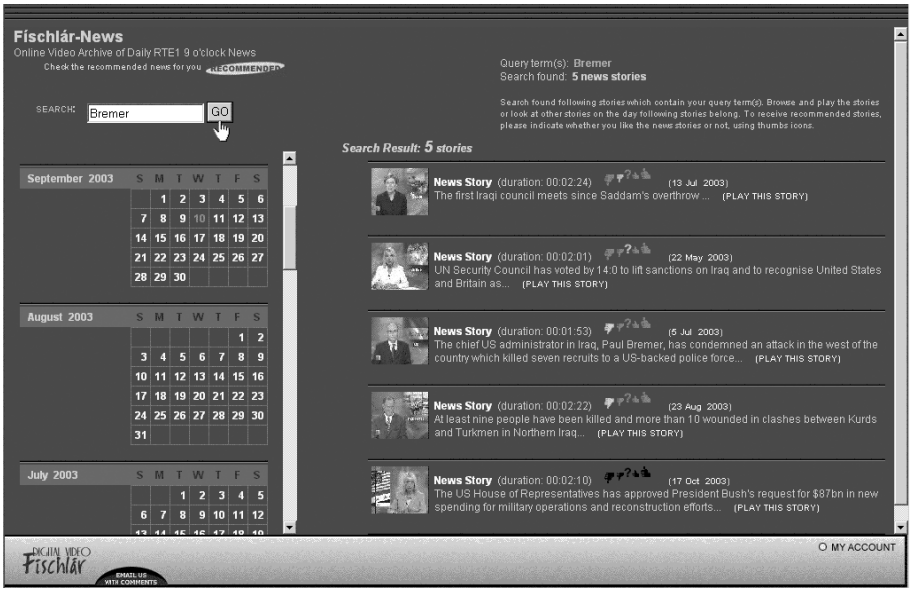

Fig. 4. Text-based Searching and List of Stories as Search Result

start playing from that point in the story onwards, enabling playback at a particular point within a story.

Fig. 5. Browsing Shot-Level Detail of a News Story

Fig. 6. Browsing Related Stories

The user can also browse through other similar stories within the archive, to trace the development of the story in last few months or to browse other stories related in some way (e.g. the same company is involved or same person mentioned) for more serendipitous browsing. Below the storyboard of a news story with keyframes and closed-captions, the user can see the ten stories that are related to that story as shown in Figure 6. This list of "related stories" is generated by taking the closed-captions of the currently opened story as query text and the top ten results arereturned at browsing time. Thus, in Figure 6 the related stories shown may be again stories about Paul Bremer but may be about something else. In this way, the user can start by browsing a news story detail, followed by jumping between related stories that the system automatically generates links to. Users can also access the news stories by automatic recommendation in which they indicate their preference for a particular news story using a 5-point thumbs-up and -down scale icons located beside each story, and as this information from the users accumulates over time the system can recommend some of the newly appeared stories as well as older stories in the archive to individual users by way of collaborative filtering [18].

A more elaborate presentation and interaction scheme to support an effective news story navigation would be possible on top of the interface that current Físchlár-News features, for example a timeline presentation of related stories [28], visualising the topic thread over time [9], use of clustering techniques to visualise clusters of topically related stories and to highlight recently

added stories in each of the clusters [6], use of time and locality to automatically generate an interactive collage of images, video and text of news stories presented with timeline and map [3], and visualisation of news themes in a virtual thematic space where a user can navigate through appearing and disappearing textual themes in a highly interactive way [20]. These potentially useful presentation techniques will require more investigation on their usability before deployed and used by real users.

Físchlár-News has been operational on our University campus for over 3 years continuously capturing, processing and archiving daily TV news, to support media and journalism students and staff as well as other users who want casual news updates during the day. Apart from serving as an experimental platform for the various content-based analyes described above, the value of the system over watching the news on TV is that using these techniques the system automatically turns the sequential, time-based medium of many hours of news video (currently several hundred hours of video content) into an easily browseable and searchable commodity which allows convenient story-based access at any time. More details of technical aspects of the system and its envisaged usage scenario have been drawn in [25], and a long-term user study on people's actual usage of the system in the workplace can be found in [15]. There are many commercial online news websites that are highly up-to-date and feature photos and video footage with links to related stories, but their human indexing and authoring means a high cost of manual work and a difficulty in maintaining indexing consistency among indexers over time.

