Indexing, Browsing and Searching of Digital Video

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

1. Introduction

Video is a communications medium that normally brings together moving pictures with a synchronised audio track into a discrete piece or pieces of information. The size of a "piece" of video can variously be referred to as a frame, a shot, a scene, a clip, a programme or an episode, and these are distinguished by their lengths and by their composition. We shall return to the definition of each of these in section 4 this chapter. In modern society, video is very commonplace and is usually seen by the majority of people as television, movies, or home video produced by a video camera or camcorder. We are also familiar with video recorded from closed circuit TVs for security and surveillance, as part of our daily lives. In summary, video is everywhere, and is increasingly becoming embedded within society.

Digital video is, as the name suggests, the creation or capture of video information in digital format. Most video produced today, commercial, surveillance, domestic, and so on, is produced in digital form although the medium of video predates the development of digital computing by several decades. The essential nature of video has not changed since or because of the advent of digital computing as it is still moving pictures and synchronised audio. However the production methods, and the end product, have gone through significant evolution in the last decade especially.

Video information is multi-dimensional. It consists of the obvious visual and audio streams and in addition each piece of video will have some metadata such as title, date and time of creation, actors, characters or objects appearing within, plus a whole host of other information which can be derived automatically from the video, including camera motion, colour histograms, identification of places for shot and scene bounds, the dialogue or transcript as spoken by those in the video, classification of audio type, classification of location, faces or text overlay which appears on-screen, etc. Because it is such a rich type of information, there is a huge array of metadata which can be derived from raw video and it is this derived metadata which can form the basis for content-based access to digital video.

The fact that video is a communications medium used as a carrier of messages to people, and the intensified developments in technology and the diverse range of applications in which video is used (entertainment, security, learning, etc.), it follows that we now have access to a huge amount of video information in our work and for our leisure pursuits. This is good because free and open access to information is a desirable feature in a modern society and it also means that the ability to provide effective and efficient access to this information becomes critical. Computing and its supporting technologies such as networks, storage devices, raw number crunching capacity and human-computer interfaces, plus the huge progress in software engineering and developments in computer science, have had a massive impact on the development of effective and efficient access to information in general, as past volumes of this Annual Review of Information Science and Technology have shown. These developments are also having major impact on the development and use of video as a communications medium. Using modern computing technology and the video compression and storage formats which have been developed recently and now have widespread use, we can now capture, store, edit, transmit and play back or stream digital video information quite easily. This is because the engineering challenges associated with these tasks have received most of the attention of the research community and quite recently we have found that our computers can comfortably manage quite large volumes of digital video information. Indeed announcements of the development of video digital libraries has now started to appear in the literature and video digital libraries research is now receiving research and development funding. Papers at the major Digital Libraries conferences in the US (JCDL, 2002), Europe (ECDL, 2002) and in Asia (ICADL, 2002) now regularly feature papers on the availability and development of video digital libraries.

Just as the advent of computing and networking has led to massive developments in the creation, availability and dissemination of text documents we are now starting to see those developments in the creation, availability and dissemination of digital video information. The advent of computing has also seen massive developments in direct access to text documents. This has come about after decades of research into information retrieval with web searching

being the most obvious example of such direct access to content by users. Analogously, if we examine how access to video has been progressing as a result of developments in computing we find that we are now only just starting to see mechanisms for content access to video based on features extracted directly from the video. Content access to video is really in its infancy and such direct access to digital video information is what is covered in this Chapter.

With all this video content now available to us and the emergent demand for content-based access, plus the techniques being developed to derive metadata directly from video, we now find that this is creating the opportunity for the development of sophisticated and effective techniques for content-based access to video. In this chapter we ask what techniques are there now for such content-based navigation operations and what future developments are we likely to see. When we consider content access to text documents we have indexing, interactive user searching, user browsing, automatic summarising, automatic linking together of related documents, and so on, and here we examine whether there are analogous operations for video. However, before we look at what content access there is, we must be aware of the constraints and limitations imposed upon us because of the ways in which video information is encoded and stored. For this reason, in the following section we present an overview of video coding and standards and we follow that with a section on conventional access techniques to video libraries. In section 4 we examine how video information can be, and needs to be, structured automatically into shots, scenes and so on in order to provide any kind of really useful access. Section 5 gives an overview of how video can be searched and browsed and how browsing is such an important part of video access, moreso than for other media. In section 6 we look at how video information retrieval is being evaluated, specifically within the TREC framework. TREC is an annual benchmarking exercise co-ordinated by the National Institute for Standards and Technology which now features a specialist track or activity, on evaluating different aspects of video information retrieval. Section 7 looks briefly at how video can be streamed to platforms other than desktop computers, specifically to mobile platforms. This is important for video information retrieval as the context for information access from a mobile device is far different to the fixed desktop and it is expected that mobile platforms will present a huge opportunity for access to video information. A final section of the Chapter presents a summary of the main trends in video information retrieval research.

It is important for the reader to note that what we cover in this chapter are only the technical aspects of video retrieval, i.e. what we need to do to actually build and deploy systems, rather than the conceptual aspects of sense making from video and human information behaviours. This chapter is also related to other chapters published in the Annual Review of Information Science and Technology because they are all in some way related to information retrieval. The strongest link, however, is to (Rasmussen, 1997) which provides a comprehensive review of indexing and retrieval from still images and video IR can be regarded as an evolution from that.

2. Video Coding and Video Standards

There are many fundamental aspects to video production, including video aspect ratio, sync, horizontal and vertical resolutions, frame rates necessary to give the illusion of smooth motion, colour fundamentals and the RGB colour space, analogue video formats such as NTSC, PAL and SECAM, video performance measurements, colour test cards, etc. These are the technical aspects of video production which predate the digitisation of video as we know it and are quite comprehensively covered in (Koegel Buford, 1994) and in (Poynton, 2003).

As we know, basically, video is a sequence of images relayed at a constant speed, normally 25 to 30 frames per second, with a synchronised audio track. In order to give the illusion of smooth motion for “normal” video (i.e. with smooth camera motion and/or smooth motion of objects in the camera frame), up to 25 frames of video per second are required, each frame showing an increment in motion from the previous one. A frame rate of 30 fps is common in the US (e.g. US television) but the digital encoding of video, especially MPEG encoding, seems to have adopted 25 fps as a default setting. To display a single image of TV resolution video at 8 bits per pixel requires 0.844 MBytes of uncompressed storage and this means that 20.8 MBytes of storage is required for each second of a video at 25 fps. For a 90 minute film this equates to 112 Gigabytes ! It also has the alarming implication that a CD-ROM with a storage capacity of 648 Mb and a data transfer rate of 150 Kbytes-per-second (the data transfer rate for original CD-Roms, though with 20x, 30x and even 40x CD-Rom drives, the transfer rate is now much faster) would only be able to store 31 seconds of video, and it would take five seconds to download and display each frame. These rough calculations clearly illustrate that the display and manipulation of TV quality video on computer screens requires massive compression of the video in order to be usable.

There are several commonly available formats or standards for encoding video information each of which includes some kind of compression, and the situation with respect to interoperability among those formats is far better than is the case for interoperability among image or audio information. For both image and audio information there are literally several dozens of encoding formats, most of them based on lossy compression and the sheer number of them

makes interoperability and exchange difficult. For video encoding, the main encoding formats which all use some kind of video compression are AVI from Microsoft, Quicktime from Apple, H261/p*42 which is used for encoding TV signals over phone lines and has application in video conferencing and other visual telephone applications (Koegel Buford, 1994), and the MPEG family of standards. AVI is important because of the involvement of its corporate developers, and QuickTime is popular because it is well-supported on the Apple MAC which is used a lot in the film production industry. The most important set of standards are the MPEG family though H.261, MPEG, QUickTime and AVI are not regarded as alternatives but compliments for video encoding. These all use similar but not identical encoding algorithms and have different niches for their use and at one time it was conceivable that they will not converge into one format but all remain in place and interoperable players would become commonplace. Now, however, most people believe that the MPEG family will eventually dominate video encoding.

A fundamental part of all kinds of video compression is motion compensation, which involves identifying the motion in adjacent video frames in order to spot and transmit only the differences between frames (this does not apply to scene or camera changes of course). To do this based on pixel-to-pixel comparisons is too simplistic because cameras are seldom fully stationary and can pan or zoom, or be noisy, or have slight movements which would make pixel-to-pixel comparisons across adjacent frames unreliable. To overcome this, frames are divided into blocks and motion compensation is tested between the appropriate blocks from adjacent frames.

