Indexing, Browsing and Searching of Digital Video and Digital Audio Information

Alan F. Smeaton
School of Computer Applications
and Centre for Digital Video Processing
Dublin City University

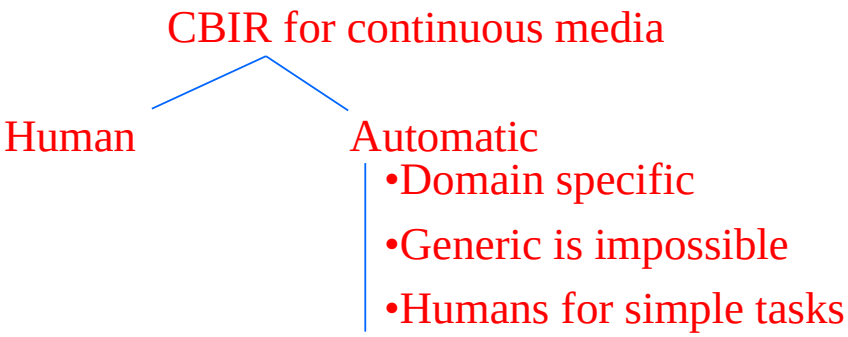
Alan.Smeaton@DCU.ie

- Good news ….
  - We can go on until 17:30 if we want
- Bad news …
  - we won't !
  - Not a mathematical formula in sight !
  - Only 43 slides
  - No more animation

1. Introduction

- Content-based operations on continuous medium data … audio and video, covering retrieval, filtering, alerting, summarisation, clustering, etc.
- Principles and trends behind CBIR of multimedia
- IR on digital audio, mostly spoken audio, and the 4 approaches taken, using Taiscealaí as example
- Digital video …formats, MPEG, segmentation, shot bound detection, keyframes, browsing, video searching
- Conclusions … for continuous medium ... browsing is important

CBIR



CBIR Principles

- General principles across all kinds of objects being retrieved is that CBIR is based on understanding of content and there are 2 approaches:
  - Human interpretation generating captions, keywords etc., but inconsistent over time, over users, no agreed format and very expensive … example manual web page classification in Yahoo !
  - Automatic interpretation is low-level, cheaper and often wrong, but consistently so
- Audio and video can be indexed using the first approach, but that is covered elsewhere this week, and we are interested in the second.

CBIR on non-text

- Current trend based on 3 ideas:
  - Successful non-text CBIR applications are domain-specific though they can be ported;
  - Generic automatic indexing tools are (to date) impossible and user intervention is required;
  - Human involvement should involve primitive tasks for consistent performance;
- This is a bleak picture and real, semantically-rich CBIR on video and on audio is far away, nothing like current text-based.
- However, considerable progress has been made

MM Object characteristics

Further characteristics of audio/video which impact CBIR:

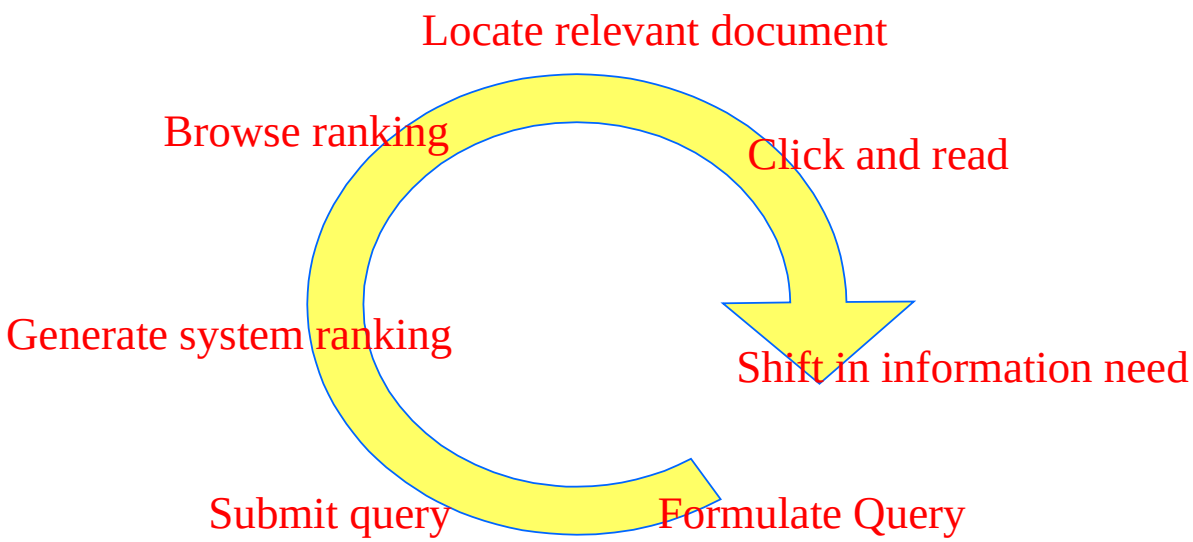
- … have multiple dimensions and how we view, our task, what we seek, etc., all elicit different properties; different features interest us at different times.
- … we may eventually require retrieval based on properties not initially captured, so we should have query-time "re-indexing", unlike with text;
- … we should develop suites of retrieval techniques for sub-groups of features based on inexact match, which can be combined into overall ranking;
- … allow for and handle indexing which is incomplete, inexact and possible erroneous;

MM Object characteristics

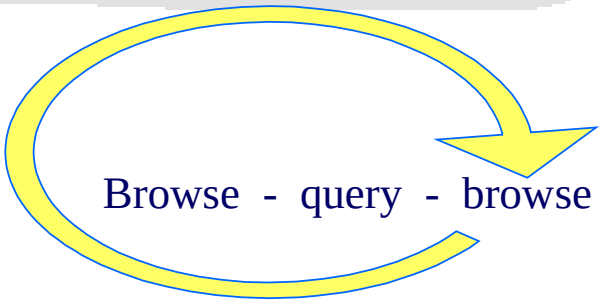
- … must understand that queries are incomplete (as with text) but because video esp. is so more content-rich than text documents, the concept of "relevance" may be much more difficult to model and capture;

Thus CBIR on video is hugely more difficult than for text, image or audio.

Typical text-based IR interaction



Typical video-based IR interaction



… comes from the fact that one cannot easily "gist" a audio/video clip compared to a text document;

… it is this inability to easily "gist" audio/video that makes CBIR from continuous media so very different to text-based or image-based IR

CBIR and encoding formats

- Final point before considering overview of audio-IR …

- Computing technology developments have made huge strides in engineering the capturing, creating, editing, storing, transmitting, rendering and displaying of multimedia;
- Foremost are the encoding and compression standards which target max. compression for min. loss of quality, without any consideration for content manipulation;
- Thus the tail is "wagging the dog" and we find ourselves wrestling with CBIR on audio/video constrained by (a) the compression formats and (b) the cost of decoding

2. IR on Digital Audio

- Digital audio … speech, music, other specialist or sound effects;
- Sound is a continuous vibration sampled at a rate leading to quantization of the analog waveform into digital format;
- Higher sampling rate - less quantization noise - better quality;
- Audio CD = 44 kHz, 16 bits, 2 channels;
- Once digitised, there are scores of encoding formats:

Audio encoding formats

- WAV is common, raw & uncompressed;
- AU, Vox, RealAudio, TSP, VMF, AIFF … all achieve compression;
- MP3 uses perceptual compression, compressing parts of the spectrum there human hearing is at its least discerning;
- MIDI, for music, represents notes from different musical instruments (pitch, duration, etc) and much work on IR from MIDI has been published;

Audio encoding of speech

- 2 utterances of the same word by same person in same place/time, will have different waveforms, so waveform matching is a non-starter because of variances in loudness, pitch, brightness, harminosity, etc.
- Speech document retrieval applications are based on some kind of recognition;
- Speech is composed of phones, where a phone is a unit of pronunciation, e.g.

