# Content-Based Access to Digital Video: The Físchlár System and the TREC Video Track

*Alan F. Smeaton*

School of Computer Applications and Centre for Digital Video Processing
Dublin City University, Glasnevin
Dublin 9, IRELAND.

## ABSTRACT

This short paper presents an overview of the Físchlár system - an operational digital library of several hundred hours of video content at Dublin City University which is used by over 1,000 users daily, for a variety of applications. The paper describes how Físchlár operates and the services that it provides for users. Following that, the second part of the paper gives an outline of the TREC Video Retrieval track, a benchmarking exercise for information retrieval from video content currently in operation, summarising the operational details of how the benchmarking exercise is operating.

## 1. INTRODUCTION

The development of technology to support digital multimedia has, to date, concentrated on the creation, capture, storage, compression, transmission, presentation, streaming and delivery of digital information including text, image, audio and video. If we examine digital video information in particular we can see that progress made in computing power, memory, storage and networking have all contributed to making the acquisition, storage and playback of digital video now a commodity operation. This has left us in our present position in which huge libraries of digital video information are becoming available to us on fixed and soon on mobile platforms for a variety of applications such as education, entertainment and communications. However comparatively little research and development has been done on organising digital video information so that it can be indexed, searched, browsed, summarised, filtered or otherwise manipulated by content. The area in which the Centre for Digital Video Processing at Dublin City University works is in developing techniques to allow compressed digital video information to be analysed automatically in order to support advanced content-based operations.

This short paper is divided into two parts. In the first part we present a number of principles which apply when considering information retrieval from multimedia objects and which differentiate MMIR from information retrieval on text. We shall then briefly outline the Físchlár system - an operational system at Dublin City University which is used by over 1,000 users on campus for teaching, learning, and for entertainment. Físchlár serves as an illustration of the kind of content-based operations which can be performed on digital video information and which we are finding that our campus users are using in various applications. In the third part of the paper we briefly summarise the work ongoing in the TREC Video Retrieval track taking place this year. While this work is still very preliminary and the results and impact of the track are still awaited we believe it is instructive to include an overview of the TREC activities in content access to digital video in this workshop.

## 2. PRINCIPLES OF MULTIMEDIA INFORMATION RETRIEVAL

Before briefly describing the Físchlár system and the problems of information retrieval from multimedia in general and video in particular, it is probably instructive to examine the characteristics of multimedia information which makes retrieving information from such information, so difficult.

Any kind of multimedia information has multiple dimensions because of the richness of its information content – far moreso than for text – and how we view a video clip or image, what our task in viewing it ultimately is, what information we are seeking and why, etc., all elicit different properties from a MM object. Different features of an image or video will interest us at different times depending on what we are looking for and why. This is true for text also but to a lesser extent as text is not as *information rich* as the visual media, and we have enough difficulties performing effective information retrieval on text anyway !

Given that a visual multimedia object can have such a "moving target" when it comes to its interpretation by us it follows that we may eventually require retrieval of and from multimedia objects based on properties of the objects which are not initially captured when we index them. We should really account for this by having dynamic, on-the-fly, query-driven "re-indexing" of multimedia but for large collections of information this is prohibitively expensive so we compensate by indexing images and video by as many different types of characteristics as we can and using whatever and whichever of these is appropriate at retrieval time. This is in stark contrast to information retrieval from text where we index only once and perform all retrieval at query time on this index. In indexing multimedia information for subsequent retrieval we should be aware of the following

- We should develop suites of retrieval techniques for sub-groups of features, each based on an inexact match between query specification and the index,

which can then be combined into an overall assessment or ranking of multimedia objects if that is what the system provides for users;

- We should allow for, and handle, indexing of multimedia information which is incomplete, inexact and possibly erroneous;
- We must understand that queries provided by users are incomplete (as indeed is the case with retrieval from text) but because image and video especially are so more content-rich than text documents, the concept of "relevance" may be much more difficult to model and capture;

It follows from these points above that content based information retrieval on video is hugely more difficult than retrieval on image which is more difficult than for audio which is more difficult than for text.

