Challenges for Content-Based Navigation of Digital Video in the Físchlár Digital Library

Alan F. Smeaton

Centre for Digital Video Processing,
Dublin City University, Glasnevin, Dublin, 9, IRELAND.
Alan.Smeaton@dcu.ie

Now that the engineering problems associated with creating, manipulating, storing, transmitting and playback of large volumes of digital video information are well on their way to being solved, attention is turning to content-based and other means to access video from large collections. In this paper we present an overview of the different ways in which video content can be used, directly, to support various ways of navigating within large video libraries. Some of these content-based mechanisms have been developed and implemented on video already and we use our own Físchlár system to illustrate many of these. Others remain beyond our current technological capabilities but by sketching out the possibilities and illustrating with examples where possible, as we do in this paper, we help to define what challenges still remain to be addressed in the area of content-based video navigation.

1. Introduction

The technical challenges associated with the capture, compression, storage, streaming, transmission and playback of digital video information have commanded most of the attention in the areas of digital video research and development over the last decade. This has been necessary in order to allow real video libraries which can deliver video content to users in an efficient manner, to be developed. The engineering problems and challenges associated with this are being solved or are close to being solved, thanks to the development and adoption of standards such as those from the MPEG family. Now that we can comfortably engineer and construct the actual video library systems, we can turn our attention to the problems faced by users who wish to use these systems, and foremost among those problems are the issues of how to effectively navigate through a video library. While the first video libraries were based around indexing and content navigation based on video metadata such as titles, actors, dates, etc. [1] such access does not fully leverage the advantages of having video content in digital format. What are also needed are techniques which operate *directly* on the video content, in the same way that text-based information retrieval searches are performed *directly* on the documents or web pages rather than on some metadata information.

As we begin to contemplate the possible ways in which digital video information can be manipulated, directly, based on its content, our natural inclination is to limit ourselves to the kinds of content based operations which we already perform on analogue video and to regard the benefits of being digital as simply being related to ease of access, playback, free ridership and replication. This, however, is under-exploiting the full advantages of video being digital and this is what we want to address in this paper.

Our aim in this paper is to introduce and summarise most of the functions which have been developed to allow content-based operations on text, image and audio, and to present these in a rough classification. The purpose of this is to then view digital video information as just another kind of media, and to examine the possibilities for content-based operations on video using the rough classification from content-based operations from other media. It is hoped that by examining video navigation from this more abstract position, we may gain some insights into what kinds of video navigation tools are possible, what is still missing based on current technology, and what remains to be discovered and built. We use our own video library system, Físchlár, to illustrate by example wherever possible.

The rest of this paper is organised as follows. In the next section we present – in an abstract way – the different kinds of content-based navigation operations that can be done on text, image and audio documents. Section 3 gives a brief summary of the Físchlár system while in section 4 we re-examine the content-based navigation tools which have been developed for video. For this section we illustrate from the Físchlár system wherever possible. Finally, a concluding section sets out the challenges which remain to be addressed in order to achieve the goal of truly flexible, effective, efficient, and scalable video navigation on large volumes of video content

2. Content-Based Navigation through Text, Image and Audio

In examining the broad amount of work done in content based retrieval we will deliberately exclude work done on retrieval using metadata such as descriptor captions or structured data fields, as we are interested on retrieval *directly* from information and we are also not concerned with issues of scale or size of the data being searched.

We have divided the whole gamut of content-based navigation operations into three broad categories, namely searching, grouping related objects, and summarising.

2.1 Searching

Searching through text, image or audio collections is based on matching a user's query against some set of units of information whose granularity can vary from an entire whole object (document, image, audio recording) to a small element within (a fact within a document, an object in an image or an utterance in an audio recording. Searching of text databases is well established and there are well-known techniques for retrieving whole documents (web searching for example), for retrieving parts of documents (such as passage retrieval) [2] and within recent years we've seen the emergence of question-answering [3], where a "factoid" such as a person name, monetary amount, a date or a location, is retrieved from a document database, in response to a query. In a sense this corresponds to retrieval of the "atomic" entities within a text document as retrieval of text units of information smaller than a factoid would simply not make sense.