  m oo r d ii t ei l z
- Phone recognition is a pre-process to word recognition and usually outputs a lattice of phones with probabilities rather than a single stream;

Recognition of spoken audio

- The phone lattice is then processed against a pronunciation dictionary to determine allowable words, with the two processes self re-enforcing;
- Word segmentation is a complication because of the way we speak … “wreck a nice beach” =/= “recognise speech”;
- Full speech recognition is speaker-dependent, expensive, and requires training in order to be in any way effective; even phone recognition requires some training;
- An approximate grouping of speech IR is the following:

1. Word Spotting

- Cambridge/Olivetti VMR application for video mail … pre-defined vocabulary of some tens of words, search for these words only;
- works because there is a pre-defined vocabulary of terms so recognition becomes feasible;

2. Speaker Recognition

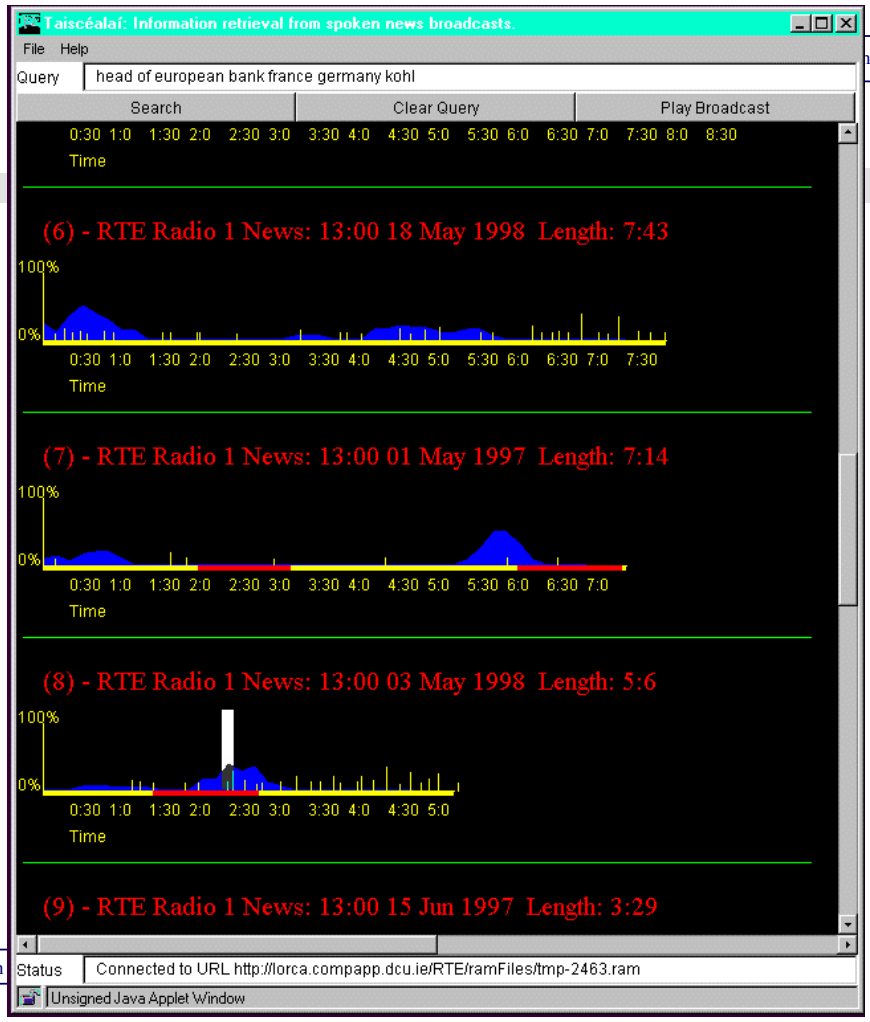
- Jabber proj at U. Waterloo is an example where speaker recognition was useful;
- Applied speaker-independent continuous speech recognition to recordings of meetings but speech recog. wasn't up to the task …
- They were able to recognise **who** did the speaking;
- A visual summary of the turn-taking in dialogue in a multi-person meeting was a sound enough basis for browsing meeting records