What makes the MPEG standards attractive is that they are a set of standards, agreed upon by large committees with a broad representation, where nobody on the committee has a particular video standard they want to push. This is because the various MPEG standards developed are far more computationally and conceptually complex than anything in place at the time that the standards are finalised. For example, MPEG-1 was finalised in 1992 but at that time MPEG-1 playback was not possible on the then standard desktop PC and it is only within the last 5 years that MPEG-1 encoding hardware has become available at reasonable cost. In order to promote the development and use of video on PCs, chip designers are known to design chip instruction sets to make video encoding and playback run faster, such as the MMX (multimedia extension) instruction set introduced by Intel into their Pentiums from 1997 onwards (Brooks and Matonsi, 1999), (Intel, 2002) and now a standard integrated feature on desktop PCs. MMX consists of 57 specific CPU instructions built into the instruction set of the Intel chip which enhances the performance of demanding numerical calculations in certain types of applications, especially video decoding/encoding (Kratchenko, 1998). This means that matrix multiplication, chroma keying, alpha blending etc., will run faster on MMX-enhanced chips because their fundamental arithmetic operations are run directly on chip hardware, which are now part of the Intel standard.

MPEG-1 encoding turns a 3D video sequence into a 1 dimensional bit stream for transmission. It uses a frame size of 352x288 pixels at 25 fps giving VHS quality at a fixed rate of 1.5 Mb/sec or just under (which is the data transfer rate for the original CD-Rom), though larger frame sizes and different frame rates can also be encoded. MPEG decoders are common and operate comfortably on PC or advanced handheld devices and can decode in real-time in software, but encoders are still usually hardware-based. Each frame is compressed by breaking it into 8x8 pixel blocks for inter-frame and 16x16 pixel macroblocks for intra-frame motion compensation. Macroblocks are then strung together to form slices which are combined into a picture. A number of pictures are grouped together (into a group of pictures, or a GOP) to form a random access unit to allow forward/rewind with no dependencies between GOPs meaning that decoding (or editing) can be done from any part of the video file.

In MPEG-1 there are 3 types of frames

1. I-frames, or intracoded frames, which are encoded block-by-block independently of the context of adjacent frames as if they were still images. I-frames are encoded with a lossy compression known as JPEG (Wallace, 1991).
2. P-frames are forward-predicted frames, which are encoded with reference to the most recent previous I- or P-frame;
3. B-frames are bi-directional predicted frames coded with reference to previous and next I- or P-frames with motion-estimation and encoding similar to P-frames;

An example of a frame pattern in an MPEG-1 stream would be the following which corresponds to approximately 1.5 seconds of video:

- I - BBB P BBB P BBB - I - BBB P BBB P BBB - I - BBB P BBB P BBB - I –

When P-frames or B-frames are encoded “with reference” to previous frames that means that each equivalent macroblock in the different frames are overlaid to test for differences between them. Under common circumstances (i.e. no change of shot in the video) the differences between equivalent macroblocks might be minor shifts to the right

or left in the case of the camera or an object moving right or left, or there might be no change in the case of a stationary camera and stationary objects in the frame, or if the camera is tracking an object. The MPEG encoder then computes the vector difference between the two macroblocks and it is this single vector consisting of a value for direction and a value for magnitude which forms part of the encoding, instead of the entire macroblock of 16x16 pixels. This helps to achieve huge savings in storage costs by taking advantage of the minor incremental difference between equivalent macroblocks from adjacent frames. There is further compression as each GOP or frame pattern of I-, B- and P-frames generates a bit stream which is further compressed using Huffman coding. A review of MPEG-1 encoding can be found in (Le Gall, 1991).

Beyond the MPEG-1 standard there is MPEG-2 with data rates of between 2 and 10 Mbps but with a greater quality of picture and variability in data rate. MPEG-2 is the encoding which is used for transmission of digital TV and for encoding movies on DVDs. The default frame size is 720x576 pixels though we cannot see this quality of resolution on our legacy analogue TV sets. The approaches used in encoding of MPEG-2, which was released as an agreed standard in 1994, are broadly the same as for MPEG-1, except for the level of detail and quality of the picture.

There was an attempt to develop an MPEG-3 standard which was supposed to cater for high definition TV but the MPEG-2 specification proved adequate for this so the development of an MPEG-3 was dropped because the development of an MPEG-4 standard had already been started. MPEG-4 is another in the family of MPEG standards which was only recently finalised after years of development. MPEG-4 is targeted at very low bit rate coding of audio-visual interactions and requires completely new approaches to encoding which are based on human-computer interactions. Part of this involves identifying objects which move in a video sequence as being coloured and textured shapes and tracking these objects from frame-to-frame and then applying a very effective shape compression to them. This is all to be done without the encoding knowing what the shapes actually represent except that they are overlaid on a fixed background. Encoding video in this way allows for the development of future multimedia applications with extended interactive functionality and access to the actual content, i.e. the objects appearing in the video. Encoding of objects also allows deconstruction and reconstruction of the video in an object layer, where the rendering of the frames is done by the client's player and can be personalised dynamically.

All this offers very interesting possibilities for interaction with MPEG-4 encoded video but we are still some way from achieving this. The MPEG-4 standard was finalised only in 2000 and as with other MPEG standards it has been finalised ahead of the technology to deliver it being available. The development of true MPEG-4 shape based encoding of general, natural scene video is a topic currently receiving huge interest in the image processing and video coding communities, but true MPEG-4 encoders for general video are not yet available. MPEG-4 encoders, players and video is currently marketed as being available but this is based on encoding a frame as an MPEG-4 background image and is not based on overlaid shapes or objects which are compressed separately, or it is based on encoding of synthetic rather than naturally occurring video. Details of MPEG-4 coding can be found in (Puri and Eleftheriadis, 1998), (Koenen, 2001) and (Avaro *et al*, 2000).

The compression achievable with the MPEG family, and with others such as AVI, Quicktime and RealNetwork's proprietary formats is impressive when we consider the amount of storage required for uncompressed video as mentioned at the start of this section. For example, if we were to record an entire human lifetime in a reduced quality MPEG-4 video format at 150 kbps it would total just over 30 Terabytes and the quality would be browsable, but still a good bit short of broadcast TV quality. When we consider the cost of disk space it is frightening to think of an entire lifetime of video stored and available online on disk space costing approximately €50,000 !

The final MPEG standard to be mentioned here is MPEG-7 which has visual, audio and content descriptor streams. MPEG-7 does not encode video per se but can be used to encode manually or automatically-derived descriptions of video which can then be used for subsequent content-based operations. MPEG-7 syntax looks like XML and MPEG-7 compliant video library systems will have the same advantages as XML-compliant information systems in terms of interoperability and flexibility. What MPEG-7 offers is a standard format for encoding a description of a video or indeed of any multimedia object. At a low level of abstraction for video, MPEG-7 allows features such as shapes, motion, texture, colour or camera motion, to be encoded and for audio, MPEG-7 can be used to encode harmonicity, timbre, dialogue and even musical notes. Features such as these can be automatically derived directly from the content (as we shall see later) but MPEG-7 can also be used to describe other metadata not derivable from the content, including aspects like creation and production information, dates, times, locations, names of actors, etc. Finally, high level semantic information such as manual descriptions of content, can also be encoded in MPEG-7. Thus, MPEG-7 can be used as an all-encompassing vehicle to describe all aspects of content (in video) and it is the fact that the encoding of content description across videos is in this standard format that offers the many opportunities for exploitation in content-based operations. Overviews of MPEG-7 can be found in (Smith *et al*., 2000) and in (Chang *et al*., 2001) and MPEG-7 will be discussed later on in the context of the TREC video track.

3. Conventional Approaches to Accessing Digital Video

Because video is a temporal medium, we can use the conventional play, pause, fast-forward and rewind techniques familiar from consumer VCR, cassette tape and DVD devices to navigate through a single video file as shown in Figure 3.1. Techniques have also been developed to allow the fast forward operation to automatically remove pauses or silences in the video and to let the user adjust the playback speed (Li *et al.*, 2000) and (Drucker *et al.*, 2002). This does allow rapid playback of concentrated areas of a video archive and users have been shown to be able to watch more video in a shorter period of time when using this as opposed to browsing a single video. However there is a high cognitive load and users do suffer from mental fatigue when doing this rapid playback.