For image retrieval, retrieval of entire whole images from a database in response to a query or sample image is the most common application [4]. Retrieval of parts of whole images or image fragments in response to a query image is generally restricted to specialist applications such as medicine, astronomy or geographic information systems where the original data source (the images in the database) consists of images which are much larger in size than that which is required by the user. An example from the medical domain can be seen in [5] where regions of an image in the database are retrieved in response to a query. Finally, retrieval of the "atomic" units making up an image is analogous to identifying and then retrieving objects from an image and this is possible only if the objects to be identified and retrieved, are homogenous. For example, it is possible to index a database of images of faces, or fingerprints or trademarks and to perform retrieval of the actual objects (faces, prints or trademarks) identified in those images. It is important to make the distinction between matching against and then retrieving the objects in an image, as opposed to retrieving the whole image which is easier.

If we restrict our analysis to audio recordings of speech then we can say that indexing and retrieval of recorded audio is less developed than on other media. This is because the analysis of the raw data into something meaningful requires speech recognition, or phone recognition at least. Systems such as the Taiscealaí system [6] which retrieve clips or segments of audio recordings, have been developed and these are analogous to retrieving passages from within text documents or fragments from within whole images. Retrieval of entire audio "documents" or recordings in response to a user query on content is generally not a useful operation, and retrieval of discrete and identified spoken "factoids" from an audio archive, is surely now within reach, although this should not be confused with retrieving arbitrary audio clips which just might contain "factoids".

2.2 Grouping Related Objects

The task of automatically grouping objects related by virtue of having related content and then using this as part of retrieval, broadly falls into three specific areas.

In the first case, object pairs (text documents, images) can be linked based on their direct similarity, independent of any superimposed or overarching structure. This would correspond to "find more like this" which uses a given whole text document or an image as a query and finds the documents or images which are most similar to the given one. Such links can be computed dynamically, or pre-computed and stored, and while this is reasonably commonplace for text and images it is not generally done for spoken audio. Following such links is akin to standard web browsing, following static hypertext links.

Another use for grouping of related objects is in automatically calculating a clustering of the set of objects in the collection. This is initially based on calculating pairwise object-object similarities and then applying a clustering algorithm to detect groups of related objects which form clusters. Sometimes outliers or small clusters of objects remain but most objects are generally brought into the overall classification.

Like the calculation of independent pairwise links, clustering works well on text documents and on images but not audio data.

Somewhere in between independent pairwise linking and overall collection clustering lies the concept of automatically linking some related objects from the collection and superimposing some kind of higher level local structure. This is local in the sense that it is not applied to the entire collection but to some subset such as the output of a web search in order to impose some structure on the result of a query [7]. Another example of this would be a dynamically generated guided tour [8] and although we can find no examples of this applied to images there is no reason why it could not be applied.

2.3 Summarisation of Content

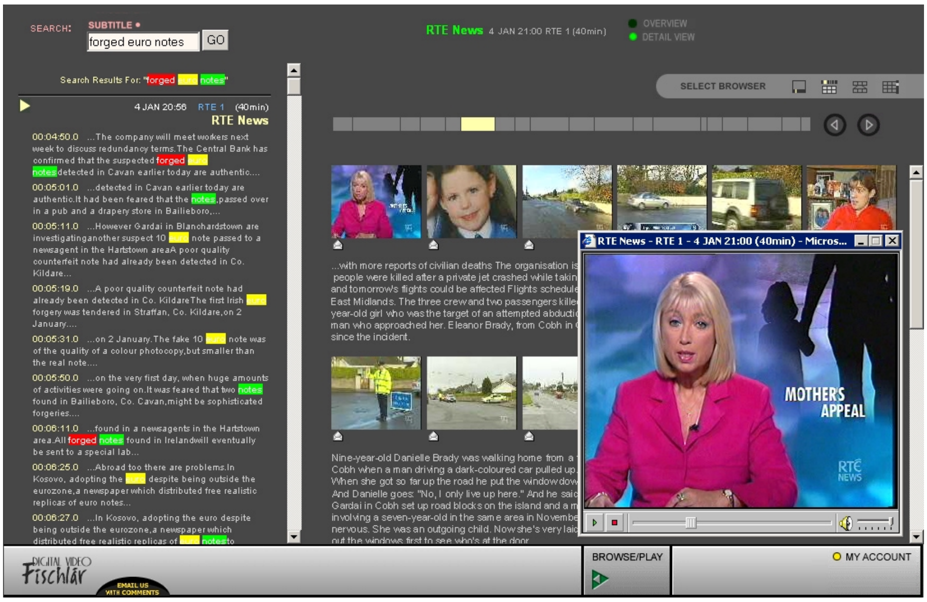
The summarisation task can be regarded as fairly straightforward. Given a single object (document, image, video program), automatically generate a smaller or shorter version which encapsulates the most important content from the original in such a way that an approximation of the meaning of the doucment, image or video, can still be obtained. Automatic summarisation of text documents is now an achievable task [9], although systems which do this are not yet robust and accurate enough to be in widespread use. Automatic summarisation of individual images is a trivial engineering task and involves generating a thumbnail or lower resolution and quality to the original. Automatic reduction of spoken audio information can involve detecting and eliminating silences and pauses in the original, and playback at a faster speed, and in this way the audio is reduced and takes less time to play. There is good potential for combining text summarisation with audio summarisation, and using audio features such as intonation and prosody, though this is still very much a research area.

