Information Navigation in Digital Video Archives

Alan F. Smeaton
Centre for Digital Video Processing
Dublin City University

Presented at the 4th European Summer School on Information Retrieval, Aussois, France, September 2003

Organisation of this Talk

- A wide-ranging presentation covering …
  - The nature of (digital) video information, compression formats, standards, MPEG-7, etc.
  - The nature of current approaches to video navigation & automatic structuring of digital video, extraction and identification of video features, video analysis, objects in video;
    - Teletext search and keyframe browsing … the Físchlár Systems
    - Video Search based on feature extraction … the Físchlár-TREC2002 System
    - Video navigation based on objects … Físchlár - Simpsons !
  - TRECVID

1. Introduction to Digital Video Encoding

- Video is 25/30 fps of synchronised images and audio;
- To display a single image of TV-quality video requires 720 Kbytes, so without compression this is 100 GBytes for a 90 minute movie -> video must be compressed !!!
- There are formats such as .AVI, QuickTime, .ram, .rm (Real Networks), .wma, .wmp, but the ones that matter are the MPEG family;
- Before we look at IR and video we should have some understanding of how video is encoded;

Video Encoding principles

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

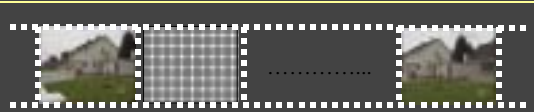
Motion compensation

Camera panning right

Evolution of MPEG standards

- MPEG-1 ('93) … for 1.5 Mbit/s transmission, near VHS-quality. Encoding based on macro-blocks;
- MPEG-2 ('95) … allows data rates up to 100 MBit/s. Used for digital TV, DVDs and broadcasting units;
- MPEG-4 ('01) … suitable for lower data rates between 10 KBit/s and 1 MBit/s. Good quality object-based encoding that will enable object-based interactions for users, but just out of reach for now (see later);
- MPEG-7 has recent adoption, basically, an XML-like content descriptor stream;

MPEG-1 encoding



- 352x288 pixels per frame @ 25 fps giving near-VHS quality at 1.5 Mbps; can be decoded 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 and in B- and P-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 (1 per macroblock) can sometimes be incorrect and have rogues;
- MPEG-2 is the same principle except 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 interaction;

What's great about MPEG ?

- MPEG/video standards are great for …
  Recording →
  Authoring →
  Compression →
  Transmission →
  Playback
- MPEG/video standards do nothing for …
  Searching →
  Browsing →
  Linking →
  Summarising

Technical Challenges

- Many / most technical challenges associated with capture, compression, storage, archival, transmit, rendering of digital video are solved, or nearly so;
- Remaining issues are scale, deployment, business models, killer applications;
- Other challenges of developing object-based analysis and compression are image processing challenges;
- Current applications of DV are production-quality video recording, digital TV and DVDs, consumer (home) video, CCTV, TiVo etc. home platforms;
- Costs are plummeting, with a lifetime of video on HDDs costing little more than $50k

Video Summarisation ?

- Video is linear, typically not marked up with structure, and takes some fraction of linear to view, gist, or summarise;
- Marchionini et al. work on 125x, 250x F/F for gisting;
- Video summarisation is (now) a hot topic … it is achievable e.g. sports summarisation (CIVR had very many papers);
- Summarisation of (sports) video is a low-hanging fruit because it can be low-level signal processing of audio and visual;
- We don't know enough about grammars for movie trailers, or other summaries;

Who needs video searching ?

- With all this video information available, it follows that information management / organisation / retrieval / navigation is required, but who needs it ?
- Journalists, producers, film & TV program makers need to search, the BBC archive has +500k queries plus 1M new items … per year;
- From the BBC …
  - Police car with blue light flashing
  - Government plan to improve reading standards
  - Two shot of Kenneth Clarke and William Hague
  - Bullying at school
  - X ray machines at airports
  - Failing schools
  - UN peace keeping forces in Angola
  - Interior of UN Security Council
  - Actuality of UN General Secretary Kofi Annan
  - Exteriors of commerical banks

Who <u>else</u> needs video searching ?

- This can be done with keyword captions and indices and is laboriously done at c.10x real-time in almost all TV archives;
- However, the development of video navigation is not necessarily about replacing or improving existing applications … its about creating new ones;
- Video content is plentiful … its now available digitally … we can work on it directly … so it follows that digital video navigation is required;

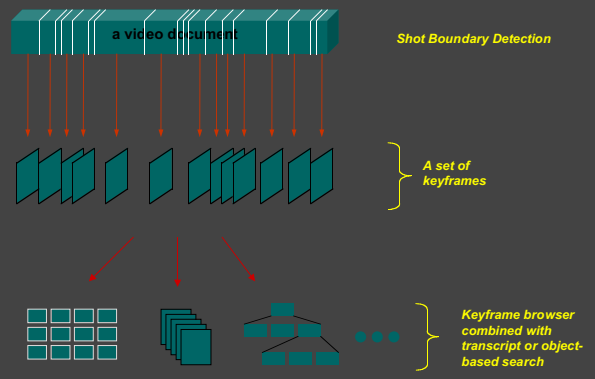
2. How do we do video "navigation" ?