3. Phone based retrieval

- Instead of recognising words, recognise phones;
- ETH-Zurich was first to do this on German news radio and out own Taiscealaí system extended this;
- Taiscealaí recorded radio news, twice daily, recognised phones, divided the broadcast into overlapping windows with the phones in each 30s window being a retrievable "document";
- Phone recogniser was HTK HMM recogniser which was trained on 24 hours of transcribed broadcast, 20x realtime;
- We augmented a CMU pronunciation dictionary with new words and idosyncratic Hiberno-English pronunciations … Taoiseach, Tainiste, Drogheda, Kehoe, Maher, Taiscealaí, ...

3. Phone based retrieval (cont)

- 30s windows were represented by triphones .. Triples of adjacent phones
- User's (typed) queries were looked up in the dictionary and (bag of) phones for query terms matched against (bags of) phones for 30s windows using standard weighted IR;
- We could have gone much further with the IR aspect of this
- We extended this further to trawl online newspapers for new newsworthy terms which were flagged to be added to the pronunciation dictionary;
- Taiscealaí was operational for over 18 months;
- It stream RealAudio and looked like this ...



3. Phone based retrieval (cont)

- Taiscealaí gives us our first glimpse of IR from continuous media … instead of a ranked list of 30s windows, we aggregated window scores into scores for broadcasts and ranked broadcasts, providing within-broadcast navigation via time series of scores;
- Analogous to passage retrieval in text … but very different;
- It is more difficult to get a quick overview of an audio clip than the equivalent sized text document;

4. Word-based retrieval

- Despite difficulty of speaker independent continuous speech recognition, there are several examples of this work;
- The big catalyst behind this has been the TREC track on spoken documents;
- TREC is …
- In TREC spoken document, many groups train a recogniser and then take account of speech recognition errors in the retrieval task;
- The data is radio/TV news broadcasts, and the task is to locate a (or the) relevant story;

3. Information Retrieval from Digital Video

- Video is 25fps images and synchronised audio;
- To display a single image of TV-quality video requires 720 Kbytes, so without compression this is almost 100 Gbytes for a 90 minute movie.
- Video information must be compressed !!!
- There are some formats such as QuickTime, DVI, H.261, etc., but the ones that matter are the MPEG family, overview later;
- Before we look at IR on video we must have some understanding of how video is encoded in order to realise the limitations we face;

Video Encoding principles

- All video encoding standards use motion compensation, identifying motion between adjacent frames and xmitting only the differences … except across shot bounds;
- Doing this on a pixel basis is too fine-grained because cameras boom, tilt, pan, zoom, shake, so frames are divided into pixel aggregates called “blocks” and motion compensation is tested between equivalent blocks;
- This allows a graceful and effective encoding of deliberate camera motion, as well as object motion

MPEG family

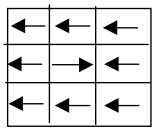
- At present there are 4 MPEG standards finished or under construction and what makes MPEG attractive is that they are standards defined **ahead** of the technology, so no proprietary material included;
- MPEG groups are open and have large participation;
- MPEG-1 around longest and most common … details shortly;
- MPEG-2 is higher quality but same principle as MPEG-1, requires hardware to encode and decode;
- MPEG-4 is specified but technology not yet developed except in test environments;
- MPEG-7 under development, includes a stream to describe content, which will be something like XML;

MPEG-1 encoding

- 352x288 pixels per frame @ 25 fps giving VHS quality at 1.5 Mbps; can be decoded in realtime on an (old) PC, encoding requires hardware;
- MPEG stream has I-, P- and B-frames in a given pattern;
- I-frame is a JPG image; each frame divided into 16x16 pixel macroblocks (22x18 of them) and in B-frames and P-frames, which follow I-frames, equivalent macroblocks are compared and a motion vector generated if possible;

MPEG-1 encoding

- I, P and B frames form a pattern depending on the encoder used … ours has an I-frame every 12 frames (2 per sec) but it does not have to be like this;