Figure 3.1: Conventional video player control (from Microsoft Media Player)

In terms of searching through a potentially large archive of video material, much practical development work has been done on supporting the manual creation of descriptive metadata which is then used as a basis for video searching. The support for searching in large, operational video archives such as TV or specialist video libraries is almost exclusively based upon manual annotation of the video which is coupled with structural metadata. Typically, video is manually segmented into in units called *shots* and for each shot there is a text description generated by a trained professional librarian. In TV and news archive applications, aspects such as the names or people, places or objects on-screen are important, plus their interaction if any, plus some outline of camera or perspective activity. For example, the following are all valid shot descriptions from a BBC TV documentary and from a BBC TV news program:

- Pres. Bush, White House Lawn, walking towards camera, camera fixed; 90 frames.
- Pres. Reagan greets Margaret Thatcher, embrace with R gives T kiss on cheek, camera fixed, 132 frames.
- Full shot of camel loaded with baggage, desert scene, moves slowly to left, camera panning left, 165 frames
- Camera pan right and zoom out across room of teenage students sitting exam in school hall, 135 frames

Each indexed shot would have some additional automatically generated metadata such as date, time, location of the actual video (physical shelf, tape number, offset, etc.) and the combination of manual annotation and structural metadata would form the basis for user searching. Some applications will use controlled vocabularies in the manual annotations, some will have thesauri, but there is no universally agreed standard or vocabulary and every installation seems to have evolved its own description language. Users such as TV producers or researchers seeking video clips from the archive or from some commercially available stock footage from organisations such as Getty Images (Getty Images, 2002) will formulate queries as text and if they are lucky they may get some visual preview of the videos the search engine has returned, but more likely they will get the descriptive annotations on which they must make judgements on whether to retrieve the full video or not. That is the way that video information retrieval works in practice in most commercial video IR applications.

The basic idea of a video archive as a collection of independent shots indexed and retrieved by their text captions has been taken a step further by Abe and Wakimoto (1996) who describe a multimedia authoring environment which features content-based management of video. Here, there is support for applications to annotate video scenes and clips and also objects within those video segments, such as a person, and not just the video clip of that person. Video clips can then be retrieved by keyword search on the "captions" or annotations and also by navigational search, i.e. by following hyperlinks.

Clearly, video IR applications such as the ones mentioned above do not directly exploit the fact that video content can be digital and as with all digital media, great progress can be made in terms of effective access, when we can process the media directly. Two examples of this in the digital video domain which do exploit the fact that video is digital but do so in a lightweight way are the Jabber project (Kazman *et al.*, 1996) and the Video Mail application (Jones *et al.,* 1996). The Jabber application is to transparently, and without any human involvement, digitise person-to-person meetings, face-to-face, by video and also capturing the audio (speech) and then afterwards allow users to query

what happened during the recorded meetings. This is done by performing speaker recognition, namely classifying which person is speaking at each point in the video without identifying them or recognising what they have said. This can then be used to generate a meeting "signature" of the people who have spoken and that can be useful for video retrieval.

The VMR application is a retrieval system based on the VMR video mail retrieval project at Cambridge University. The video documents are about a half-dozen hours of video email messages mostly consisting of "talking heads". VMR works by using word spotting from a pre-determined set of 35 useful indexing words in the vocabulary and this was then subsequently extended to look at large and open word vocabulary indexing.

While Jabber and VMR are functionally speaking, video IR systems, they are similar to the caption-based video IR systems in widespread deployed use in TV and video archives in that they do not leverage information directly from the visual aspect of the video information that they operate on, and with the advent of video in digital format, that is their limitation. In order to really take advantage of video being digital we need to process the content directly rather than just retrieving from structural metadata, or rather we need to automatically generate our own additional metadata to describe video from which we wish to do retrieval, and we need to make that derived metadata content descriptive. We discuss some examples of this later in section 6 when we present details of how this is done in the TREC2002 video track (Smeaton & Over, 2003). Before we can do that however, we need to structure the video in some way, and how to do that is addressed in the next section of this Chapter.

4. Automatically Structuring and Indexing Digital Video

To provide anything more than linear navigation through video information we need to structure the video in some way and this needs to be done automatically. Video is made up of frames grouped together into *shots* which are defined as the contiguous set of frames taken by a single uninterrupted camera over time. During a shot, the camera may move by zooming in or out, panning to the left or right, tracking, booming up or down, tilting, or a combination of any of these motions. Shots are often grouped into logical or semantic units called *scenes* which will have some interpretation in terms of the overall story being related in the *program* or *episode* which is the complete video file. A *clip* of video is any unit which may be as large as a set of multiple shots or scenes, or as small as a shot fragment.

There are a variety of techniques for automatically dividing a video clip into shots. The task is called *shot boundary detection* (SBD) and most techniques are based on computing similarity between adjacent frames in some way and when that similarity drops below some threshold then that is indicative of a likely shot change. This is in fact equivalent to content-based image retrieval (CBIR) which is itself an important kind of information retrieval (Maybury, 1997). However, as of quite recently, CBIR techniques are finding great application in the video SBD task. Some SBD techniques are based on extracting each frame as an image from the compressed video stream and these include comparing colour histograms for the frames (Bouthemy, 1996) and performing edge detection and comparing edges across frames (Zabih *et al.*, 1995). Other techniques which operate directly on the compressed MPEG video files are based on colour blocks (Boon-Lock and Lin, 1995) or based on examining motion vector patterns (Arman *et al.*, 1993) and will run much faster than those that require image decoding such as histogram or edge detection methods.

In terms of effectiveness of the SBD task, there have been several reported comparisons including (Boreczky and Rowe, 1996) and (Browne *et al.,* 2000) and many others, all reaching more or less the same conclusion. The SBD task is now one of the tasks supported in the annual TREC benchmarking exercise, as described in section 6 of this chapter. The techniques mentioned above are effectively image retrieval (Maybury, 1997) and SBD works reasonably well when the video has shot bounds which are known as *hard cuts*, i.e. the first frame of the new shot follows directly from the last frame of the preceding shot. However, much video is post-produced to include more gradual transitions from one shot to another in order to make the video aesthetically more pleasing (Myers, *et al.*, 2001). Fade-in and fade-out, dissolving, morphing, wipes and many other such chromatic effects are surprisingly commonplace in TV and in movies, occurring especially in gardening and cookery programmes, in live sports transmissions when introducing action replays, in TV adverts, in educational and training materials, etc. If a gradual transition takes place over, say, a 4 second period, then the incremental difference between frames during the transition will be quite minor as the overall transition will span 100 frames at 25 fps. Some SBD techniques such as edge detection tend to do well on gradual transitions but are more liable to have false positives during soft, or out of focus shots. On average, and over most video types, SBD seems to be a fairly well-solved problem, and tends to have precision and recall percentage figures in the low to mid 90s. To improve accuracy even further by combining all techniques into one unified process is regarded as not being worth the computational effort for the small payback in improved performance (Browne *et al.*, 2000), though some groups have found it useful (Zhong and Chang, 2000).

Automatic grouping of shots into logical scenes is currently attracting much attention but is proving to be difficult except in well-structured video domains. Broadcast TV news normally has a well-defined structure involving introductory credits followed by a series of separate stories, each of which will involve an anchorperson in a studio panoramic or close-up view plus perhaps some outside footage, live studio interviews, telephone interviews with a static picture or image on-screen or duel picture interviews between an anchorperson and somebody remotely. Higher-level analysis of broadcast news effectively equates to segmenting the broadcast into story bounds and this has received much investigation (Stokes *et al.*, 2002), (O'Connor *et al.*, 2001). While not quite yet a completely solved problem, this kind of segmentation can be done fairly reliably. The segmentation of other types of less well-structured video into scenes is much more problematic and part of the difficulty lies in the fact that scenes are hard to define anyway. A scene in a TV drama, movie or comedy program is an artefact inherited from plays and dramas which are staged live and which need to have this well-defined structure. Recorded video, whether digital or not, has less need for strict boundaries on logical scenes and so automatically segmenting video into such higher-level units will always be difficult.

Once a video has been segmented into shots it can be browsed by choosing a single frame as a keyframe from those shots which are longer than a certain threshold duration. Keyframes are intended to be selected as representative indicators of the shot from which they are drawn and there have been attempts to develop sophisticated algorithms for keyframe determination based on finding the frame which is closest in some image retrieval sense to the average set of frames from a shot. On the other hand, the keyframe can be chosen simply as the one in the middle, the start or the end of the shot. The danger of choosing from the start or end of a shot is that there is a likelihood of picking up artefacts from the previous shot if there has been gradual rather than a hard shot transition. There are also some heuristics which can be incorporated into the keyframe selection process. For example, if a video producer has decided to focus on some object by zooming in or panning the camera to that object then frames at the end of the zoom or pan are more likely to be meaningful and should be chosen. To date, there have been no studies reported of which techniques for keyframe selection work best in practice. It is regarded as a "black art", and the most common approach involves choosing the keyframe as an I-frame from the middle of the shot. There is certainly some investigative work required here.

Shot and scene boundary detection are part of the indexing process for video and while shot/scene bounds are almost universally used, there are other indexing primitives which can be computed and used in retrieval. These include using speech recognition to determine the dialogue spoken (Witbrock and Hauptmann, 1998), classification of the audio into speech/music (Jarina *et al.*, 2001), speaker segmentation (Roy and Melamund, 1997), camera motion (Stein and Shasua, 2000), amount of object motion in the shot, detection of slow motion as in sports action replays, face detection (Rowley *et al.*, 1998), detection of the number of faces on-screen, face recognition, overlay text detection, segmentation and OCR, primitive object recognition, and classification of a shot into indoor/outdoor or cityscape/landscape scenes. The potential list of these primitives is huge and feature detectors can be developed for domain-specific video. For example, (O'Connor *et al.*, 2001) report a technique for detecting the appearance of an anchorperson in TV news broadcasts. Many of these feature detectors can work directly in the encoded (compressed) domain and are thus very efficient while some of the visual processing techniques require decoding into images. The trend in the field is to encode these automatically detected features into MPEG-7 for greater flexibility and interoperability and to decode directly from encoded domain where possible.