4 Físchlár-TRECVid: System Description

The Físchlár-News system, that we have just described, is an operational video retrieval system with a campus-wide user base, hence the user interaction supported is clearly defined and easy for any user to understand and operate. Físchlár-TRECVid, on the other hand, is an experimental retrieval system, the aim of which is to evaluate alternative techniques to interactive video search and retrieval. This evaluation is conducted annually as part of the TRECVid workshop.

4.1 TRECVid: Benchmarking Video information Retrieval

The history of (text-based) information retrieval is one where empirical investigation and experimentation has always been fundamental. Information retrieval draws its background from a combination of computer science, information science, engineering, mathematics, human-computer interaction and library science and throughout its 40 years of history theoretical improvements have always had to be validated in experiments before being accepted to the IR community. This philosophy has led to the emergence of the annual TREC (Text REtrieval Conference) exercise which, since 1991, has facilitated the

comparative evaluation of IR tasks in an open, metrics-based forum. TREC is truly global with 100+ participating groups in 2004.

In 2001, TREC featured yet another "track" or activity, on tasks related to video information retrieval, including shot boundary detection, feature extraction, and interactive searching. This has now spun out as a separate independent exercise known as TRECVid [29, 27]. The operation of TRECVid, and TREC, revolves around the organisers, National Institute of Standards and Technology (NIST) gathering and distributing video data to signed-up participants (60+ participants in 2005). In 2001 this was only 11 hours of video content and in 2005 it is 200+ hours of video. Although the tasks of shot and story bound detection and of feature identification are of importance to video IR, we are interested in the interactive search task here, where users are given a multi-modal (multiple media) topic as an expression of an information need and a fixed time window (15 minutes usually) to complete the task of finding as many shots likely to be relevant to the topic as they can. Note that the user task is to locate shots, not news stories, which are likely to be relevant.

TRECVid participants use the same video data, run the same topics (descriptions of an information need) against this using their own systems and then the relative performance of systems/groups in terms of retrieval effectiveness is measured. The TRECVid excercise has participation from dozens of research groups worldwide and is a true benchmark of the effectiveness of different approaches to video retrieval. When NIST receive the identified shots from each participating site and for each topic, these are then pooled together and duplicates eliminated. Remaining shots are then presented to assessors who examine each retrieved shot and then make a binary judgement as to relevance. This establishes the ground truth for each of the topics and with this information available, the organisers are then able to measure the performances of the retrieval runs submitted by each participating site and compute retrieval performance figures in terms of precision and recall, for each.

In the 2004 edition of TRECVid, the video data distributed to participating groups was broadcast TV news, from CNN and ABC. The interactive search task was to retrieve *shots* which matched the topic, not news stories and the nature of the search topics illustrates this. Topic 144 asks users to "Find shots of Bill Clinton speaking with at least part of a US flag visible behind him" and Figure 7 shows the 2 images and keyframes from the 2 video clips which form part of the topic definition. Associated with the TRECVid 2004 broadcast data there were three types of text information distributed to participants; the original closed-captions which give an accurate summary but not an exact replication of the dialogue as spoken in the video, the output from an automatic speech recognition system, and "video OCR" which corresponds to the character recognition of any text appearing in the video frame, such as a sub-title or text overlay.

Finally, several groups evaluated their own feature (semantic concept) detection algorithm(s) for each of the 33,367 shots in the collection and submit-

(a) Sample image

(b) Sample image

(c) Sample video keyframe

(d) Sample video keyframe

Fig. 7. Sample images ((a) and ((b)) and keyframes from sample video clips ((c) and (d)) for topic 144 from TRECVid 2004.

ted the results of their analysis to NIST as part of one of the TRECVid tasks, but also made the results of these feature detections available to other participants for use in their own search systems (distributed in MPEG-7 format). The features whose detection performance was evaluated and whose results were used in some TRECVid search systems, including our own, were boats or ships of any type, the presence of Madeleine Albright, the presence of Bill Clinton, trains or railroad cars, a beach with the water and the shore visible, a basketball score with the ball passing down the hoop and into the net, an airplane taking off, people walking or running, physical violence between people and/or objects, and finally, a road of any size, paved or not. Some of these are really difficult to do and represent very challenging tasks while others are more achievable. Some of the donated features were used by us in the system described below and by other participating groups in their interactive retrieval experimentation.