- Encoders are not perfect and the 396 motion vectors in a frame can sometimes be incorrect and have rogues;
- The pattern of different frame types allows random access, FF, REW and reverse playback;
- MPEG-2 is 720x576 pixels and is used for digital TV;
- MPEG-4 is object based compression, based on identifying, tracking and encoding object layers which are rendered on top of each other, with huge potential for video interaction;

Information Retrieval on video

- It is straightforward to treat video as a binary blob and index/retrieve via its metadata … but that's not for now;
- Most work on IR on video streams, has concentrated on the visual stream … IR on the audio stream defaults to being audio-based IR … here we look at the visual stream;
- The way to make progress with video IR is to structure the video and above the frame, the next basic unit is the shot;
- A shot is a sequence of frames from a single camera motion over time;
- Structuring a video begins by identifying shot boundaries automatically;

Structuring video

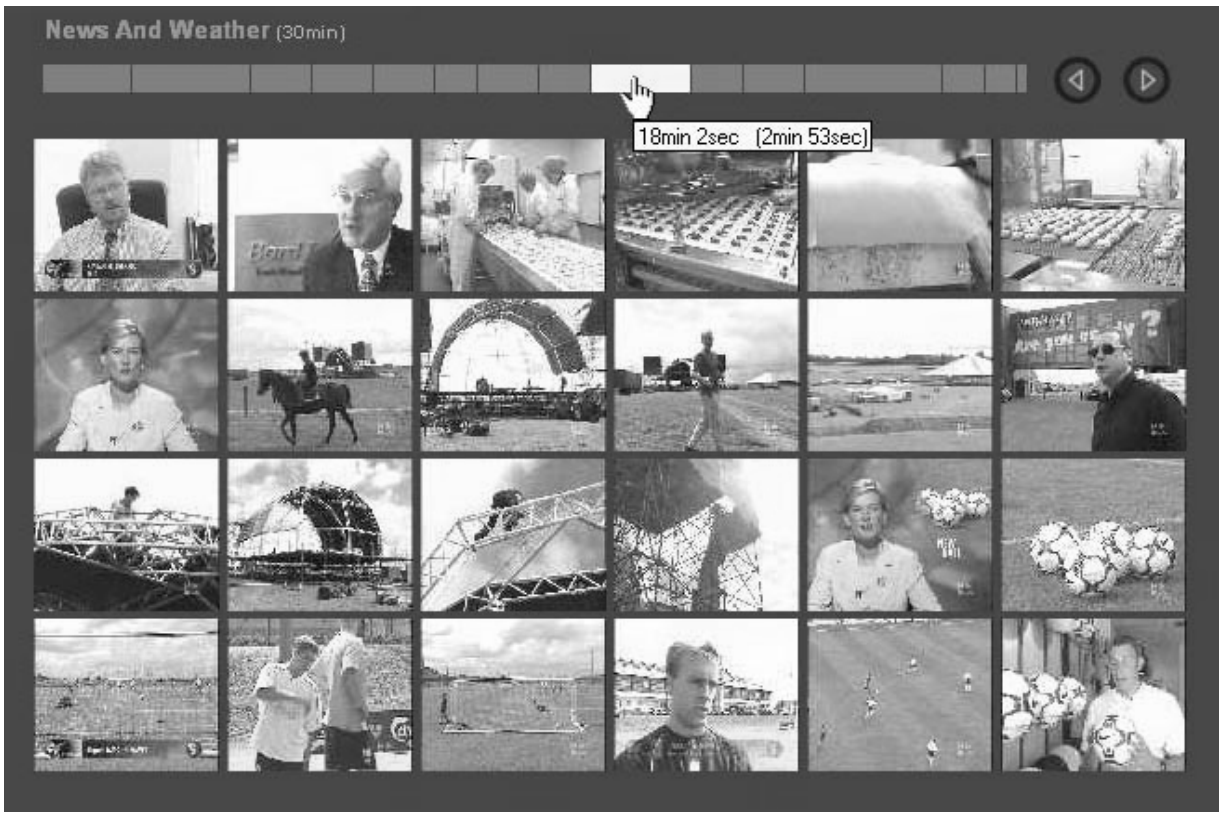
- Usual approach to SBD is to compare adjacent frames to see if they are very dissimilar … if so, then that is a likely shot bound.
- For hard cuts, the most usual technique is to compare adjacent frames based on colour histograms but shots can be joined using more sophisticated techniques like fades to black, dissolves, wipes or computer-enhanced “swooshes” !
- For these, colour histograms are less successful because the shot transitions occur over time, i.e. over multiple frames;
- Other SBD techniques are based on edge detection (good for dissolves), or from the encoded stream, based on macroblock types or on motion vectors;