Because video and image information, as visual media, have such a richness of interpretation, one of the characteristics of image and video retrieval systems is the integration of browsing and querying in the searching interface. This comes from the fact that one cannot easily "gist" an image or a video clip as easily as one can quickly summarise a text document because it has so many interpretations. Much of the significant work on information retrieval from video and from image data addresses techniques to allow us to quickly browse and "gist" image and video. For video the difficulty is compounded by the challenge of having to locate and "gist" content from within possibly long video clips and as a result, effective and efficient information retrieval from digital video information, a medium which is both visual and which is continuous, is probably the most challenging of tasks. It is against this backdrop that we have built the Físchlár system which tries to address some of these difficulties.

## 3. THE FÍSCHLÁR SYSTEM

When providing retrieval on digital video it is straightforward to treat video as a binary blob and to index and retrieve via its associated metadata such as title, date, etc., but that is not what we are interested in here. Most work on IR on video streams has concentrated on analysis of the visual stream as information retrieval on the audio stream defaults to being information retrieval on spoken documentsand our particular interest here is to look at the visual stream;

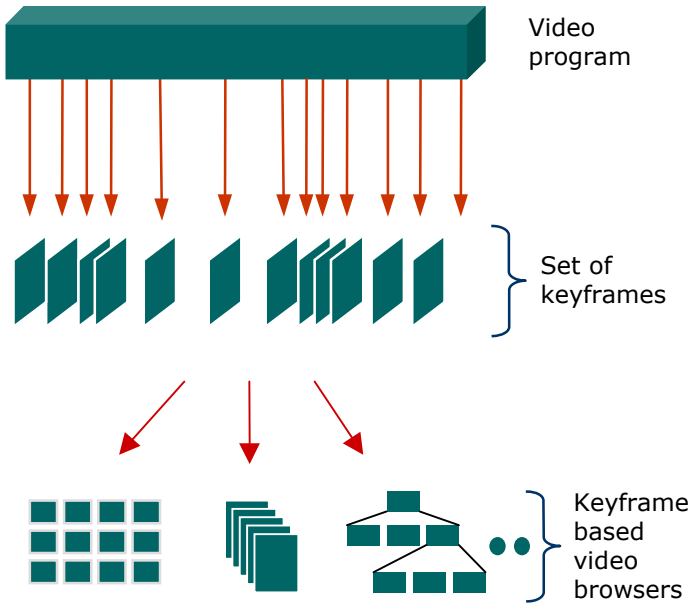
The way to make progress with IR on video is to structure the video in some way and above the level of the single frame, the next basic unit is the shot followed by the scene. A shot is a sequence of frames from a single camera motion over time and automatically structuring a video begins by identifying shot boundaries automatically. The usual approach to shot boundary detection (SBD) is to compare adjacent frames to see if they are dissimilar over a certain threshold and if so, then it is likely that shot bound is present. For hard cuts, the most usual technique to accomplish this is to compare adjacent frames based on their colour histograms but shots can be joined using more sophisticated cinematic techniques like fades to black, dissolves, wipes or computer-enhanced transitions. For these, colour histograms are less successful because the shot transitions occur over time, i.e. over multiple frames. Other SBD techniques are based on edge detection (which are good for dissolves), brightness or directly from the encoded stream, based on macroblock types or on motion vectors;

In our own work we have evaluated several SBD methods on a collection of 8 hours TV broadcast (720,000 frames, manually marked up), evaluating the performance of each both individually and in combination [1]. Our results show that the best individual technique yields over 90% in both precision and recall, and the best results are achieved by a combination of techniques but this is only a couple of % above the best individual approach. It may be that incorporating information from the associated audio track could help in SBD but not always as there are silences between some shots, but not always.

In considering the SBD task the computational cost of the processing must be taken into account. Most of the techniques mentioned above run in approximately real time on general purpose hardware while those operating on the compressed domain are naturally much faster.

Once structured, what should happen next ? The usual approach is to present video visually and to choose a keyframe for each shot and use this as a basis for browsing. Keyframe selection can be based on choosing the middle, first, last, average colour, etc. but there is no best technique. Video retrieval systems could de developed based on image retrieval on shot keyframes but the norm is to present keyframes for browsing but a problem here is the sheer number of shots or keyframes … a 30 minute program can have hundreds of keyframes for example.