In the next section we shall examine how video which has been structured into shots and for which some other features may have been detected, can be searched and browsed.

5. Searching, Browsing and Summarisation of Digital Video

In this section we examine how people can search, browse or otherwise navigate through libraries of digital video information. We start by looking at how video libraries are managed in large commercial applications such as video stock footage agencies and most TV archive departments. Most such organisations resist publishing details of how they index and support user navigation through their archives but a recent report by Enser and Sandem (2002) presented an analysis of user search and information needs in a video search environment. This showed that current searching is based around searching for specifically named persons, objects, places and events and is more search than exploration of a video archive.

Indexing video materials by manual annotation of shots as described in section 3 of this chapter, is sufficient to support the kind of retrieval that is presently used in applications like those mentioned above, albeit that it is expensive, not scalable, and not always very effective. Thus video navigation could be made much better as the techniques mentioned above are limited.

Video navigation based directly on video content can be broadly divided into searching, browsing, and summarising and although the three tasks are distinct, in experimental systems they tend to be woven together almost seamlessly. Video searching is based on matching a user's query or information need against a video database which has been structured or partitioned in some way. Most commonly, the video will have been structured into shots and retrieval of "relevant" or matching shots is one of the tasks in the TREC video track since it started in 2001 (TREC, 2002). TREC and the video retrieval track within TREC are described in more detail in the next section of this chapter. The video shot, or unit of retrieval, can be matched against a searcher's query in a number of ways. The simplest, and most common, is to match the text of a user's query against a transcript of the spoken dialogue. This is effectively spoken document retrieval (Sparck Jones *et al.*, 1996) but with pictures and has been well-explored elsewhere. Shot retrieval based on matching against the text transcript of the spoken dialogue is based on using the closed captions now provided with most TV and movie broadcasts if it is available, or is based on speech recognition of the audio (McTear, 2002). Even with a word error rate as high as 50%, spoken document retrieval can still be very effective (Sparck Jones *et al.*, 1996) and thus so can shot retrieval based on recognised speech.

The best example of a video retrieval system which supports transcript searching or searching through closed captions is the landmark Informedia system developed at Carnegie Mellon University (Hauptmann and Witbrock, 1997). This is a most successful pioneering video retrieval project which integrates video, audio and other feature extractions to provide video search and browsing facilities on a very large archive of broadcast TV news. Wactlar *et al.* (1996) describe the Informedia project while the highly accurate speaker-independent speech recogniser is described by Witbrock and Hauptmann (1998). Development of the Informedia system is still on-going and new advances are still being made by the CMU group (Informedia, 2002).

Another video retrieval system which supports searching through recognised audio transcripts or closed captions is the CueVideo system developed at IBM Almaden (Poncelon *et al.,* 1998). Here the domain of the video application is video-recorded technical talks and presentations, the same application domain as the video retrieval work at the FX Palo Alto Laboratory (Foote *et al*., 1999).

Video shot retrieval systems have also been developed based on images as the user queries or even based on full shot-to-shot matching. Image-to-shot retrieval can be based on matching the query image against a shot keyframe using conventional image retrieval approaches (Maybury, 1997) and there is great potential to match shots and ultimately lead to video retrieval based on other shot features such as camera movements, objects or object movements, or combinations thereof, though this is beyond what is currently feasible in large scale. Other reported work can match shots based on 3-dimensional colour corellograms (Darwish *et al.*, 2002) which are like 2-dimensional histograms from still images but spread over all frames in a shot, and shot retrieval can even be based on virtual 3-dimensional models of the real world derived from a shot which has camera movement (Schaffalitzky and Zisserman, 2002). Techniques such as these are experimental for now and do not work on a large scale, but development in these areas is ongoing. Ultimately this will lead to video retrieval where the user provides a sample video shot as part of their query and while not all user's information needs have sample clips as part of their formation, some do, and this will empower such searching.

While searching though video can be enabled based on matching a user's query directly against the video content, it can also be supported by allowing searching on attributes automatically derived from the video. A transcript of the spoken dialogue is the most obvious example of this but systems have also been developed which support search through other features. Figure 5.1 shows a screendump from the Físchlár system developed for the TREC2002 video track. Here we can see in the top left part of the screen that the user has requested shots which have "healthy office environment" as the query to the spoken dialogue, have some speech dialogue rather than monologue or music, and have a face on-screen and an indoor location, though some of these parts of the query are hidden behind selection tabs. The search has resulted in a ranked list of video *programmes* shown in the bottom left quarter of the screen with the top-ranked program being "How Much? Ca. 1963" which is 7 mins, 32 seconds in duration. The searcher has chosen to examine this programme in more detail and the right side of the screen shows a ranked list of shots from that programme, each shot represented by a keyframe, a transcript of the dialogue spoken (including speech recognition errors), and some iconic indicators of features automatically extracted for those shots. For example, the top-ranked shot has the keyframe with text "How Much!" emblazoned across the screen (note the TV with caption icon), has music in the audio (note the music icon), and has the dialoge transcript "GREETING HEALTHIER OFFICES AND FACTORIES WE KNOW ALL THIS ABOUT (sic!) HOW MANY PEOPLE KNOW". The vertical bars to the right of the shots indicate the strength of matching in the "people", "location", "audio" and "text" categories for each shot, and the wide horizontal bar above the shot listing is an iconic representation of the degree of similarity between the query and different shots, spread across the entire program. This can be used for navigation through the program, and the buttons on the top of the screen allow the shot listing to be ordered by degree of match based on people, on location,

etc. This system (Lee *et al.*, 2002) is an example of one which allows the user to use automatically-derived video features, marked up in MPEG-7, as part of shot retrieval.

Figure 5.1: Searching interface of Físchlár-TREC2002

When users have information needs sometimes their needs can be formulated as a query in which case a video IR system can be of use, but sometimes an information need is so broad and vaguely defined that we need to *browse* instead of perform directed searching. If a keyframe is an indicator of the contents of a shot, then browsing through sets of keyframes can be used to quickly *gist* the contents of a video program. This browsing can be through some narrow set of search results or can be through a large video segment. Most video browsing implementations present sets of keyframes in a readily-absorbed format and Lee *et al.* (2000) have defined a categorisation of such keyframe browsers. There has been little work published which actually measures the effectiveness of keyframes as a content indicator, but recent work by Goodrum (2001) has addressed this on a small scale. An example of one such browser can be seen in Figure 5.2 which is a screendump from the Físchlár system (Lee and Smeaton, 2002b). Figure 5.2 shows a library of video programmes on the left side of the screen from which the user has selected the last program "Informatics DVD". The main part of the screen shows 24 keyframes taken from adjacent shots and above this is a navigation bar allowing the user to move from this page of 24 keyframes to the next page of keyframes, and so on. The page of keyframes currently on display is highlighted on this horizontal bar.

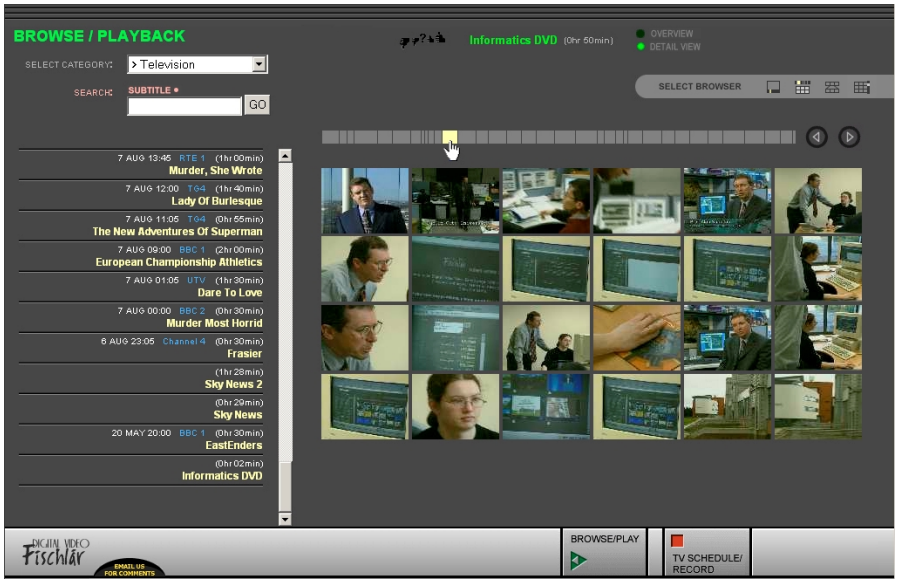


Figure 5.2: Browsing interface of Físchlár-TV

There is a substantial body of work on developing operational browsing interfaces to libraries of digital video information from the Baltimore Learning Community and the Open Video Project at the University of North Carolina. This work is summarised in a recent article in D-Lib Magazine (Marchionini & Geisler, 2002). The Open Video website provides access to over 1,600 digitised video segments and provides browsing interfaces to these also (Geisler *et al.*, 2002) and is another good example of what is currently feasible in terms of browsing and searching archives of digital video.