The summarisation task may be applied to a set of objects and the task is much more difficult, to automatically generate a single, object, reduced in some way, which approximates the meaning of the set of objects. Such summarisation is only at the research stage, for all media.

3. Overview of the Físchlár Video Library System

Before we examine how searching, linking and summarisation can apply to video information, we will now introduce our own video digital library system, Físchlár, and use it for illustrating some content-based operations on video in the section to follow. Físchlár is a library of digital video information which records, analyses, indexes, stores and provides streamed playback as well as browsing and searching, on broadcast TV materials in a University campus environment. At any point in time there are between 300 and 400 hours of video content available to a userbase of over 1,500 registered users, though only about 1,000 of those are actually active and regular participants. The video streaming technology behind Físchlár is capable of supporting over 250 simultaneous streams of MPEG-1 encoded video and the interface to Físchlár is via a conventional web browser with a plug-in for video streaming.

There are two versions of the Físchlár system in use on our campus, one which allows recording, browsing and playback of TV programmes transmitted on any of 8 terrestrial TV channels and "Físchlár-News" which is the system of interest here. This automatically records the main evening news from the national broadcaster's main TV station and the archive of TV news has now grown to almost 1 year in size (c.170 hours of news content). Each program is digitised into MPEG-1 and then submitted to an analysis phase during which we automatically detect the boundaries between different camera shots. For shots over a threshold length we automatically select a single frame as the keyframe whose content is somehow indicative of the content of the whole shot. The entire News programme can then be presented to a user as a storyboard of keyframes, either the full set of keyframes (perhaps 200-250 such keyframes per broadcast) or an abstracted subset of 20 or 30. To support user browsing through these keyframes we have developed several keyframe browser interfaces [10] from which the user can choose. To compliment the keyframe browsing facilities in Físchlár, we have developed a text-based search tool for video digital libraries. We capture the closed captions associated with 6 TV channels, 24x7 and we link these to the archived broadcast. In this way we can support text-based searching through the text caption archive with relevant video clips as "answers" to queries.



Search Results For: "forged euro notes"

4 JAN 20:56 RTE 1 (40min)
RTE News

00:04:50.0 ...The company will meet workers next week to discuss redundancy terms.The Central Bank has confirmed that the suspected forged euro notesdetected in Cavan earlier today are authentic....

00:05:01.0 ...detected in Cavan earlier today are authentic.It had been feared that the notes,passed over in a pub and a drapery store in Bailieboro,...

00:05:11.0 ...However Gardai in Blanchardstown are investigatinganother suspect 10 euro note passed to a newsagent in the Hartstown areaA poor quality counterfeit note had already been detected in Co. Kildare...

00:05:19.0 ...A poor quality counterfeit note had already been detected in Co. KildareThe first Irish euro forgery was tendered in Straffan, Co. Kildare,on 2 January....

00:05:31.0 ...on 2 January.The fake 10 euro note was of the quality of a colour photocopy,but smaller than the real note....

00:05:50.0 ...on the very first day, when huge amounts of activities were going on.It was feared that two notes found in Bailieboro, Co. Cavan,might be sophisticated forgeries....

00:06:11.0 ...found in a newsagents in the Hartstown area.All forged notes found in Irelandwill eventually be sent to a special lab...