Thus some kind of structure should be applied to keyframe sets such as the following diagram illustrates:



The SBD and keyframe selection task is embodied in the Físchlár system [2]. Físchlár is a large capacity (several hundred hours), communal, shared, video on demand system, where broadcast TV programs are recorded on demand by users from SMS messages or via a web browser, and recorded programs are indexed into shots and scenes, with a selection of keyframe browsing interfaces to locate sections within videos from which streaming can begin. A personalisation and recommender system [3] which assigns program genres, can be used to organise the TV schedule or the library of recorded programs.

Físchlár has a web interface and is available from a web browser or mobile PDA such as a Compaq iPAQ. It is used by over 1000 people on campus for a variety of applications including entertainment (Físchlár is available from student residences on campus), study (a version of Físchlár which records TV news programs is available from the main University library in DCU and other Universities), research (Físchlár is used by Faculty in disciplines like journalism, communications, languages and business to make broadcast video available to a large number of students.

In our work on Físchlár we also capture the closed captions or teletext 24x7 from 6 terrestrial channels in the Dublin area (RTE1, Network2, BBC1, BBC2, Channel 4 and ITV) and we use this as a basis for searching, alerting and summarisation of broadcast TV programs. Alerting works by matching user profiles against the incoming closed captions and alerting via email and/or SMS messages. At the present time, search works by matching user's text queries against closed caption streams and locating parts of recorded programmes which are ranked.

Part of the expertise in the Centre for Digital Video processing is in MPEG-4 encoding where some colleagues have been instrumental in developing and specifying that standard. That expertise in object recognition and tracking is now being used to develop more sophisticated navigation through video content by locating similar objects across different scenes within a program or across different programs within the same genre. We already structure TV News programmes into scene and story bounds [4] and our immediate work involves hyperlinking stories across news broadcasts based initially on dialogue and then based on recognised objects. Identifying and tracking objects as per MPEG-4 encoding will allow us to compute similarities across programs based on dialogue and objects present in the visual stream.

Meanwhile, we are continuing to work on deconstructing the encoded stream to reverse engineer characteristics like camera motion (panning, zooming, etc.) and faster SBD which could, if successful, be included into the operational Físchlár system to enrich the index representation of video content.

## 4. THE TREC VIDEO RETRIEVAL TRACK

Our work on the development of Físchlár has provided us with an excellent environment for producing a usable system for a large population of users who have varied information needs but as with other research groups working in this field, evaluating the performance of the system as an information access tool is difficult. Until recently there are no test collections, no agreed evaluation metrics and no opportunities for direct comparison across systems or approaches. This year's annual TREC benchmarking exercise includes a special "track" or line of activity, on video indexing and retrieval.

The Text REtrieval Conference (TREC) [5] is an annual event which has been ongoing for 10 years and whose aim is to foster and facilitate the evaluation of information retrieval tasks. In the first TREC conference the IR task was ad hoc retrieval – text queries run against text documents – though this has now diversified over the last decade to include retrieval from spoken documents, retrieval from web pages, interactive retrieval, retrieval from different languages, retrieval across different languages and now, this year, retrieval from digital video information.

The series of TREC exercises have involved participating groups benchmarking the effectiveness of their various approaches to whatever "task" is on the agenda for that year, on a common dataset and using a common set of queries or topics. Participating groups then send the results of their system, which implements whatever aspect of retrieval that they are interested in, back to the TREC organisers who take the combined retrieval results of all participating organisations and then pool them together, eliminating duplicates, and performing a manual assessment of relevance. Once this "ground truth" is established, the performance of the participating groups in terms of precision, recall and other effectiveness measures, can be computed.