- In operational video IR systems the predominant access is manual tagging as metadata;
- Emerging automatic approaches are based on shot boundary detection or other video structuring, feature extraction and keyframe identification, followed by feature searching with keyframe browsing;
- Up to recently there has been no test collection of video, so it is difficult to compare approaches, but TRECVID (see later) is addressing this;

Automatic structuring of video

- Video "programmes" are structured into logical scenes, and physical shots
- If dealing with text, then text structure is obvious:
  - paragraph, section, topic, page, etc.
  - All text-based indexing, retrieval, linking, etc. builds upon this structure;
- If dealing with video, then first it needs to be structured, automatically;
- Automatic shot boundary detection and selection of representative keyframes is usually the first step;

Typical automatic structuring of video



After video structuring …

I classify video search/browse into 3 types:

- Text search and keyframe browsing
- Feature-based search and browsing
- Object-based search and browsing

- is with us now; (2) is starting to appear and (3) is still a bit away.

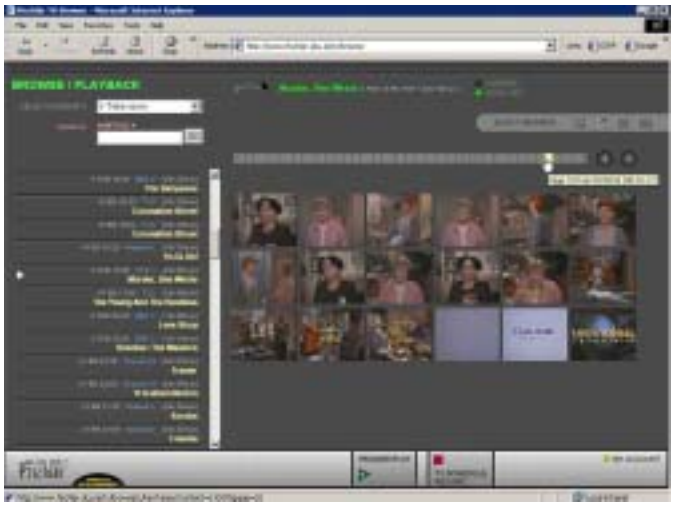
- These can be combined

2.1 Text Search & KF Browse - Físchlár-TV

- Físchlár-TV supports recording, analysis, browsing, and playback of digital video, currently, TV from 8 channels;
- Users select programmes from a TV schedule with programme genre (category) automatically assigned;
- At transmission time, we capture video, detect shots, scenes & keyframes and place videos in a library of content;
- Users browse programme genres or otherwise locate programmes, and select a program for viewing;
- Initially, users are allowed to browse keyframes and then playback;

Físchlár-TV - CURRENT Use

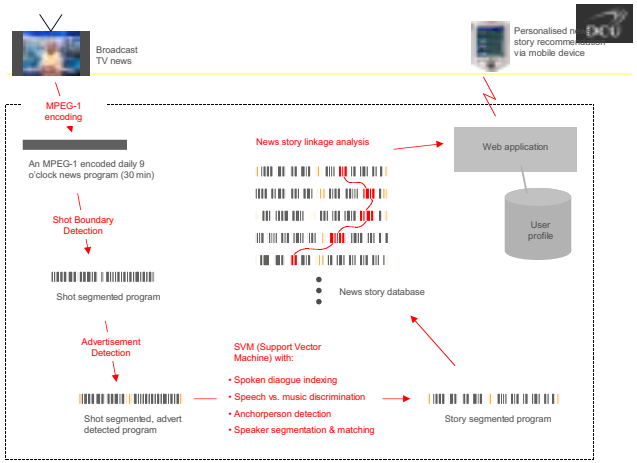
- Físchlár is very popular ...
  - 2,000+ users on campus, 850 "regular"
  - term-time recording 18-30 hours / day
- Used from computing labs, residences, library, on campus;
- Used for entertainment, research *and* for teaching & learning
- We can store 300 hours online at any one time and have 200 simultaneous playbacks, and the unit of retrieval is the entire program;