In only 5 years of operation TRECVid has grown considerably in terms of the data volume, the number of groups taking part, the tasks being evaluated, the measures used and the complexity of the whole exercise. It is within this framework that we developed a version of our Físchlár system for TRECVid in 2004, which we call Físchlár-TRECVid, which we now describe.

4.2 Físchlár-TRECVid

In Físchlár-TRECVid, search and retrieval of video shots is supported as well as the browsing of news programmes at the shot level. The shot retrieval engine employed for Físchlár-TRECVid is based on a combination of query text matched against spoken dialogue combined with image-image matching where a still image (sourced externally), or a keyframe (from within the video archive itself), is matched against keyframes from the video archive. The image matching is based on low-level features taken from the MPEG-7 eXperimentation Model (XM) [30].

Unlike Físchlár-News, which operates successfully over the closed-caption text alone, Físchlár-TRECVid employs three sources of text data to support shot search and retrieval namely closed-captions, ASR text and video OCR. It has been shown [5] that the integration of multiple sources of text improves retrieval performance over using a single source of text transcripts alone, when operating over TV news programmes. For example, the addition of closed-caption and OCR text to an existing automatic speech recognition transcript-only retrieval engine improves searching performance by 17% (MAP) and the number of relevant shots found for a typical query by 18%.

As stated earlier, the visual shot matching facilities were primarily based on using the MPEG-7 eXperimentation Model (XM) to provide shot matching services over the keyframes of the video shots in the archive. We incorporated four XM algorithms: local colour descriptor; global colour descriptor; edge histogram descriptor and homogenous texture descriptor. This allowed us, for a given shot (using the representative keyframe) or externally sourced query image, to generate a ranked list of shots that best match a query image. The user of the system was allowed to choose which of the XM techniques were to be used for any given query.

In addition to these four visual shot matching techniques, we also incorporated two additional image processing techniques to improve visual shot matching performance. The first of these was motion estimation which allowed us to rank shots from the collection with regard to similarity of motion within the shots and the second was a face filter which filtered out shots from the video archive that contain one or more faces. This would be very useful, for example, if a query was to find video of people or known persons. Of course, in a video search and retrieval system such as Físchlár-TRECVid which supports both text and image based retrieval, the facility to query using both text and image evidence is essential, especially if relevance feedback is to be supported. In our experience (and for other TRECVid participants as well), the addition of visual shot matching techniques to a text-based video retrieval system improves retrieval performance, in our experiments by 13% [5].

Similar to Físchlár-News, which supports a version of relevance feedback in the form of its "related stories" and recommendation features, Físchlár-TRECVid supports relevance feedback at the shot level. For example, if a user queries for "forest fires" using a text-only query and locates a number of

good examples of shots of forest fires, then one or more of these can be added to the query and included in subsequent searches. In this way a query can be augmented and refined by adding relevant shots to the query as the user locates and identifies these shots. As shots are added to the query, the most important terms from the shots are extracted and used to augment the text aspect of the user query and the keyframes from these shots are employed for visual shot matching.

Fig. 8. Searching using text and image examples

The interface to Físchlár-TRECVid is shown in Figure 8 and is comprised of three panels. On the left of the screen is the query panel and below this is the playback window, in which the video from any selected shot can be played back through the video player. On the right of the screen is the saved shots panel in which the user keeps track of shots that have been found to be relevant. In the centre is the search result panel which displays the results of user interaction and search. As can be seen (in Figure 8) a query which is comprised of the text 'rocket launch' and a single sample image has been entered. In response to this query, the user has been presented with the ranked

list of groups of shots (the top three shown), twenty per page with a total of five pages of results available.