00:06:25.0 ...Abroad too there are problems.In Kosovo, adopting the euro despite being outside the eurozone,a newspaper which distributed free realistic replicas of euro notes...

00:06:27.0 ...In Kosovo, adopting the euro despite being outside the eurozone,a newspaper which distributed free realistic replicas of euro notesto

...with more reports of civilian deaths The organisation is
people were killed after a private jet crashed while takin
and tomorrow's flights could be affected Flights schedule
East Midlands. The three crew and two passengers kille
year-old girl who was the target of an attempted abducti
man who approached her. Eleanor Brady, from Cobh in
since the incident.

Nine-year-old Danielle Brady was walking home from a
Cobh when a man driving a dark-coloured car pulled up.
When she got so far up the road he put the window dow
And Danielle goes: "No, I only live up here." And he saic
Gardai in Cobh set up road blocks on the island and a m
involving a seven-year-old in the same area in Novembe
nervous. She was an outgoing child. Now she's very laic
out the windows first to see who's at the door

Fig. 1. **Screenshot from Físchlár showing search for "forged euro notes" and browsing through keyframes of relevant TV News program.**

An example of teletext search through a video news archive can be seen in Figure 1 where the user has requested clips about "forged euro notes", has chosen the TV News of 4 Jan 2002, is browsing the keyframes and associated teletext and is playing back a news clip of the anchorperson. However, teletext search is just part of our work on analysing video in order to support searching and browsing. We have also developed and tested techniques for news story segmentation, segmentation of programmes into scenes which are groups of shots, counting the number of human faces in a shot, speech-music discrimination, speaker identification, camera motion detection and detection of the amount of object motion in shots. We are also working on sophisticated object and shape detection and have been able to demonstrate this as a detector which recognises the appearance of either Bart or Homer Simpson in "The Simpsons", in real time. Further details of this work can be found on the project publications page at http://www.cdvp.dcu.ie/publications/

The Físchlár system has been deployed on campus for the last 2-and-a-half years and has allowed us to build up a userbase of real users with real video retrieval needs and we have been able to observe, first hand, how people want to do content access to digital video libraries.

4. Content-Based Operations on Digital Video

4.1 Searching Digital Video

Digital video information is structured from individual still frames into shots (single camera motion over time), then into logical scenes (groups of shots which make up a semantic whole), then into programmes and when we search video we may wish to retrieve from any of these levels of granularity. Retrieval of whole programmes can really only be done using metadata and this is not of interest to us here. Similarly, retrieval of individual frames is best done by treating each frame in the video as an image and using image retrieval techniques.

When it comes to searching for parts of video then a user's query can be either a text query – as in a web search engine – or a user may use an existing video shot or an image, as the query. In response to this, video shots and/or scenes can be retrieved. This requires shot boundaries and the boundaries between logical scenes to be detected, automatically. Shot bound detection is a well-established problem with many solutions and the recent TREC video track [11] benchmarked several different approaches to this problem. Detecting the boundaries between scenes can also be done, though this is more difficult. In Físchlár we can automatically detect the boundaries between news stories in a news broadcast and will shortly be able to retrieve an entire segmented news story in response to a user's query. Such segmentation is easier than detecting and retrieving logical scenes from other broadcast TV though this is something we are presently working on. To achieve this kind of scene segmentation we have developed techniques for speech/music discrimination, speaker identification and other audio analysis, face detection, anchorperson detection, etc. and we will incorporate these into the Fischlár system. Meanwhile a user's text query in Físchlár can be used to identify video clips based on the associated teletext which for the present is used to bring the user to a relevant part of the video, with no shot or scene segmentation for now. An example of this in action can be seen in Figure 1.

At Carnegie Mellon University, the Informedia project [12] has had much success in developing techniques for video indexing and retrieval. The CMU approach is based on searching speech transcripts, where the speech recognition has been performed by the Sphinx speech recognition engine. This is then coupled with automatic identification of faces from the video, naming of those faces when they are among a database of "VIPs", automatic identification of the occurrence of text on-screen in captions or as part of the image, OCR applied to such text, and automatic identification of named locations. These locations are then input into a mapping interface allowing a user to see whereabouts on the globe or on a smaller scale, that the video content refers to.