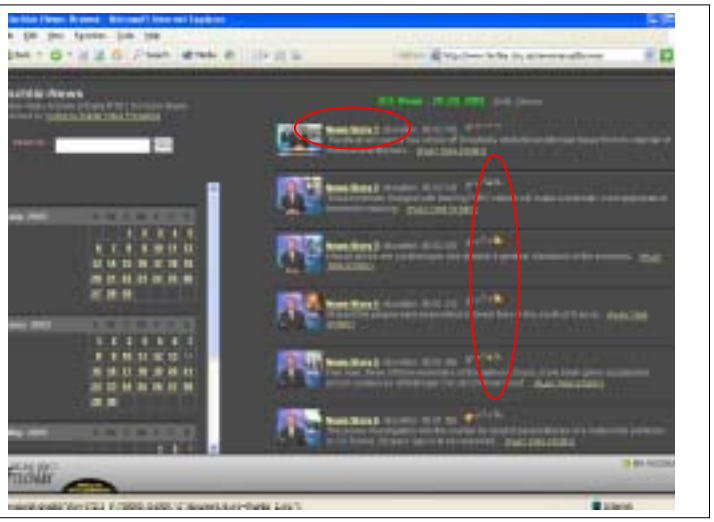


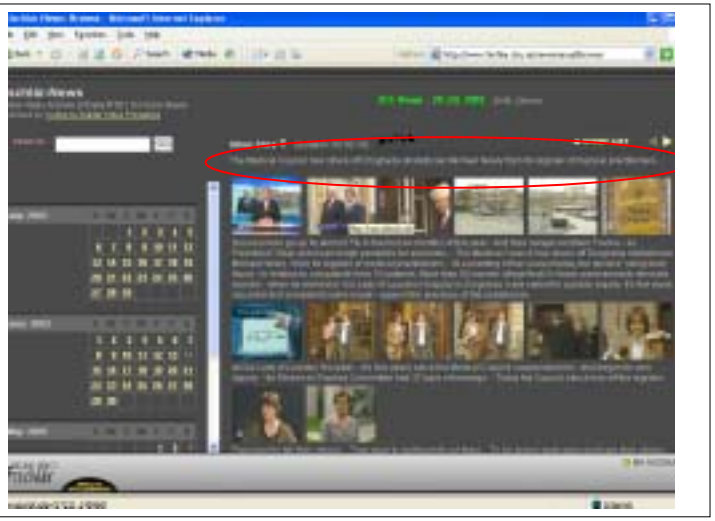
Físchlár–News-Stories

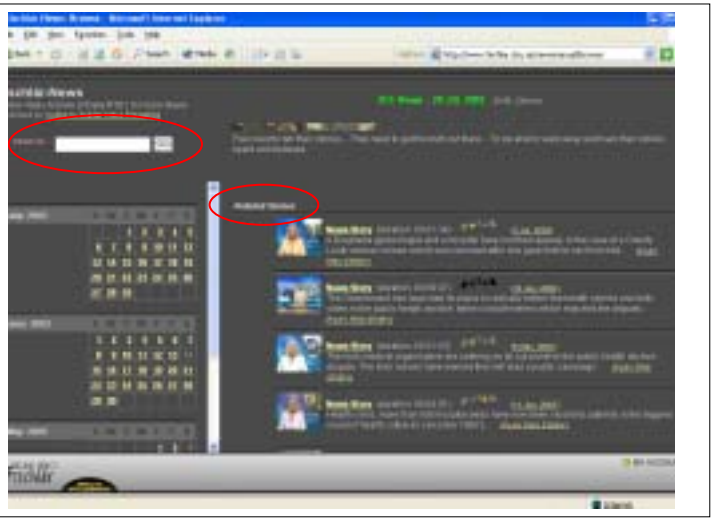
- Físchlár-TV is a large, multi-user, shared TiVo;
- Its unit of retrieval is the *program* so it does not need more than program navigation, and within-program browsing;
- Its local browser(s) are good for finding previously viewed shots;
- Its an eye-catcher, but it is limited video navigation;

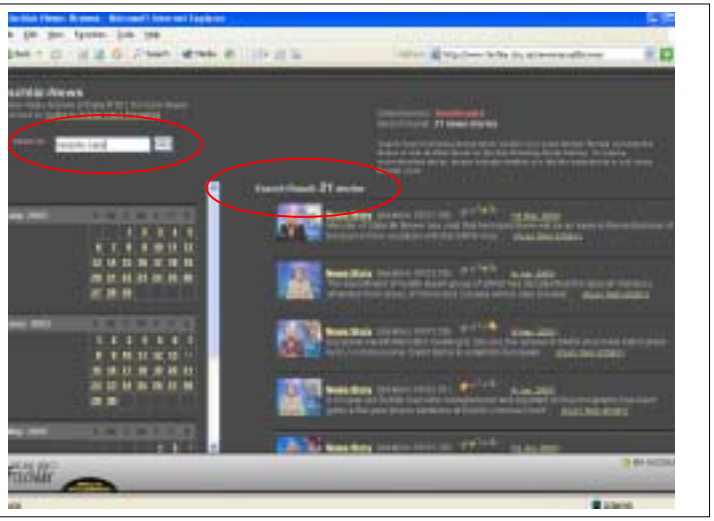
- Físchlár–News-Stories, however, is more sophisticated;