Presenting results from an archive of news stories is ideally done at the news story level. So a user looking for news concerning, for example, President Bush, would be presented with a ranked list of news stories about George Bush, as is the case with Físchlár-News. However, when operating at the shot level, there may be many useful video shots adjacent or near to a highly ranked shot, but which themselves may not be ranked highly. This is because the spoken text or closed-caption text may not be precisely synchronised with the video on the screen at that point. Our experience of running user experiments into interactive video search and retrieval suggests that relevant shots are often found adjacent to the highest ranked shots, especially when the query is composed of text alone. To overcome this problem results are presented not as shots but rather as groups of adjacent shots (see the centre panel in Figure 8) in which three results are displayed, each of which is composed of five adjacent shots, this giving the user the context of each result. Within this group of five adjacent shots, the middle shot is the matched shot for the group, with the adjacent shots providing context and only slightly influencing the shot ranking of the matched shot. The matched shot's keyframe is displayed largest (with a red border) and the neighbouring two keyframes are progressively smaller. In addition, the speech recognition transcript text of the five shots is presented immediately below the shots with the query terms highlighted to provide additional context. The two buttons below each keyframe in the 'search result' panel allow the user to either add the shot to the query or add the shots to the collection of saved shots for a given query.

Nearby a matching shot or somewhere within the broadcast, there is a likelihood that there will be more relevant shots. By providing a mechanism to see the full broadcast for any given shot, the system can allow more efficient searching. However, going into a full broadcast (25-30 minutes long) is at the same time making the users browsing space considerably larger, and thus will require more user effort. Our experiences suggest that the user is less likely to browse a full programme than scan a ranked list of results. Taking this onboard, in Físchlár-TRECVid, we support a three-level search and browse hierarchy. In addition to the first level results (Figure 8) from all programmes, the user can see all matched shots within a broadcast (by selecting the "MORE MATCHES IN THIS BROADCAST" option immediately above the five grouped keyframes). This will present a ranked list of groups of shots from a given broadcast and thus aids the user in locating further useful shots without having to browse all the keyframes in the programme. That said, if the user thinks there could be further more matches within other parts of this broadcast, she clicks on "BROWSE FULL BROADCAST" link from the second level (not shown), which will bring the user into the full broadcast browsing which is the third level of our three-level hierarchy. In full broadcast browsing, an interactive SVG (Scalable Vector Graphics) timeline is presented with the matched points in the broadcast highlighted, to allow

the user to quickly jump within the broadcast (shown at the top of the result panel in Figure 9). Our experiences from user experiments suggest that the three-level hierarchy is beneficial in that a user only need browse all keyframes from a full programme when the previous two levels suggest that there may be more relevant content to be found at the programme level.

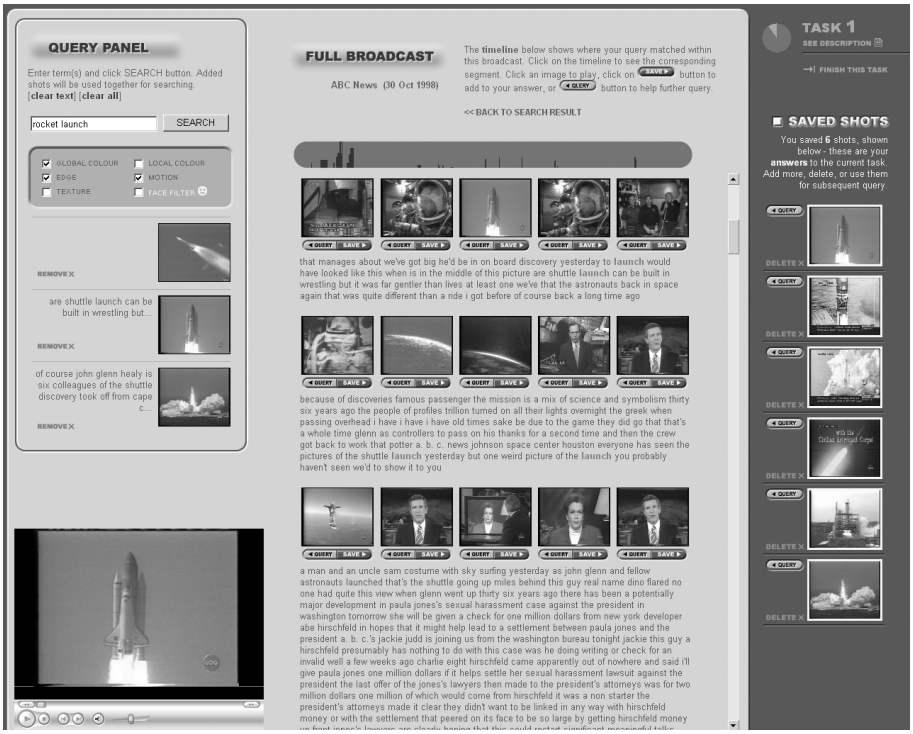

Fig. 9. Browsing a full broadcast (with saved shots)