The third content-based video navigation operation that we are interested in is automatic summarisation of video and this can be done independently of any given user query or information need. Some of the earliest work on summarising videos was reported by Rorvig (1993) who generated summaries of NASA video materials while Elliott and Davenport (1994) described a novel 3D summary of a (short) video which showed shot bounds and object movement. More recent work on video summarisation has proved to be successful for sports videos ranging from outdoor football (Sadlier *et al.*, 2002) to indoor snooker (Denman *et al.*, 2002) where the excitation level of the crowd or commentator's excitement, plus the detection of sports events such as goals or snooker balls being potted, can point to highlights.

For videos from other domains, pulling together clips of a feature film or movie selected based on identifying the clips as containing text, dialog or gunfire/explosions, can yield a video abstract, analogous to a text abstract, and this is what the MoCA video abstracting system developed at the University of Manheim in Germany, does (Lienhart *et al.*, 1997. A video (movie, TV sitcom, documentary, etc.) is divided into scenes, each of which is made up of shots, and video abstracts are short clips containing the essence of the longer video. This is achieved by detecting special events such as the on-screen presence of the principal actors, processing of the dialogue, detecting action sequences, and other heuristics such as detecting the title music, preferring short dialogue scenes, etc. This is clearly useful in applications like multimedia archives, making movie trailers, home entertainment, and so on. Some work has been reported by Wildemuth *et al.* (2002) which evaluated the effectiveness of several video surrogates or summaries, for a variety of content retrieval based tasks such as gisting and recall. Video abstracts or gists such as those generated by Wildemuth *et al.* (2002) can be used for different purposes, documentaries to give an overview, movie trailers to entertain but not give the storyline away, etc. but as the authors state, video trailer construction is an art, and not a science, and we are a long way from automating this process for effective video summaries.

Automatic summarisation of video can also be done in the context of a given query where the results set of matching shots or scenes is to be aggregated and summarised in some way. The best example of how this can be done comes from the aforementioned Informedia project at CMU (Hauptmann and Witbrock, 1997), where the set of shots retrieved as a result of a user's query, can be organised and presented in terms of the faces of the people who appear on-screen or the most important named entities in the dialogue, the worldwide geographical locations referred to in the videos, the timeline for when the news stories were broadcast, as sets of keyframe summaries as filmstrip views (Christel *et al*., 1999).

The query-dependent video summarisation techniques present in systems such as Informedia are quite elaborate and dependent upon the sets of features extracted from the underlying videos but these various ways in which video content can be "sliced and diced" are under-exploiting the potential that exists for automatically linking video content. The twin operations of search and browse are well-established in text-based navigation where we have pre-constructed hyperlinking of the text documents and we browse through document sets by following these document-document links. This is commonplace for those of us who regularly use the web. For video, such automatic linking is now feasible, but only recently, and few research groups seem to be advancing this concept. The idea here is that in addition to supporting searching through video archives, there also exist some automatically constructed links between video clips that can be followed by users as they navigate the video library. This is distinct from the task of browsing through keyframe sets from *within* a video clip as described earlier. The reason why this is now possible is because video is so information-rich. The automatic detection of features like on-screen text, faces, objects, camera motion, speaker segmentation and so on, is seen to be a useful aid to help users search video libraries by providing a view on other facets of video content but this rich description can also be used to help in automatically creating links between related video clips across programmes. This appears to be a challenging but useful way in which video information retrieval can develop.

Finally, before wrapping up this section on video searching, browsing and summarisation, it is worthwhile to remind ourselves of the limitations on the processing we can perform on digital video because of the way in which it is encoded. As mentioned earlier in this chapter, all digital video formats use some kind of motion compensation to achieve massive compression, in order to make data sizes manageable and any kind of analysis which operates on the compressed rather than the uncompressed domain will be far more computationally efficient. For video encoded in MPEG-1 or MPEG-2 which has camera and/or object motion, the I-frames are the ones which have the highest level of

quality because they are effectively JPEG images, and the quality of the frames then slowly degrades as we move through P- and B-frames before arriving at the next I-frame. That means that I-frames rather than other frame types, should be the ones chosen as keyframes. It also means that video analysis which requires fine-grained processing of the image should be targeted at those I-frames in preference to others.

For MPEG-4 encoding, clearly effective and efficient object segmentation and tracking is critical if MPEG-4 encoding is to really take off, and this is one of the hottest topics in the image processing area. Object-based encoding is then, in turn, the key to object based interactions between people and video libraries including searching based on using an object in a picture as opposed to a full picture/keyframe as the query. As object based searching becomes achievable then object based linking and object based browsing becomes possible. Initially this may only be implementable on synthetic or artificial video such as animated cartoons but it could develop on to natural video from there. An interaction scenario with such a system might involve a user beginning a search by pointing at an object on a screen – such as a car or the Eiffel tower or a butterfly, and using that object as a query to find other video instances with the same or similar objects. Such interactions are indeed a huge advancement over what we currently have in place.

6. Measurement and Evaluation of the Effectiveness of Video Information Retrieval

Information retrieval is a discipline that has always been a mixture of the theoretic and at the same time the empirical, and we tend to regard contributions from both these ends of the research spectrum as equally important. That is one of IR's strong points and in any report on information retrieval from video information, the issue of measurement and evaluation of the effectiveness of such retrieval, must be addressed.

Without doubt, the greatest impact on practical evaluation and measurement of a range of information retrieval tasks has been made by the series of TREC conferences, co-ordinated by the National Institute of Standards and Technology (NIST). TREC (Text REtrieval Conference) is an annual activity which benchmarks the effectiveness of a variety of information retrieval tasks and which has been ongoing for over the last decade (Voorhees, 2001). These tasks have included retrieval on text documents, documents in a variety of natural languages, spoken audio documents, web documents with hyperlinks, documents corrupted by an OCR process, documents in one language with queries in another language, and so on (TREC, 2002). In 2001, TREC included a "track" or activity which explored different approaches to searching through a collection of digital video information. This was followed by a more elaborate track in 2002 with greater participation and a greater range of specific video retrieval tasks.

The goal of the TREC video track is to promote progress in content-based retrieval from digital video by using open, metrics-based evaluation within a collaborative framework and using publicly available video. Participating groups were asked to index a test collection of video data and were asked to return lists of shots from the videos in the test collection which met the information need for a set of topics distributed to all participants. The boundaries for the units of video to be retrieved were not predefined in the first year but a common set of shot boundaries were used in 2002, making evaluation easier. The sets of shots returned by the participants were then pooled together and manually assessed for relevance by the TREC co-ordinators. With this ground truth of relevance judgements available, the effectiveness of each of the participating groups' submitted results could then be measured in terms of some variation on precision and recall.

The TREC2001 video track had 12 participating groups and was divided into two distinct tasks namely shot boundary detection and searching (Smeaton *et al.*, 2002a). The shot boundary detection task involved automatically structuring the video into shots, a task described earlier in this Chapter. The searching task involved running real user queries against the video collection and what made the queries particularly challenging was that they were true multimedia queries as they all had either video clips, images, or audio clips as part of the query, in addition to a text description. These query topics were designed as multimedia descriptions of an information need, such as someone searching an archive of video might have in the course of collecting material to include in a larger video or to answer questions. While this could be done largely by searching associated descriptive text created by a human when the video material was added to the archive just as is done in many TV archives, the TREC track's scenario envisioned allowing the searcher to use a combination of other media in describing his or her information need.

The TREC2002 video track (Smeaton 7 Over, 2003) had an even greater participation with 17 groups taking part and the tasks involved (1) shot boundary detection as before, (2) searching through a video archive larger than the previous year and (3) automatically detecting some from a set of 10 different features directly from the video (TREC, 2002). These features include the presence of faces on-screen, classification of a shot into indoor or outdoor, landscape or cityscape, or classifying the audio into speech or music, the presence of text in a shot as part of a text overlay

(caption) or as part of the video (writing on a shop front for example), and the automatic recognition of the spoken dialogue.