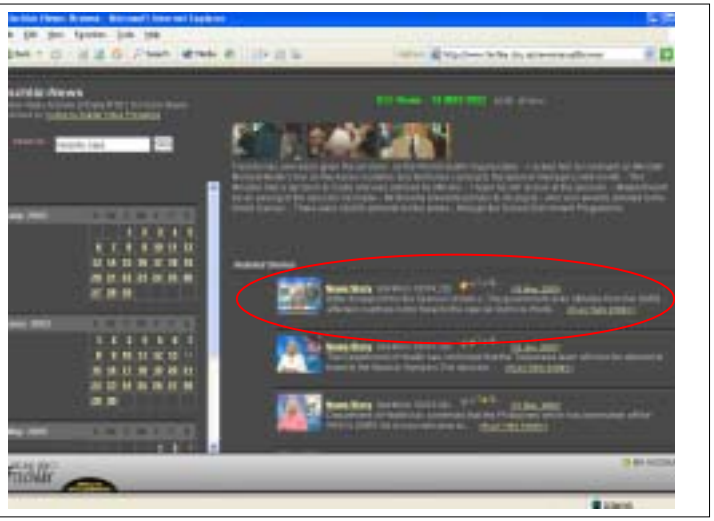






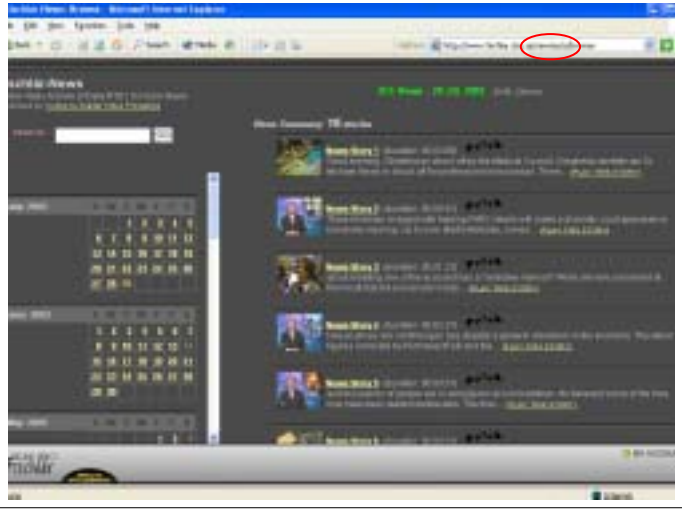






Físchlár–News-Stories

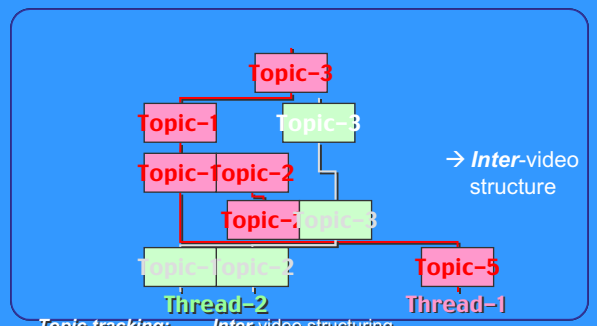
- We have manual story bound segmentation, to kick-start, but we also have automatic story bound segmentation



Físchlár–News-Stories

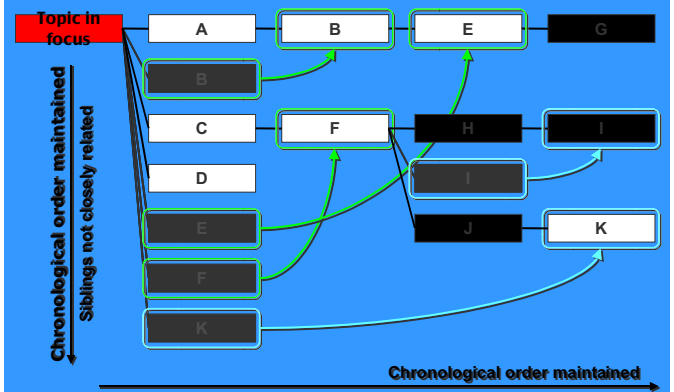
- Físchlár–News-Stories is a very simple, 1-step, no lookahead linking of stories based on CC text;
- Story-story linking currently constrained by CC lagging, but there is scope to include other story features … face recognition, speaker segmentation, more elaborate text match, etc.
- There is no threading, no lookahead,

Inter-Video Structure (Ichiro IDE et al. National Institute of Informatics, Japan, in AAAI2003 Spring Symp.)

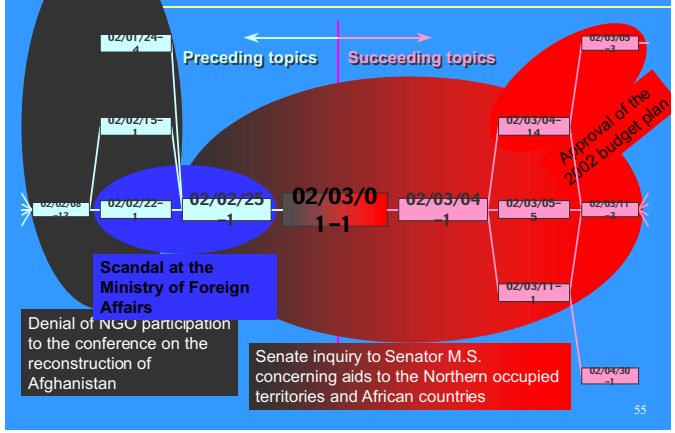


- ***Topic tracking:*** ***Inter***-video structuring
  → Reveals implicit content-based structure across videos