As we have mentioned, at any point in the search session, shots (represented by keyframes and associated text) can be added to the query in a process of relevance feedback. In Figure 9 there are two shots added to the query (in the query panel on the left of the screen) in addition to the original query image. The text associated with these two shots is displayed beside their keyframes. The 'search result' panel now shows a user browsing an entire broadcast, which presents a temporally organised listing of keyframes and the text transcript from these shots.

After a phase of search and browse, the 'saved shots' panel (rightmost panel of Figure 9) will contain shots that the user considers relevant to the information need. These shots can be added to the query for further relevance feedback, or removed from the 'saved shots' panel. More details of the technical

apsects of Físchlár-TRECVid, a more detailed user iteraction scenario and its comparative performance are presented in [5].

5 Analysis of Video Searching and Browsing

In the case of Físchlár-News and the Físchlár system developed for our participation in TRECVid in 2004, we have seen how they analyse and index video and support different ways for a searcher to navigate video archives. Físchlár-News supports searching and browsing video where the unit of information is the TV news story. Each news broadcast can contain between 10 and 20 individual stories and Físchlár-News is designed and built to support people searching for news information. Sometimes users want to get a high level gist of *all* the news on a given day in which case browsing the archive by calendar and seeing a summary of all news stories on a given date is sufficient. Other times a user wants to search for news on a particular topic, in which case a keyword search against the spoken dialogue will result in a ranked list of stories for the user to browse. When a user finds a story which is of interest and wants to locate other stories on the same topic then the automatically-generated links to related stories provides this. As we have shown in an extensive user study of Físchlár-News [15], the system has enough functionality to satisfy its users as a tool for searching and browsing TV news video on a regular basis.

The Físchlár-TRECVid system supports users searching for video *shots* using a combination of text from the closed-captions or the automatic speech recognition, and/or using sample images which in some way illustrate or capture the information need, collectively. Video clips can also be used as part of the search criteria but in this case it is the keyframe from the clip rather than the clip per se, that is used. Relevant or useful video shots can also be added to the search and used as part of an expanded search which combines text searching and video/image searching into one. In searching for video shots in Físchlár-TRECVid, raw metadata such as date, location or program name does not offer any kind of useful support for searching and is not used since the information need addressed is entirely *content-based*.

For the two systems the user needs addressed are very different, and thus the two sets of specific functionalities offered are different. Browsing among news stories in Físchlár-News is very rapid and users can easily jump from one story to another in several ways. Browsing among shots in Físchlár-TRECVid is also equally fast with support for rapid visualisation across shots and the rapid location of required shots.

Searching in both systems uses text derived from closed-captions or speech recognition, and this is sufficient where the search is for information which appears in the dialogue as in TV news. Where the information need is partly based on what appears in the video then we use some aspects of image searching by extracting visual features from the video content. However, the visual features we use are low-level characteristics like colour and texture , but these

have no real semantic value and higher-level feature detectors, such as airplane taking off, or beach scenes, have proved difficult to detect with a high degree of accuracy. To move beyond feature extraction from an *entire* video frame as a basis for image searching we need to use sub-parts of an image, and ideally these should be the major objects appearing within a frame, or within a shot. There has been recent preliminary work on using video objects as a basis for video retrieval in [11, 23] and while this shows promise and appears to work well it has yet to be tried on really large collection sizes. Object-based video retrieval is one of the key challenges and areas for future research, but the main hurdle to achieving this remains the automatic identification of video objects, which has been a challenge to the video analysis community for some time and can currently be done only semi-automatically [1].

With more than one search option available (closed-captions, ASR, various low-level visual features like colour and texture, and higher-level features and possibly even video objects) one issue is how should we use these different search options in combination. A recent study of different combination methods has provided some insights [12] but the best approach appears to be to blend different search types together in a weighted combination where the weights depend on the type of query [7] as used by the Informedia system at TRECVid 2004.