In 2001, 11 hours of MPEG-1 video and 74 topics were used as the search collection and this was increased to 40 hours in 2002. Participating groups in the TREC video track used a variety of techniques to match multimedia topic descriptions against the video collection, some running fully automated matching techniques and others involving users in interactive search experiments. Topic descriptions in the TREC video track were true multimedia topics, including text, image, audio or even video clips as part of the topic description. Participants were free to use whatever indexing and retrieval techniques they wished, had available, or could develop, though the search task was divided into two distinct classes, one for interactive retrieval which involved some human in the search loop, and one for automatic retrieval where the retrieved shots had been matched and ranked by a system automatically.

As might be expected for the first running of an evaluation framework with so many complexities, the TREC2001 results are most useful only for small-scale comparisons - within-topic and between closely related system variants. Results from the TREC2002 video track are more useful in terms of allowing cross-system comparisons and comparisons across different approaches taken. In terms of absolute performance results and comparisons however, a multiplicity of difficulties in terms of aligning submitted shots determined by different groups, differing frame numbering caused by different MPEG decoders user, and user interpretation of the meaning of relevance of a video shot in terms of a query image or video clip, meant that the performance results were not as good as was hoped. A review of the TREC2001 video track activities can be found in (Smeaton *et al.*, 2002b)

For TREC2002 there was a greater degree of collaboration among participating groups with the features mentioned earlier being detected by a number of groups who shared their detection results with any other groups in the track who could use them. This sharing of detected features was done by exchange of MPEG-7 encoded descriptor schemes and the TREC video track actively promotes the adoption of MPEG-7 among participating groups to facilitate easy information exchange.

It is likely that the TREC video retrieval track will continue for some more years, with an increase in the size of the data collection, the number of participants, and the complexity of the search task. What the TREC video track has demonstrated to date, however, has been that collaborative research in video retrieval can be supported, on a world-wide basis, and that there are several research groups who have the capacity and experience to develop effective, scalable video retrieval tools. However, there is a danger that the TREC video track becomes entirely technology-focussed or focussed on just one or a small number of the different aspects of the complex entity that is video navigation. A broad perspective on video navigation and a list of criteria for measuring the impact of automatic video (or any media) content analysis and retrieval algorithms has been provided by Chang (2002) and this broad perspective should always be borne in mind.

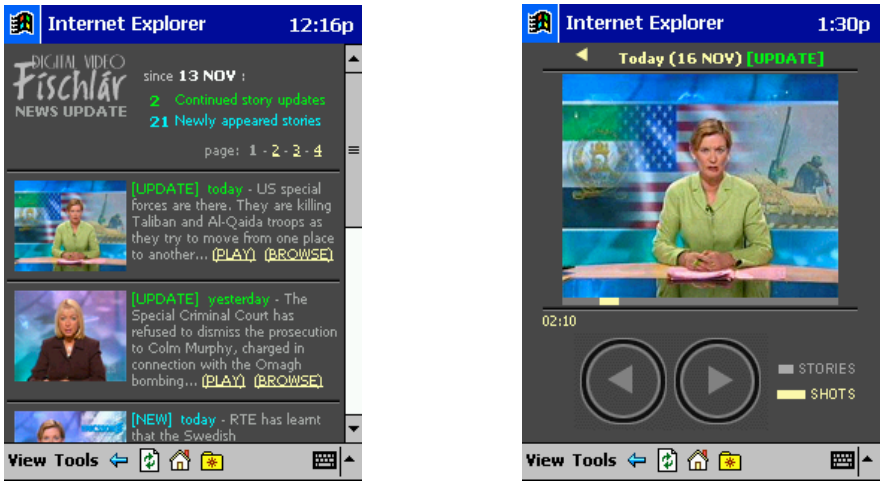
7. Mobile Platforms for Video Access

The development of mobile computing devices and infrastructure to support networked access from such devices is one of the topics in research and development receiving much attention at present. 3G (Kaaranen, 2001) and GPRS networks (Kavanagh and Beckmeyer, 2002) already deployed in parts of Europe and in widespread use in Japan, are creating opportunities for applications which require high bandwidth and which provide users with useful services. Access to, and streaming from video archives is one such application and the mobile platform and context in which users' access such archives creates a need for effective content-based access to video information. Such access will vary from people wanting to see the latest episode of their favourite soap, highlights of sports programmes, or updates on breaking news, and for all this there will be a need for high-quality video summarisation, and for high-quality information retrieval and matching. Whether this is a case of the development of the technology leading the development of the applications, or the demands of users forcing the direction of technology development is not known – indeed it may be a little of each – but the bottom line is that the topic is receiving a lot of interest and development.

In designing an application that provides mobile access to video archives we need to be aware of the limitations that mobile platforms present. The obvious ones are small screen size and absence of a keyboard, but there are others. Mobile users are normally not able to concentrate on the handheld device for long periods of time, certainly not as long as on desktop machines, and have a more limited timespan for its use. Mobile users are open to distractions and interruptions as they are on the move. Finally, mobile users generally cannot perform multiple tasks on the handheld platform compared to working on a desktop environment (Brown and Jones, 2001).

In terms of developing any system for a mobile device which is to support searching and information retrieval tasks, all these limitations point to more pre-processing on the system's side in order to determine which pieces of information a particular user will most likely to want to see at a given point in time. This encourages the development of systems which proactively recommend a particular piece of information (or pointers) to the user, and consequently demand less interaction on the user's part. Though the current literature alerts to the fact that we do not have any established or known methodology on which to base an interface design for a mobile platform, there are some rough design guidelines for doing this (Lee and Smeaton, 2002a). Clearly we need to use a layout that does not require a large space by converting spatial information into a temporal format where possible, as in the RSVP system (Bruijn and Spence, 2000). We should try to minimise user input by providing yes/no selection options rather than asking for input, using simple hyperlinking by tapping and not using visually demanding browsing that requires a careful inspection of the screen. We should filter out information as best we can so that only a small amount of the most important information can be quickly and readily accessed from the mobile device and this can be achieved by using personalisation and recommender systems (Smyth and Cotter, 2000). Finally, a system on a mobile device should proactively search and collect potentially useful pieces of information for a user and point these out, rather than trying to provide full coverage of all information via an elaborate searching/browsing interface.

If we follow these guidelines when developing a video information retrieval system then this leads us to a completely different kind of video IR system than one developed for access to the same material from a desktop environment. One example of this is the interface developed for the Físchlár system mentioned earlier, which runs on a mobile platform and is used for accessing an archive of broadcast TV news. The system is known as mFischlár (Lee and Smeaton, 2002a) and runs on a Compaq iPAQ using a wireless LAN and some screens taken from the system are shown in Figure 7.1. Here we can see, in (a), a personalised summary of the TV news with 9 news stories recommended for the user's attention, 8 of which are new and 1 of which is an update on a news story that this user is already aware of. Figure 7.1(b) shows the user flicking through the keyframes generated for a particular news story using the -> and <- (right arrow and left arrow) buttons, and by simply tapping on a keyframe, streaming of the video to the mobile device will commence from that point in the video. mFischlár is a somewhat contrived system in that a PDA with a wireless LAN card is not a streamlined user-friendly product such as the modern GPRS and 3G enabled mobile handhelds such as the XDA, but as a demonstrator it is a good illustration of what is possible on mobile platforms.



(a) Personalised news story recommendation

(b) Within-story browsing

Figure 7.1: mFischlár-News for a PDA

In computing the personalised news summaries in mFischlár just as in any video IR application for a mobile platform we can see that the system does most of its information retrieval work at the back end in computing the personalised recommendations for each user. This requires information retrieval functionality in segmenting the news broadcast into stories, and in computing similarities between different stories based on a transcript of the dialogue as well as other video-specific features automatically extracted from the video such as anchorperson detection and speech-music discrimination. This likely to be one of the ways in which future video information retrieval will operate. We shall examine other future trends in the final section of this Chapter.

8. Trends

Predicting future developments in such a fast-moving and recently-developed area as digital video navigation is somewhat dangerous but there are some trends which we can be confident will continue. The first of these trends is that the deployment of GPRS and 3G mobile communications networks, or perhaps hotspot wireless LANs based on 802.11b in public places, will spawn video streaming applications to the handset. These may include summaries or highlights of sports events, breaking TV news, or personal communications such as video messaging but all will deliver video to the mobile platform. This will help to raise the awareness of video as a personal information artefact and partly as a result of this the volume of online and available video content will increase.

In parallel to the development of personal video we will see the development of digital libraries of more public digital video content which will be accessible via the web. These will take advantage of falling storage costs, stable video coding standards and faster networking and will create new applications and demands for video information retrieval. The present commercial environments in which video information retrieval is used such as searching through TV archives and video libraries like those of Getty Images, will start to use some of the automatic feature extraction techniques described earlier such as face detection, detecting camera motion, and so on. However, it may be some time before we see public access to broadcast TV archives. In most countries, such archives are regarded as a kind of national treasure and the biggest handicaps to improving their availability are copyright and rights management issues and these need to be worked out before their value can be exploited.