Inter-Video Structuring: Topic Tracking
— An Example of Topic Threading —

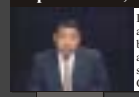


Inter-Video Structuring: Topic Tracking
— An Example of Extracted Topic Threads —



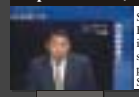
Visualizing the Topic-Based Structure
— An Example of a Tracked Topic Thread (1) —

September 12, 2001; Topic#1



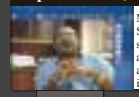
In the United States, hijacked airliners slammed in the World Trade Center in New York, and the Pentagon in Washington on Tuesday. Rescue efforts are on the way at the sight, but the work is not proceeding smoothly. The death toll from the series of terrorist attacks could top several thousands. Ministry of Foreign Affairs has confirmed the safety of some 300 Japanese employees of 36 companies housed in the World Trade Center in New York, but there are still 18 unaccounted for. Last night, before 10 .....

September 12, 2001; Topic#5



Since the incident occurred in New York's financial center, the New York Stock Exchange was closed yesterday, and will continue to be closed today the 12th, too. That is all from New York. OK. That was the latest report from New York. On the other hand, suburban Washington was also attacked. Local fire department estimates up to 800 people may have been killed at the Pentagon. A report from Washington is by Kenji Sobata. Mr. Sobata? Yes. Is the Pentagon still burning? Yes, can you see the .....

September 13, 2001; Topic#3

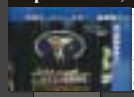


Mr. Degawa, there is a certain man's name in the suspect group whispered in the United States government. His name is Osama Bin Laden. May we suspect that he is behind this series of terrorist attacks? Well, we cannot say anything for sure, yet, but investigators are focusing on Mr. Laden, Islamic fundamentalists, certain Arabs and Middle Easterns, and so on. Osama Bin Laden is a leader of Islamic fundamentalists, and he is said to be in the back of an international network. We interviewed an Egyptia n professional .....

September 13, 2001; Topic#4

Visualizing the Topic-Based Structure
— An Example of a Tracked Topic Thread (2) —

September 13, 2001; Topic#4

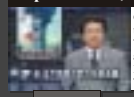


Now, what will be the next target of the investigation? Well, FBI is investigating houses in Florida and Boston, which are suspected to have been used by the hijackers, and is inquiring several people. The target will be how far Mr. Laden's involvement could actually be tracked. On the other hand, the Bush administration is preparing for military retaliation in case the background of the attacks become clear. Secretary of States, Collin Powell stated that diplomatic consensus is becoming formed amo ng .....

September 14, 2001; Topic#1

The United States government says that at least 18 people were involved in the attacks, and it is becoming increasingly convinced that an Islamic Fundamentalist leader, Osama Bin Laden was behind the attacks. The Bush administration has said it is planning to launch comprehensive military retaliation for the attacks against the terrorist organizations responsible and any nation that supports them. I'm looking at those terrorist organizations, who have the kind of capacity that would have been .....

September 15, 2001; Topic#1



Good evening, it is 7 PM, Saturday September 15th. Tonight's program will be extended to 8 o'clock. We have extensive coverage of the terrorist attacks in the United States. The United States Congress has approved the resolution allowing the Bush administration to use force to retaliate against Tuesday's terrorist attacks. President Bush is preparing seriously for the military retaliation to the terrorist organizations. The resolution allows full-measure military attacks to terrorist organizations .. ....

Físchlár and Standards

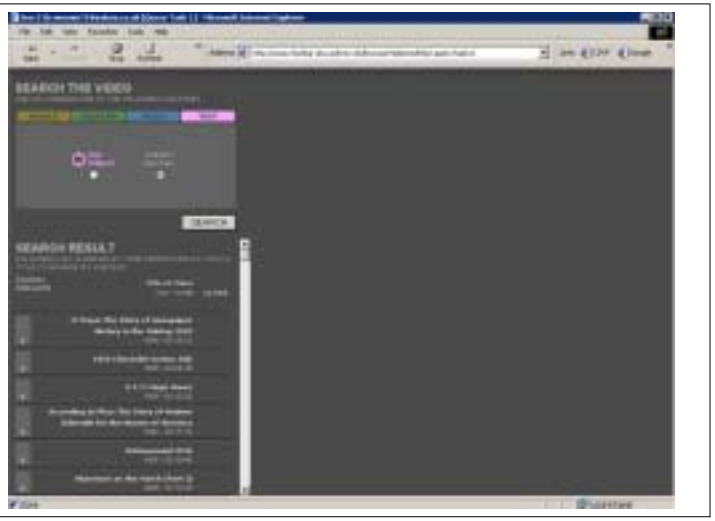
- All Físchlár systems are standards-compliant … feature and shot descriptions generate MPEG-7 and Físchlár can process any MPEG-7 described object;
- Físchlár internally produces user responses as XML documents allowing XSL transformations to browsers;
- This allows us to develop iPAQ and *xda* interfaces with XSL processing to strip out unwanted details on the mobile platform;