However, in general we can say that while we use image features in video retrieval, we don't really use much of the visual features of video in video retrieval. We use keyframes only and we rarely use the temporal aspect of video, no inclusion of camera motion, no inclusion of object motion (though there are exceptions [19]) and so we have a long way to go in video search to develop it to a comparable level as, say, web searching.

6 Conclusions

It is inevitable that content-based information retrieval, including searching, browsing, summarisation and highlight detection of video information, is set to become hugely important as video becomes more and more commonplace. During 2003 alone, Google ran 50 billion web page searches and during 2004 AskJeeves ran more than 20 million web page searches per day, globally.[2] These figures indicate how embedded the web and web *searching* have become into our society. If video is to become even a fraction as important as we believe it is, then video searching and video browsing are critical technologies.

At the present time the development of effective video IR is decades behind text-based IR, but is catching up fast. What will accelerate this is what has accelerated web page searching, namely commercial interest, and it will be across a range of video genres and a range of applications. Searching CCTV

[2] Tuic V. Luong, Sr. VP Engineering and Technology, AskJeeves, at the 9th SearchEngine Meeting, The Hague, The Netherlands, 19-20 April 2004.

for possible suspects, searching broadcast TV news archives for past stories about tropical storm damage in Florida, searching a person's recorded TV programs on their own DVR, searching an online archive of past episodes of "Here's Lucy" to find the hilarious scene where she is working in a jam factory or searching through personal (home) video to find clips of your son and daughter together over different family vacations. These are all examples of the kind of searches we will want to do, and the need for which will drive the development of video IR.

From our experiences we can conclude that video navigation consists of search (with relevance feedback being of high importance), local browsing and collection-wide link traversal. The search techniques employed rely heavily on old text search technology with some help from visual shot matching techniques. However the video search technology could be so much more. Over the coming years we will see many advances in video search and retrieval in a number of 'hot topics'. For example, the visual shot matching techniques we have employed in the research presented in this chapter are still at the early stages of development. A user does not intuitively think of an image or video in terms of colours, textures and edges or shapes, rather the user understands semantic concepts (cars, explosions, etc.) and would like to query using these. Hence, object based search and retrieval will be a key technology where a user can define and select an object from a video clip and search for that object across an entire archive. Also key advances will come in the area of security and intelligence, where huge archives of digital video footage will be gathered, indexed and objects identified, which in so doing will help to solve another video IR problem at present; that of searching extremely large archives of tens or hundreds (or more) of thousands of hours of video content. What the search engine has done for text retrieval, security and intelligence requirements may do for digital video retrieval. Other 'hot topics' include summarisation and personalisation of video content, so that a user only gets a summary of important or novel video, in response to a query or information need.

Acknowledgements

This work is supported by Science Foundation Ireland under grant 03/IN.3/I361. The support of the Enterprise Ireland Informatics Directorate is also gratefully acknowledged.

References

1. T. Adamek, N. O'Connor, and N. Murphy. Region-based segmentation of images using syntactic visual features. In *WIAMIS2005*, 2005.
2. P. Browne, A.F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut. Evaluating and combining digital video shot boundary detection algorithms. In *IMVIP2000*, 2000.