The IBM CueVideo project at IBM Almaden also does searching of video content and in this case it is based on speech recognition, smart browsing and other video analysis techniques [13].

Retrieval of video clips based on queries in media other than text was also supported in the TREC video track where user queries could include actual video shots and the task was to find other video shots similar to the query. Some exploratory approaches which have been taken include matching based on colour correlograms for a video shot (a multidimensional data structure incorporating colour histograms for frames spread throughout a shot) and automatic pre-classification of video shots into a semantic hierarchy [11]. While these, and other similar approaches represent the "high end" in terms of sophistication of approaches taken they do allow first attempts at automatic video retrieval based on video queries.

4.2 Grouping Related Video Clips

As we saw earlier, grouping pairs of objects on the basis of relatedness depends upon being able to compute similarity between objects – in the case of video, either shots or scenes. In the Físchlár system we are developing techniques to automatically link related news stories which have been automatically segmented. Our current work does this based on linking teletext associated with each story, which is effectively the same as link creation in text documents, but we are also working on incorporating image and audio analysis into this process. Automatically detecting anchorperson, image similarity and speaker identification are all being used to lead to a "higher level" of similarity matching between news stories, so we move away from a dependency on text (closed caption) similarity. Most of the difficulties associated with text-based information retrieval revolve around word polysemy and other language ambiguities [14] and adding in audio and image based analysis will help to redress some of these difficulties.

In our initial work, pairwise linking of news stories is done independently of all other stories but we hope to move on to more elaborate local clustering to sets of news clips, perhaps as part of a user's video navigation session in the same way that guided tours on text documents have been dynamically created in our previous work [8]. Following that we can address the possibilities of global clustering of an entire video archive, though we are not aware of any work reported to date which attempts to group related video clips. The hurdles to doing this are simply the ability to segment video accurately into meaningful units such as news stories, and the ability to calculate similarity values between such units. Our work on video searching, via text, plus image and audio analysis, gives us the foundation for such work.

4.3 Summarisation of Video

The most widespread type of video summarisation is based on generating keyframes from video shots and presenting these via some kind of keyframe browser interface. In Físchlár we have several keyframe browsing interfaces which allow a user to browse the several hundred keyframes from a TV program [10]. For example, a 30 minute news program may have between 300 and 400 such keyframes in one of these interface we also automatically reduce this large set to a summary of only about a dozen keyframes, one per news story.

The way in which one automatically summarising video into a shorter video clip is wholly dependent on the genre or type of the video. For some types, such as TV documentaries, soaps, etc., this is very difficult but for others – such as sports programmes – we have been able to make progress. In Físchlár we can generate summaries of certain types of sports programmes which have crowd noises and a rolling commentary based simply on audio analysis [15]. By detecting and measuring the level of crowd noise, and detecting and measuring the excitation level of a sports commentator, we have been able to identify the "most exciting" events in a sports event and bundle these together into a summary of whatever length a user requires. As a first approximation this is very simple, but effective, and we are working on identifying slow motion (which is also indicative of significant sports plays), camera long shots vs. close ups, as well as teletext analysis, each of which will further improve the quality of our summary.

Work on video summarisation and video skimming – which is like summarisation in that it involves identifying the most salient parts of a video programme and presenting them to a user, can be found as part of the VACE project, funded by the US ARDA programme [16].

While summarisation of sports and TV news is achievable, summarisation of other types will require the kind of discourse analysis which makes text summarisation so difficult, and we have a long way to go before we can do this effectively.

Finally, a set of online links to video information retrieval projects can be found at [17].

5. Challenges for Content-Based Navigation for Video Libraries

Current approaches to video navigation are based primarily around retrieval based on associated teletext but even moreso, browsing via selected keyframes. In comparison to text or other media we have not really started to explore grouping or linking together of related video clips and a summary of the different capabilities for different media can be seen in Table 1 . In Físchár, as in other projects, all our image and audio analysis techniques on video run automatically and as they are developed into robust and effective implementations they open up for us the possibility of automatically linking together related "chunks" of video, just as we have been linking between related pieces of text on the www. The pace of development of the www has shown us how comfortable and natural it is for us to browse from web page to page by following pre-constructed hypertext links, albeit that most of them are constructed manually. It is our belief that a similar kind of video navigation which seamlessly combines searching for objects, shots or scenes, browsing and following hyperlinks between related video elements, and summarisation based on generated summaries or sets of keyframes, will offer the most appropriate and effective type of navigation through video libraries. We have already achieved some of these technologies and now we know what remains to be done.