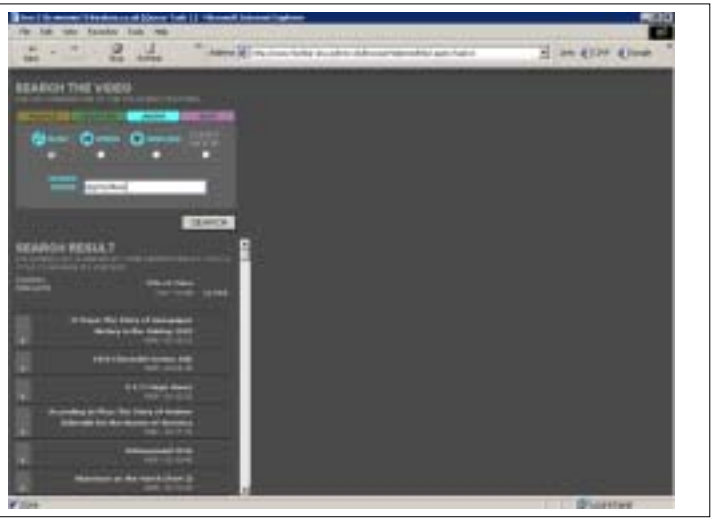
2.2 Feature-Based Search & Físchlár-TREC2002

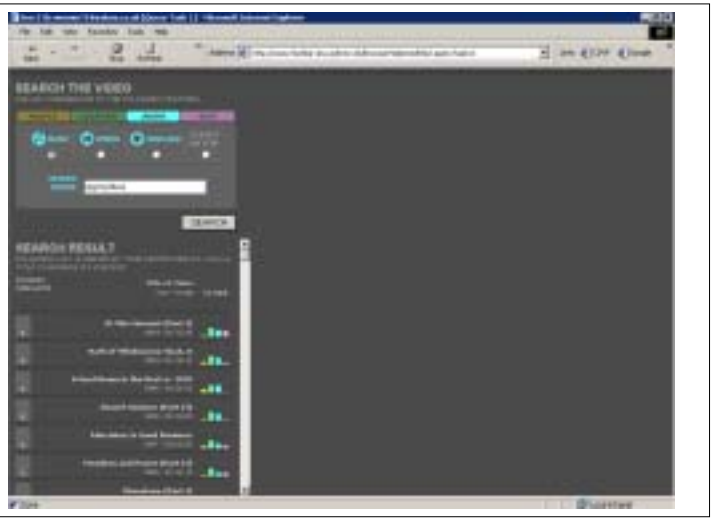
- Feature-based search uses features, derived from video, as search criterion;
- Ideally, features are automatically derived, robust, accurate, and useful for retrieval, but must be computed in advance.
- It would be nice to build a feature detector for each query at query time, but not possible;
- In TRECVID-2002 we used some of our own features, plus imported features and speech transcript donated by other groups (imported in MPEG-7 format);
- Físchlár-TREC2002 supports video shot retrieval based on user-selected features, allowing fine-grained video retrieval of shots;
- What does it look like ?

The 10 Features Chosen

1. Outdoors
2. Indoors
3. Face - 1+ human face with nose, mouth, 2 eyes
4. People - 2+ humans, each at least partially visible
5. Cityscape - city/urban/suburban setting
6. Landscape - natural inland setting with no human development such as ploughing or crops
7. Text Overlay - large enough to be read
8. Speech - human voice uttering words
9. Instrumental Sound - 1+ musical instruments
10. Monologue - 1 person, partially visible, speaking for a long time without interruption

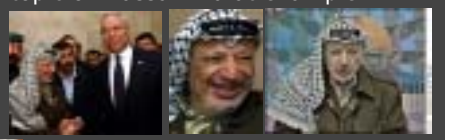






Feature-based video retrieval

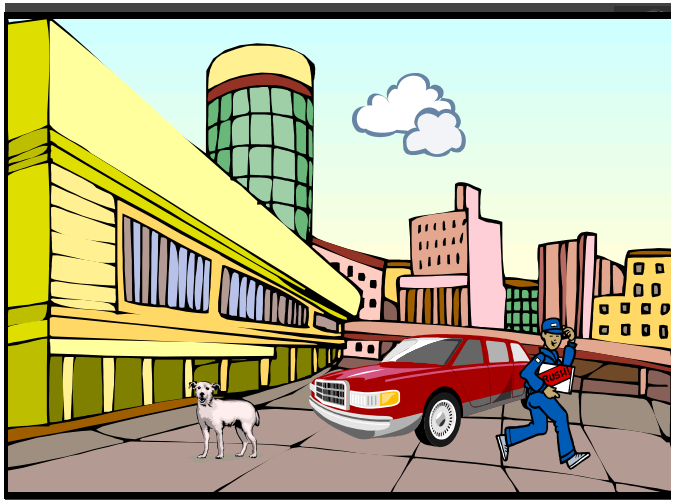
- Good quality and useful feature-based retrieval requires a broad set of features which are useful for the query;
- Narrow, specific features would be great … e.g.TREC2003 topic on Yassar Arafat example images