In this year's TREC evaluation, video browsing and retrieval systems can be used in an information-seeking task as well as in the more straightforward task of shot boundary detection. Specifically, the following three tasks are available for participating groups:

- Shot boundary detection - the rationale for this is that it is a function which is needed for higher-level tasks and provides an easier entry to the TREC process due to the existing base of example work, software,...
- Known-item(s) search which reflect a significant type of user need where a user knows that the item being sought exists in the collection and the task is to find it. This task has the advantage of having lower evaluation costs and human assessors are not needed since the "answers" are identified by the authors of the topic who, in the case of the video track, are the participating groups;
- General statements of information need which represent the most diverse type of video searching and which is the most difficult and the costliest to evaluate, but ultimately will represent the most important type of video searching for real users in real applications;

The data used in the TREC track consists of approximately 11 hours of MPEG-1 encoded video originating from NIST itself (2

hours of US government videos), the Open Video project [6] and some stock footage from the BBC. The topics to be used in the known item(s) and general statement retrieval were formulated by the participating groups themselves and some of the topics are below:

- "Speaker talking in front of the US flag"
- "Astronaut driving lunar rover over lunar surface"
- "Ronald Reagan reading speech about Space Shuttle"

Although the outcome of these activities are not known at the time of writing, the TREC video track has already brought together some of the major research groups which develop systems for content based video manipulation and the impact of this activity is likely to be significant. Participating groups in the TREC video retrieval track this year come from at least the following institutions: Fudan University (China), Microsoft (China), Université Josef Fourier (France), Dublin City University, University of Amsterdam University (Netherlands), University of Glasgow (UK), Imperial College (UK), IBM TJ Watson Research Labs (USA), Johns Hopkins University (USA), University of Maryland (USA) and Carnegie Mellon University (USA), though others may yet contribute

## 5. CONCLUSIONS

Multimedia information retrieval, especially retrieval from digital video, is a research topic for a relatively small number of research groups and to date, each has been able to measure the effectiveness of its techniques in isolation. This is true of our own work on the Físchlár system described earlier.

However, there are several initiatives afoot which have identified the need for a collaborative evaluation which allows direct comparison of the various approaches and techniques qppearing in the literature. The MIRA working group was an EU-funded Network of Excellence [7] whose aim was to explore issues related to evaluation of multimedia information retrieval. One of its goals was to create a multimedia test collection, but this was never achieved. The DELOS Network of Excellence [8] is another EU-funded working group, this time in the area of Digital Libraries, and this group has spawned off a working group to investigate the development of a test suite. The influential President's Information Technology Advisory Committee report on Digital Libraries also calls for work on building test environments.

All these efforts serve to illustrate how important the development of an evaluation environment is, but the most successful such effort is easily TREC.

With regard to information retrieval from digital video, we cannot take text-based IR and apply it to continuous media and we must re-think the whole user-system interaction and combine the search-browse interaction seamlessly. As digital TV achieves greater penetration and the use of devices such as TiVo and Replay boxes spread, the demand for content access to video will soar. People will want to be able to access video on their 3G mobile phones and from their STBs and this will create huge demands, and more importantly, huge markets. The existence of these emerging markets will drive the development of video retrieval, in the same way that the web searching application dominates text-based information retrieval research.

## 6. REFERENCES


[1] Evaluating and Combining Digital Video Shot Boundary Detection Algorithms. Browne P, Smeaton A, Murphy N, O'Connor N, Marlow S and Berrut C. In Proceedings of IMVIP 2000, Belfast, Northern Ireland, September 2000.

[2] O'Connor, N., Marlow, S., Murphy, N., Smeaton, A., Browne, P., Deasy, S., Lee, H. and Mc Donald, K. Físchlár: an On-line System for Indexing and Browsing of Broadcast Television Content. In *Proceedings of ICASSP 2001* (Salt Lake City, UT, May, 2001).

[3] Smyth, B. and Cotter, P. A Personalized Television Listings Service. *Communications of the ACM*, **43**(8), 2000, 107-111.

[4] O'Connor N, Czirjek C, Deasy S, Marlow S, Murphy N and Smeaton A. News Story Segmentation in the Físchlár Video Indexing System. *In Proceedings of ICIP 2001 - International Conference on Image Processing*. Thessaloniki, Greece, 7-10 October 2001.

[5] http://trec.nist.gov/

[6] http://www.open-video.org/

[7] http://www.dcs.gla.ac.uk/mira/

[8] http://www.iei.pi.cnr.it/DELOS/delos2/TestSuite/testsuite.htm