Shot boundary detection

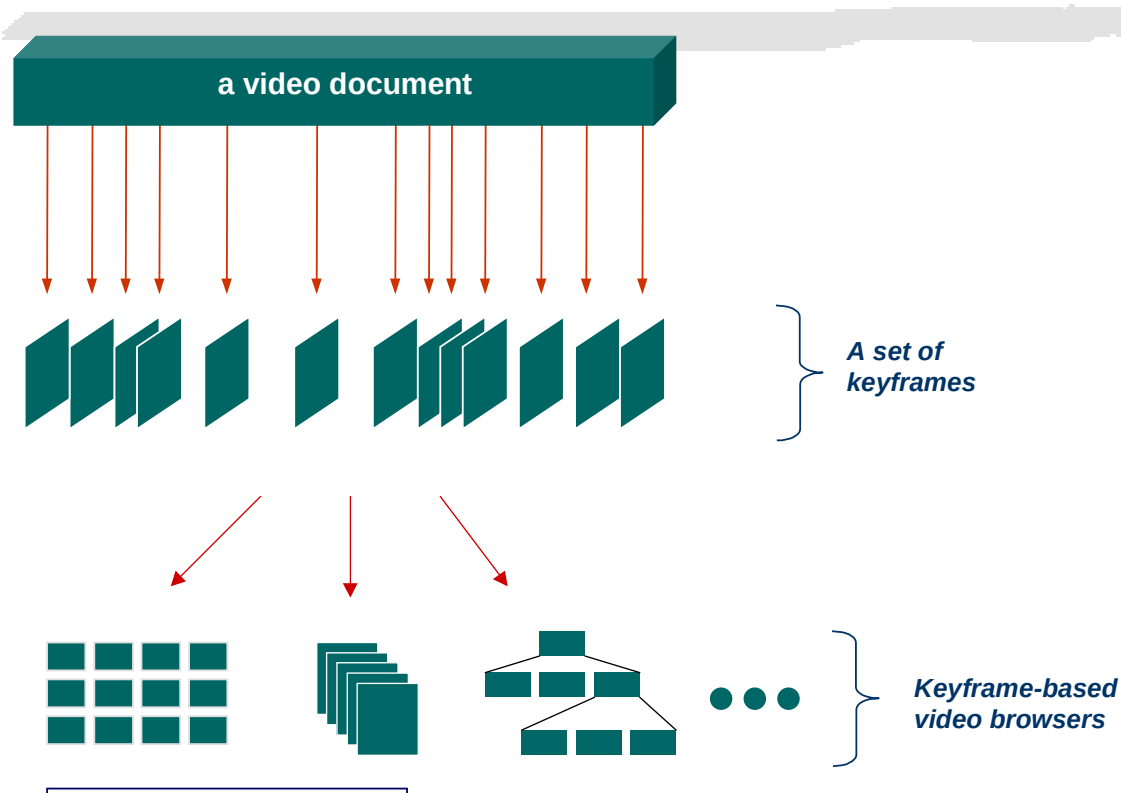
- We have evaluated all these SBD methods on a collection of 8 hours TV broadcast (720,000 frames, manually marked up), individually and in combination … best individual is 85% to 90% in both precision and recall, best combination is only a couple of % above;
- Incorporating audio could help, or not … silences between some shots, but not all
- Computational cost of SBD must be considered … most run in over real time, those on the compressed domain are quicker, but this is substantial;
- Físchlár is … TV, VCR, browse/play but not yet search;