- So build a  detector !
- We'll see later how TRECVID selected features

2.3 Object-Based Video Retrieval – Físchlár-Simpsons ?

- Dominant approach to video IR is to adapt video navigation around conventional IR retrieval, to rely on text and limited input from feature detection … not much of a future there !
- Object segmentation, shape matching and tracking = the consuming passion of the image processing and video analysis community;
- End-goal is compression (MPEG-4), object tracking, object-based compression



Qimera: Semi-automatic Segmentation

How good is it on natural video ?

Qimera =
- DCU
- QMUL
- ITI
- UPC
- and others

Figure 3. Dialog window for specifying the initial mask. The user can simply scribble (left/right mouse buttons) the foreground and background objects.

Semi-automatic object tracking

Qimera: Automatic segmentation

Original frame, region contours, region mean colour) for PFZLA (TUM).

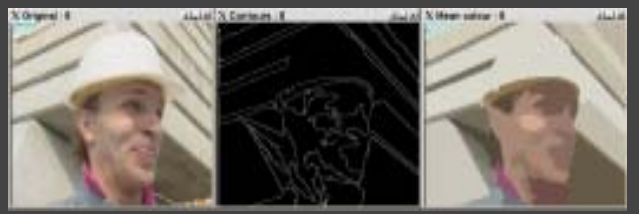
Qimera: Automatic segmentation

Original frame, region contours, region mean colour for RSST (DCU).

Qimera: Automatic segmentation

Original frame, region contours, region mean colour for EM6D (QMUL).

Qimera: Automatic segmentation



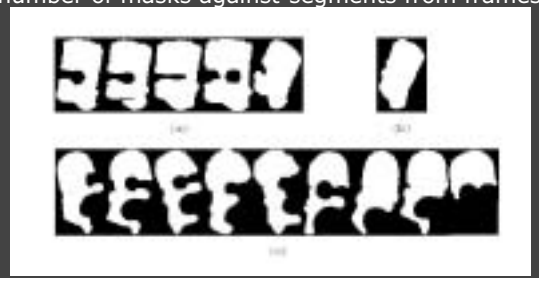
Original frame, region contours, region mean colour for KMCC (ITI).

How good is object segmentation ?

- Object segmentation on natural video is not yet great;
- Object segmentaiton on video has greater potential than on images because of movement and context;
- Object segmentation on synthetic video – animations – can be done;

CDVP Object segmentation

- We have a robust shape matching algorithm, invariant to size, rotation, scale, inversion and mild deformation;
- We use it to find Simpson characters, using a number of masks against segments from frames



Object-based Simpsons

- Using the easy environment of Simpsons, we can detect and track objects over frames, over scenes, do shot-shot similarity, query-shot retrieval, and shot-shot linking using closed caption text, and objects (heads of characters);
- That's a stepping-stone to the kind of retrieval we eventually want to do on natural video;

4. TRECVID – Benchmarking Video Retrieval

- TREC is an annual exercise which has grown over the last 12 years to be the largest, collaborative experiment in information retrieval;
- Some argue that TREC has been the single most influential factor influencing the devlopment of IR over the last decade;
- TREC is global, with nearly 100 participant groups in 2002;
- TREC facilitates comparative evaluation of IR tasks in an open, metrics-based forum;
- TREC started with ad hoc text retrieval and has spun our many "tracks", like SDR, CLIR, non-English IR, web IR, OCR-IR, QA, interactive IR, high-precision IR, novelty detection, etc.

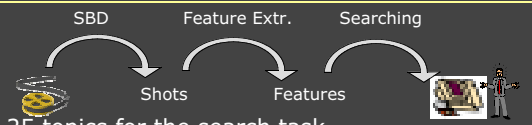
TREC operation

- All TREC tracks have the same organisation;
- An email list agrees the outline, and the details;
- NIST source data and distribute it to registered participants (web download, or ship DVDs/HDDs)
- Participants index/install this locally;
- Sometimes there is data for training available;
- NIST formulate (25, 50) search topics and distribute to participants with a deadline to return top-ranked items;
- NIST pool submissions and manually evaluate, creating the ground truth;
- Standard IR evaluation measures submissions against the ground truths;
- November workshop gathering compares results;

TRECVID – 3 year progression

- TRECVID introduced in 2001, 2002, 2003
- Participants:
  - 12, then 17, now c.35 participating teams;
- Video Data:
  - 11, then 73, now 120 hours;
- Tasks
  - Shot boundary determination
  - Semantic feature extraction
    - features defined jointly by the participants & task is to identify shots with those features
    - 10 features in 2002, 17 in 2003
  - News story segmentation
    - introduced in 2003
  - Searching for shots