3. M. Christel, A. Hauptmann, H. Wactlar, and T. Ng. Collages as dynamic summaries for news video. In *ACM MM'02*, pages 561–569, 2002.
4. T.S. Chua, S.F. Chang, L. Chaisorn, and W. Hsu. Story boundary detection in large broadcast news video archives: techniques, experience and trends. In *ACM MM'04*, pages 656–659, 2004.
5. E. Cooke, P. Ferguson, G. Gaughan, C. Gurrin, G. Jones, H. Le Borgne, H. Lee, S. Marlow, K. Mc Donald, M. McHugh, N. Murphy, N. O'Connor, N. O'Hare, S. Rothwell, A.F. Smeaton, and P. Wilkins. TRECVid 2004 Experiments in Dublin City University. In *TRECVid2004*, 2004.
6. D. Frey D, R. Gupta, V. Khandelwal, V. Lavrenko, A. Leuski, and J. Allan. Monitoring the news: a TDT demonstration system. In *HLT'01*, pages 351–355, 2001.
7. A. Hauptmann, M-Y. Chen, M. Christel, C. Huang, W-H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, and H. Wactlar. Confounded Expectations: Informedia at TRECVid 2004. In *TRECVid2004*, 2004.
8. IBM. MARVEL: MPEG-7 Multimedia Search Engine. website available at: http://www.research.ibm.com/marvel/ (last visited august 2005).
9. I. Ide, H. Mo, N. Katayama, and S. Satoh. Topic threading for structuring a large-scale news video archive. In *CIVR2004 (LNCS3115)*, pages 123–131, 2004.
10. Informedia. Informedia digital video understanding research. website available at: http://www.informedia.cs.cmu.edu/ (last visited august 2005).
11. J. Sivic J and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV 2003*, 2003.
12. K. McDonald K and A.F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *CIVR2005 (LNCS3569)*, pages 61–70, 2005.
13. W. Kraaij, A.F. Smeaton, and P. Over. TRECVid2004 - an overview. In *TRECVid2004*, 2004.
14. H. Lee, A.F. Smeaton, N. O'Connor, and N. Murphy. User-interface to a CCTV video search system. In *IEE ICDP2005*, pages 39–44, 2005.
15. H. Lee, A.F. Smeaton, and B. Smyth. User evaluation outisde the lab: the trial of Físchlár-News. In *CoLIS5 Workshop on Evaluating User Studies in Information Access*, 2005.
16. C-Y. Lin, B.L. Tseng, and J. Smith. Video collaborative annotation forum: establishing ground-truth labels on large multimedia datasets. In *TRECVid2003*, 2003.
17. N. O'Hare, A.F. Smeaton, C. Czirjek, N. O'Connor, and N. Murphy. A generic news story segmentation system and its evaluation. In *ICASSP2004*, 2004.
18. D. O'Sullivan, B. Smyth, D. Wilson, K. McDonald, and A.F. Smeaton. Improving the quality of the personalized electronic program guide. *User Modeling and User-Adapted Interaction*, 14(1):5–36, 2004.
19. M. Rautiainen and D. Doermann. Temporal color correlograms for video retrieval. In *ICPR2002*, pages 267–270, 2002.
20. E. Rennison. Galaxy of news: an approach to visualizing and understanding expansive news landscapes. In *UIST'94*, pages 3–12, 1994.
21. R. Ronfard. Reading movies - an integrated DVD player for browsing movies and their scripts. In *ACM MM'04*, pages 740–741, 2004.
22. D. Sadlier, S. Marlow, N. O'Connor, and N. Murphy. Automatic tv advertisement detection from mpeg bitstream. *Journal of the Pattern Recognition Society*, 35(12):2719–2726, 2002.

23. S. Sav, H. Lee, A.F. Smeaton, N. O'Connor, and N. Murphy. Using video objects and relevance feedback in video retrieval. In *SPIE Vol. 6015*, 2005.
24. A.F. Smeaton. Indexing, Browsing and Searching of Digital Video. *Annual Review of Information Science and Technology (ARIST)*, 38:371–407, 2004.
25. A.F. Smeaton, C. Gurrin, H. Lee, K. Mc Donald, N. Murphy, N. O'Connor, D. O'Sullivan, and B. Smyth. The Físchlár-News-Stories System: personalised access to an archive of TV news. In *RIAO2004*, 2004.
26. A.F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, H. Lee, K. Mc Donald, P. Browne, and J. Ye. The físchlár digital video system: A digital library of broadcast TV programmes. In *ACM+IEEE Joint Conf. on Digital Libraries*, 2001.
27. Alan F. Smeaton, Paul Over, and Wessel Kraaij. TRECVid: evaluating the effectiveness of information retrieval tasks on digital video. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 652–655, New York, NY, USA, 2004. ACM Press.
28. R. Swan and J. Allan. Timemine: visualizing automatically constructed timelines. In *SIGIR2000*, page 393, 2000.
29. TRECVid. TREC Video Retrieval Evaluation. website available at: http://www-nlpir.nist.gov/projects/t01v/t01v.html (visited august 2005).
30. A. Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J.R. Ohm, and M. Kim. Mpeg-7 visual part of experimentation model version 9.0, 2001.