Acknowledgement: The Support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

References

1. Little, T.D.C. and Venkatesh, D.: Prospects for Interactive Video-on-Demand *IEEE Multimedia, 1(3), 14-24, 1994*
2. Salton, G., *et al.:* Approaches to passage retrieval in full text information systems. In Proceedings of ACM SIGIR, Pittsburgh, pp 49-58, 1993.
3. Flexible Query Answering Systems : Recent Advances : Proceedings of the 4th Intnl. Conf. on Flexible Query Answering Systems, H.L. Larsen (Ed), Physica Verlag, 2000.
4. Faloutsos, C., *et al.*: Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems, 3, 3/4, July 1994, pp. 231-262.
5. Wang, J.Z.: SIMPLIcity: A Region-based Image Retrieval System for Picture Libraries and Biomedical Image Databases.in *Proceedings of ACM Multimedia 2000,* Los Angeles, Ca., USA, October 30 to November 4, 2000.
6. Smeaton, A.F., Morony, M., Quinn G. and Scaife, R.: Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. in *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, Crete, September 1998.
7. Zamir, O. and Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results. In Proceedings of WWW8, June 1999.
8. Guinan C. and Smeaton, A.F.: Information Retrieval from Hypertext Using Dynamically Planned Guided Tours. in: *Proceedings of ECHT'92 (European Conference on Hypertext),* Milan, Italy, pp.122-130, D. Lucarella *et al.* (Eds.), (1992).
9. I. Mani and Maybury M.T. (Eds).. Advances in Automatic Text Summarisation. MIT Press, 1999.
10. Lee, H. *et al.*: Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video. In *Proceedings of the Fourth European Conference on Digital Libraries (ECDL*), Lisbon, Portugal, Springer-Verlag LNCS 1923, J. Borbinha and T. Baker (Eds), pp.206-218, September 2000.
11. Smeaton, A.F. *et al.*: The TREC-2001 Video Track Report. In *Proceedings of TREC-2001*, NIST Special Publication (in press), E.M. Voorhees and D.K. Harman (Eds.), 2002.
12. Wactlar, H., Christel, M., Hauptmann, A. and Gong, Y. Informedia Experience-on Demand: Capturing, Integrating and Communicating Experiences across People, Time and Space. *ACM Computing Surveys*, Vol. 31, No. 9, June, 1999.
13. IBM CueVideo Project. http://www.almaden.ibm.com/cs/cuevideo/ Last visited 19 April 2002.
14. Smeaton, A.F.: Information Retrieval: Still Butting Heads with Natural Language Processing ? in *Information Extraction*, M.T. Pazienza (Ed), Springer LNCS, 1997.
15. Sadlier, D.A. *et al.*: MPEG Audio Bitstream Processing Towards the Automatic Generation of Sports Program Summaries. Submitted to ICME Conference, 2002.
16. The VACE Project. http://www.informedia.cs.cmu.edu/arda-vace/ Last visited 19 April 2002.

17. Centre for Digital Video Processing, Dublin City University links to related projects. http://www.cdvp.dcu.ie/links.html Last visited 17 April 2002.

	Text	**Image**	**Audio**	**Video**
Searching for …				
…whole objects	Web search	Image retrieval	N/A	Use metadata to retrieve programmes
..parts of objects	Passage retrieval	Medical, astronomy & other specialist	Retrieve audio clips	Use teletext to retrieve shots or scenes Use image/shot to retrieve shots or scenes
atomic elements / object retrieval	Factoids	Works for homogeneous collections	Not widespread	Use image retrieval on frames
Grouping / linking based on …				
…independently linked pairs	"more like this"	Image retrieval	N/A	Preliminary work reported
…overall local structure	Guided tours, cluster web search results	Not done !	N/A	Not done yet
…overall global structure	Text clustering	Image clustering	N/A	Not done yet
Summarisation	Not widespread	Thumbnails	Silence detection & FFwd.	Keyframes or audio-based and restricted to sports programs

Table 1: Content operations on different media