Current video information retrieval techniques are almost exclusively frame based as we have seen earlier in that retrieval of shots can be based on retrieving keyframes or shot signatures rather than retrieving based on the 3-dimensional structure of a shot. The development of object segmentation and object tracking techniques is a huge focus within the image processing community and as these techniques develop this will lead to true object-based MPEG-4 compression which will do a lot for video streaming to low-bandwidth devices such as on mobile platforms. More importantly perhaps, the development of object segmentation and tracking will lead to object based interaction for video information retrieval. If we can segment objects in a video clip and then index that clip by those objects then we should be able to search video archives with selected objects as the query instead of having to use the entire keyframe as we do now. This will lead to a fascinating analogy with the recent developments in text-based information retrieval. When searching for information in text databases there are times when what we really want to retrieve are answers to specific, narrow questions which are ultimately satisfied by simple facts. Examples would be “what is the date of birth of President Bush” or “how high is the statue of liberty”. In recent years we have seen the emergence of question-answering systems (Hirschman and Gaizauskas, 2001) which identify and then retrieve facts as answers to such specific queries. The analogy with video is that for some of the time when we search video we are seeking to find a clip based on an object in that video, irrespective of the context or background of its occurrence. An example might be to find video clips which contain a certain kind of car, or the leaning tower of Pisa, where we can provide images of the car or the tower as examples. Object based segmentation and indexing of video by such objects will open the possibility for object-based retrieval, and this will be something completely new in the development of video information retrieval.

An important part of the development of the field of video information retrieval is the realisation that user issues, addressing user information needs, assessing relevance criteria in the context of video navigation and access, are all important aspects of video information retrieval. In the early stages of development, text-based information retrieval was technology-driven before the realisation dawned that trying to understand users, their information needs, perceptions and relevance criteria were just as important. This is being realised in, for example, the context of image retrieval (Choi and Rasmussen, 2002) but because video information is so complex, and so more multi-faceted compared to other media, this will be difficult.

The final, and perhaps the most likely trend in the development of video information retrieval is the uptake of MPEG-7 as a format for describing video features. MPEG-7 allows flexibility and interoperability which are buzzwords in today’s technology landscape, and the tools and technologies of MPEG-7 are ready for widespread deployment and have been used already in indexing and retrieval of images (Jaimes *et al*., 2000). It is to be hoped that the adoption of MPEG-7 will be whole-hearted and the integration of MPEG-7 descriptions with other structured description mark-up languages such as NewsML and VoiceML (Myllymaki, 2002) will allow video to be easily integrated with other media in various applications. This depends upon MPEG-7 having a pathway by which it can enter usage in practical application domains and the fact that there are efforts at harmonising the MPEG-7 standard with other metadata standards is encoraging.

Bibliography

Abe, H. & Wakimoto, K. (1996) Content-Based Management of Video in a Multimedia Authoring Environment. *Multimedia Tools and Applications,* 2(3), 199-214.

Arman, F., Hsu, A & Chiu, M. (1993). Feature management for large video databases. In *Storage and Retrieval for Image and Video Databases*, 2-12.

Avaro, O., Eleftheriadis, A, Herpel, C., Rajan, G. & Ward, L. (2000). MPEG-4 Systems Overview. *Signal Processing: Image Communication*, 15(4-5), 281-298.

Boon-Lock, Y., Liu, B. (1995). A Unified Approach to Temporal Segmentation of Motion JPEG and MPEG Compressed Video. *In International Conference on Multimedia Computing and Systems,* 81-83. IEE, Los Alamitos, Ca, USA.

Boreczky, J.S., Rowe, L.A. (1996). A Comparison of Video Shot Boundary Detection Techniques. *Storage & Retrieval for Image and Video Databases IV*, I.K. Sethi, and R.C. Jain, (Eds.), Proc. SPIE 2670, pp.170-179.

Bouthemy, P., Garcia. C., Ronfard, R. & Tziritas, G. (1996). Scene Segmentation and Image Feature Extraction for Video Indexing and Retrieval. In *Visual Information and Information Systems*, pp245-252.

Brooks, D. & Martonosi, M. (1999). Dynamically Exploiting Narrow Width Operands to Improve Processor Power and Performance. In *Proceedings of the Fifth International Symposium on High Performance Computer Architecture,* Orlando, Fl., January 1999.

Brown, P.J. & Jones, G.J.F. (2001). Context Aware Retrieval: Exploring a New Environment for Information Retrieval and Information Filtering. *Personal and Ubiquitous Computing*, 5(4), 253-263.

Browne, P., Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S. & Berrut, C. (2000). Evaluating and Combinating Digital Video Shot Boundary Detection Algorithms. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, IMVIP2000, Belfast, Northern Ireland, September 2000.

Bruijn, O. & Spense, R. (2000). Rapid Serial Visual Presentation: A Space-Time Trade-Off in Information Presentation. In: *Proceedings of the Advanced Visual Interface Conference (AVI2000)*, Palermo, Italy.

Chang, S-F, Sikora, T. & Puri, A. (2001). Overview of the MPEG-7 Standard. *IEEE Transactions on Circuits and Systems for Video Technology (Special issue on MPEG-7)*, 11(6) *688-695.*

Chang, S-F. (2002). The Holy Grail of Content-Based Media Analysis. *IEEE Multimedia Magazine*, 9(2), 6-10.

Choi, T. & Rasmussen, E.M. (2002). Users' Relevance Criteria in Image Retrieval in American History. *Information processing and Management,* 38(5), 695-726.

Christel, M., Warmack, A., Hauptmann, A., Crosby, S. (1999). Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library. In *Proceedings of the IEEE Advances in Digital Libraries Conference* 1999, Baltimore, MD. 98-104.

Darwish, K., Doermann, D., Jones, R., Oard, D. & Rautiainen, M. (2002). TREC-10 Experiments at University of Maryland CLIR and Video. In: *NIST Special Publication 500-250: The Tenth Text REtrieval Conference* (TREC 2001).

Denman, H., Rea, N. & Kokaram, A. (2002). Content Based Analysis for Video from Snooker Broadcasts. In *Challenges for Image and Video Retrieval, CIVR200,* M.S. Lew, N. Sebe and J.P. Eakins (Eds.), Springer LNCS 2383.

Drucker, S.M., Glatzer, A., De Mar, S. & Wong, C. (2002). SmartSkip: Consumer Level Browsing and Skipping of Digital Video Content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves,* Minneapolis, Minn. ACM Press.

ECDL, (2002). The European Conference on Digital Libraries (ECDL), Rome, Italy, September 16-18 2002, retrieved 30 July 2002 from http://www.ecdl2002.org/

Elliott, E. & Davenport, G. (1994). Video Streamer. In *Proceedings of the CHI'94 Conference Comapnion on Human Factors in Computing Systems*, Boston, Mass., ACM Press.

Enser, P.G.B. & Sandom, C.J. (2002). Retrieval of Archival Moving Imagery – CBIR Outside the Frame. In *Challenges for Image and Video Retrieval, CIVR200,* M.S. Lew, N. Sebe and J.P. Eakins (Eds.), Springer LNCS 2383.

Foote, J., Boreczky, J. & Wilcox, L. (1999). Finding Presentations in Recorded Meetings Using Audio and Video Features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Phoenix, AZ), vol. 6, 3029-3032.

Geisler, G., Marchionini, G., Wildemuth, B.M., Hughes, A., Yang, M., Wilkens, T., & Spinks, R. (2002). Video Browsing Interfaces for the Open Video Project. In: *Proceedings of CHI'2002: Changing the World, Changing Ourselves*, Minneapolis, Minn., 514-515.

Getty Images, (2002). The Getty Image Video Archive. Retrieved 10 August, 2002 from http://www.gettyimages.com/

Goodrum, A. (2001). Multidimensional Scaling of Video Surrogates. *Journal of the American Society for Information Science*, 52(2), 174-183.

Hauptmann, A. & Witbrock, M. (1997). Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In: *Intelligent Multimedia Information Retrieval*, Mark T. Maybury, Ed., AAAI Press, pp. 213-239, 1997.

Hirschman, L. & Gaizauskas, R. (2001). Natural Language Question Answering: The View from Here. *Natural Language Engineering*.

ICADL, (2002). The International Conference on Asian Digital Libraries, Singapore, 11-14 December 2002, retrieved from http://www.icadl2002.org/.

Informedia, (2002). The Informedia Project Website. Retrieved 1 August, 2002 from http://www.informedia.cs.cmu.edu

Intel Corporation, (2002). Introduction to MMX Technology, Online tutorial. Retrieved 1 August 2002 from http://www.intel.com/design/perftool/cbts/mmxintro/.