Structuring video for browsing

- Once structured, how to index for search;
- Usual approach is to present this visually and to choose a keyframe for each shot and use this … middle, first, last, average, our Físchlár system keyframe is …
- Systems **could** de developed to do image retrieval on keyframes but the norm is to present keyframes for **browsing** … c.f. earlier comments on browsing audio;
- A problem here is the sheer number of shots/keyframes … a 30 minute progam can have order hundreds
- Thus some kind of structure should be applied to keyframe sets … in Físchlár we do the following ...




News And Weather (30min)
(0,83)
(84,168)
(253,337)
(338,421)
(253,266)
(281,294)
(309,322)
(323,337)
(253,254)
(262,263)
(264,266)


a video document
A set of keyframes
Keyframe-based video browsers

Físchlár video browsing system

- These two examples are within-program browsing which is complimented by other Físchlár metadata (program name, TV station, date, time, etc.) to help users locate the program first … we also use personalised TV listings and Físchlár has (as of last week) over 400 programmes recorded, analysed and available for browse/play;

Video retrieval - Informedia

- This, however, is video browsing … for video **searching** (apart from image retrieval on keyframes) we have to use text;
- The seminal work in this area comes from CMU in Informedia project who took TV news broadcasts and did …
  - shot bound detection and rep frame selection;
  - speech recognition using Sphinx, one of the best;
  - object detection from frames looking for faces to match against a VIP face database;
  - looking for text in captions and in frame to submit to OCR;
- Informedia then supported text-based **search** through its archive, which was followed by user browsing;

Video Retrieval

- When it comes to video search, this is as good as it gets, for now …
  - text search on captions, teletext, OCR, etc.;
  - text search on the audio track based on speech recognition;
  - object match on easily-identified objects from a limited domain, e.g. faces;
  - heavy emphasis on browsing support;
- Active research is ongoing on **object-based** retrieval, (level 2 in image retrieval) identifying and tracking objects … for now this is a stepping stone towards MPEG-4 encoding but there is now an awareness of content-based operations (us !);
- Could be easier than object recognition in still images ‘cos there is context, dialogue, program metadata, motion ?

Físchlár plans ...

- As an aside … in our work on Físchlár we will do
  - closed caption/teletext/888 capture and use this for search, alert, filter, personalise and summarise;
  - develop SMS, WAP, mobile PDA and 3G phone interfaces;
  - scale up to a large, real userbase … stretch out the browsing work;
  - object recognition in frames and then object tracking/matching for restricted domains;
  - continue deconstructing the encoded stream to reverse engineer things like camera motion, etc.

Nearly finished …. only a couple more slides !

4. Conclusions and Summary

- Browsing is a big part of navigation through audio and video, much more so than for text or even image;
- We cannot take text-based IR and apply it to continuous media, we must re-think the whole user-system interaction and combine search-browse seamlessly;
- As digital TV achieves penetration, the demand for this will soar, much faster than deployment of WWW, perhaps as much as growth of mobile phones;
- People will want video on their 3G phones and their STBs and these will be huge markets which will dominate video IR, just as web searching dominates text-based information retrieval, so demand will outstrip our development;

Commercial systems

- Most of the commercial image retrieval systems have re-invented themselves as video indexing, browsing and retrieval systems using SBD, keyframe extraction and sometimes keyframe match;
- Informedia also has a spin-off;
- Most of these operate on broadcast TV news;

Further sources ...

- trec.nist.gov for info on TREC and spoken document track;
- New track in TREC-10 on video … easy to get hold of 2h
- A list of contemporary video indexing/browse/retrieval projects is maintained at (ref 19) which is a great starting point to familiarise oneself on who is doing what in video;
- Access to Físchlár from lorca.compapp.dcu.ie/Video … email us for password (browse, no streaming)

- Fin -