Jaimes, A., Benitez, A.B., Jorgensen, C. & Chang, S-F. (2002). Experiments in Indexing Multimedia Data at Multiple Levels. *ASIS SIG Classification Research Workshop, Dea Mart: Classification for User Support and Learning*, Chicago Ill, November 2002.

Jarina R., Murphy N., O'Connor N. & Marlow S. (2001). Speech-Music Discrimination from MPEG-1 Bitstream. In Proceedings of *SSIP'01 - WSES International Conference on Speech, Signal and Image Processing.* Malta, 1-6 September 2001.

JCDL, (2002). The ACM-IEEE Joint Conference on Digital Libraries, Portland, OR, July 2002. Retrieved 30 July, 2002 from http://www.jcdl2002.org/

Jones, G.J.F., Foote, J.T., Sparck Jones, K. & Young, S.J. (1996) Retrieving Spoken Documents By Combining Multiple Index Sources. in: *Proceedings of the 19th International ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR96), Zurich, Switzerland, 30-38.

Kaaranen, H. (Ed), Naghian, S., Laitinen, L., Ahtianen, A & Niemi, V. (2001*). UMTS Networks: Architecture, Mobility and Services*. John Wiley*,* 2001.

Kavanagh, A & Beckmeyer, J. (2002). *GPRS Networks*. Osborne Mcgraw-Hill, 2002.

Kazman, R., Al-Halimi, R., Hunt W. & Mantei M. (1996). Four Paradigms for Indexing Video Conferences. *IEEE Multimedia*, 3(1), 63-73.

Koegel Buford, J.F. (1994) *Multimedia Systems*. ACM Press, Addison-Wesley Publishers, New York.

Koenen, R. (2001). Object-Based MPEG Offers Flexibility. *EETimes*, November 12, 2001.

Kractchenko, V. (1998) Using MMX Technology in Digital Image Processing. University of British Columbia, Department of Computer Science, Technical report TR-98-13.

Le Gall, D. (1991). MPEG: A Video Compression Standard for Multimedia Applications. *Communications of the ACM*, 34(4), 46-58.

Lee, H. & Smeaton, A.F. (2002). Searching the Físchlár-NEWS Archive on a Mobile Device. In *Proceedings of the Workshop on Mobile Personal Information Retrieval*, ACM SIGIR2002 Conference, Tampere, Finland, August 2002.

Lee, H. & Smeaton, A.F., (2002). Designing the User Interface for the Físchlár Digital Video Library. *Journal of Digital Information*, 2(4).

Lee, H., Smeaton A.F., Berrut, C., Murphy, N., Marlow, S. & O'Connor, N. (2000). Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video. In *Proceedings of the Fourth European Conference on Digital Libraries (ECDL*), J. Borbinha and T. Baker (Eds), Lisbon, Portugal, Springer-Verlag LNCS 1923, pp.206-218.

Lee., H., Smeaton, A.F., McDonald, K. & Gurrin, C. (2002). Design, Implementation and Testing of A Video Search System. *Computer Vision and Image Understanding,* (submitted), 2002.

Li, F., Gupta, A., Sanocki, E., He, L. & Rui, Y. (2000). Browsing Digital Video. In *Proceedings of CHI 2000.*

Lienhart, R. Pfeiffer S. & Wiffelsberg, W. (1997) Video Abstracting. *Communications of the ACM*, 40(12), 55-62.

Marchionini, G. (2002). The Open Video Digital Library. *D-Lib Magazine, 8(12),* available at http://www.dlib.org/

Maybury, M. (1997). *Intelligent Multimedia Information Retrieval*. The MIT Press, 1997.

McTear, M. (2002). Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys*, 34(1), 90-169.

Myers, B., Casares, J.P, Stevens, S., Dabbish, L., Yocum, D. & Corbett, A. (2001). A Multi-View Intelligent editor for Digital Video Libraries. *In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Roanoake, Va, June 2001.

Myllymaki, J. (2002). Effective Web Data Extraction with Standard XML Technologies. *Computer Networks*, 39(5), 635-644.

O'Connor N, Czirjek C, Deasy S, Marlow S, Murphy N & Smeaton A.F. (2001). News Story Segmentation in the Fischlár Video Indexing System. In Proceedings of *ICIP 2001 - International Conference on Image Processing.* Thessaloniki, Greece.

Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D. & Diklic, D. (1998). Key to Effective Video Retrieval: Effective Cataloging and Browsing. in *Proceedings of ACM Multimedia*, '98, 99-107.

Poynton, C. (2003). Digital Video and HDTV Algorithms and Interfaces. *Morgan Kaufman and Elsevier Science*

Puri, A. & Eleftheriadis, A. (1998). MPEG-4: An Object-Based Multimedia Coding Standard Supporting Mobile Applications. *Mobile Networks and Applications*, 3(1), 5-32.

Rasmussen, E. (1997). Indexing Images. *Annual Review of Information Science and Technology*, Vol 32. Medford, NJ: Information Today. 169-196.

Rorvig, M. (1993). A Method for Automatically Abstracting Visual Documents. *Journal of the American Society for Information Science*, 44, 40-56.

Rowley, H., Baluja, S. & Kanade, T. (1998). Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), 23-38.

Roy, D. & Malamund, C. (1997). Speaker Identification Based Test to Audio Alignment for an Audio Retrieval System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.* Munich, Vol 2, 1099-1103.

Sadlier D, Marlow S, O'Connor N & Murphy N. (2002). MPEG Audio Bitstream Processing Towards the Automatic Generation of Sports Programme Summaries. In *ICME 2002 - IEEE International Conference on Multimedia and Expo,* Laussane, Switzerland.

Schaffalitzky, F. & Zisserman, A. (2002). Automated Scene Matching in Movies. In *Challenges for Image and Video Retrieval, CIVR200,* M.S. Lew, N. Sebe and J.P. Eakins (Eds.), Springer LNCS 2383.

Smeaton A.F., Over, P., Costello, C., de Vries, A., Doermann, D., Hauptmann, A., Rorvig, M., Smith, J.F. & Wu, L. (2002). The TREC2001 Video Track: Information Retrieval on Digital Video Information. In: *Proceedings of ECDL 2002 - European Conference on Research and Advanced Technology for Digital Libraries*. Rome, Italy, LNCS-2458, 266-275.

Smeaton, A.F., Over, P. & Taban R. (2002). The TREC-2001 Video Track Report. In: NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001). Retrieved August 1, 2002 from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Smeaton, A.F. Over, P. (2003) The TREC2003 Video Track Report. In: Proceedings of TREC2002 (in press).

Smith, J.R. Puri, A. & Tekalp, M. (2000). MPEG-7 Multimedia Content Description Standard, *IEEE Intern. Conf. on Multimedia and Expo* (ICME-2000).

Smyth, B. & Cotter, P. (2000). A Personalized Television Listings Service. *Communications of the ACM*, 43(8).

Sparck Jones, K., Jones, G.J.F., K., Foote, J.T., & Young, S.J. (1996). Experiments in Spoken Document Retrieval. *Information Processing and Management*, 32(4), 399-417.

Stein, G.P. & Shashua, A. (2000). Model-Based Brightness Constraints: On Direct Estimation of Structure and Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(9), 992-1015.

Stokes N, Carthy J, & Smeaton A.F. (2002). Segmenting Broadcast News Streams using Lexical Chains. In Proceedings of *STAIRS 2002 - STarting Artificial Intelligence Researchers Symposium,* Lyon, France.

TREC (2002). The TREC Video Track Guidelines. Retrieved 1 August, 2002 from http://www-nlpir.nist.gov/projects/t2002v/t2002v.html

Voorhees, E.M. (2001). Overview of TREC 2001. In The Tenth Text REtrieval Conference (TREC 2001), National Institute of Standards and Technology (NIST). Retrieved 1 August 2002 from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Wactlar, H., Stevens, S., Smith, M.and Kanade, T. (1996). Intelligent Access to Digital Video: The Informedia Project. *IEEE Computer,* 29(5).

Wallace, G.K. (1991). The JPEG Still Picture Compression Standard. *Communications of the ACM*, 34(4).

Wildemoth, B.M., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A. & My, X. (2002). In: *Proceedings of ECDL 2002 - European Conference on Research and Advanced Technology for Digital Libraries*. Rome, Italy, LNCS-2458, 493-507.

Witbrock, M. & Hauptmann, (1998). Speech Recognition for a Digital Video Library A., *Journal of the American Society for Information Science*, 49(7).

Zabih, R., Miller, J & Mai, K. (1995). A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. In *Proceedings of the 3rd International Multimedia Conference and Exhibition, Multimedia Systems*, 189-200, San Francisco, California.

Zhong, D. Chang, S-F. (2000). Video Shot Detection Combining Multiple Visual Features. *Columbia University ADVENT Technical Report #092,* December 27th 2000. Available from http://www.ctr.columbia.edu/papers_advent/00/scene_cutTR00.pdf Last Accessed 29 December 2002.