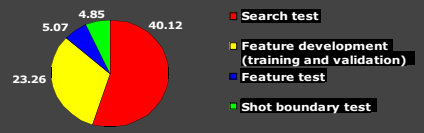
TRECVID2002 Search



- 25 topics for the search task,
  - developed by NIST
  - 4 weeks between release and submission
  - Each topic definition includes text, video, image and/or audio
- TRECVID uses a common set of shot definitions
  - donated by CLIPS-IMAG
  - Provides the common units of retrieval for feature and search tasks – not perfect but acceptable
  - allows pooling for assessment

Video Data in 2002

- Difficult to get video data for use in TREC because ©
- Used mainly Internet Archive
  - advertising, educational, industrial, amateur films 1930-1970
  - produced by corporations, non-profit organisations, trade groups, etc.
  - Noisy, strange color, but real archive data
  - 73.3 hours partitioned as follows:



The 17 groups and the tasks they completed

	Shot Bound	Feature 1	2	3	4	5	6	7	8	9	10	Search Int.	Man.
Carnegie Mellon U. (US)		X	X	X	X	X		X	X	X	X	X	X
CLIPS-IMAG (Fr)	X			X	X				X		X		X
CWI Amsterdam (NL)													X
Dublin City University (Irl)				X					X	X		X	
Fudan Univ. (China)	X	X	X	X	X	X	X	X	X	X	X		X
IBM Research (US)	X	X	X	X	X	X	X	X	X	X	X	X	X
Imperial College London (UK)	X											X	X
Indiana University (US)												X	
Institut Eurecom (Fr)		X	X	X	X	X	X	X					
Mediamill/U Amsterdam (NL)		X	X	X	X	X	X	X	X	X	X		
Microsoft Research Asia (China)	X	X	X	X	X	X	X	X	X	X	X	X	X
National Univ. Singapore (Sing.)	X												
Prous Science (Esp)													X
RMIT University (Aus)	X												
Univ. Bremen (D)	X	X	X										
U. Maryland/INSA/U. Oulu (US)								X				X	X
Univ. Oulu/VTT (Fin)					X	X	X		X	X		X	X

4.1 Feature Extraction

- FE is
  - interesting itself but when it serves to help video navigation and search then its importance increases
- Objective was to
  - begin work on benchmarking FE
  - allow exchange of feature detection output among participants
- Task is as follows:
  - given small standard dataset (5.02 hours, 1,848 shots) with common shot bounds,
  - locate up to 1,000 shots for each of 10 binary features
  - Feature frequency varied from "rare" to "everywhere"

The Features in 2002

1. Outdoors
2. Indoors
3. Face - 1+ human face with nose, mouth, 2 eyes
4. People - 2+ humans, each at least partially visible
5. Cityscape - city/urban/suburban setting
6. Landscape - natural inland setting with no human development such as ploughing or crops
7. Text Overlay - large enough to be read
8. Speech - human voice uttering words
9. Instrumental Sound - 1+ musical instruments
10. Monologue - 1 person, partially visible, speaking for a long time without interruption

Approaches taken

Using {colour histograms, colour distributions, edges, low-level audio features, face masks}

from a hand-labelled training dataset

as inputs to a {SVM, Nnet}

to develop a classifier !

- That's too short and an unfair summary
- Absolute results are not impressive, esp. since these were used in retrieval
- Groups improved upon these post-TREC

Lessons from feature extraction

- Much has been learned from 2002, not least that hand-labelling for training is tedious, and doesn't support direct comparisons across sites;
- This has been addressed in TRECVID2003 (see later);

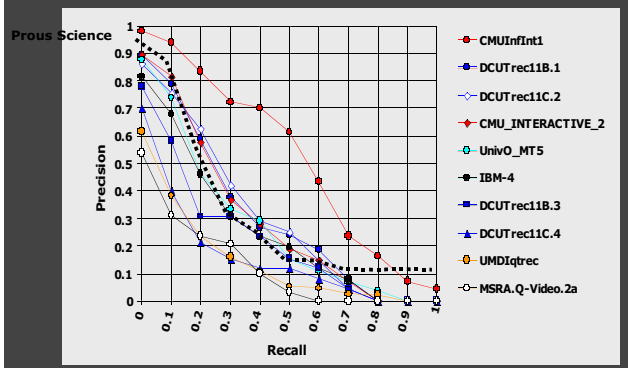
4.2 The Search Task in 2002

- Task is similar to text analogue …
  - topics are formatted descriptions of an information need
  - task is to return up to 100 shots that meet the need
- Test data: 40.12 hours (14,524 common shots)
- Features and/or ASR donated by CLIPS, DCU, IBM, Mediamill and MSRA
- NIST assessors
  - judged top shots from each submitted result set
- Used trec_eval to calculate measures

Search Topics

- Topics (25) multimedia, created by NIST
- 22 had video examples (avg 2.7 each), 8 had image (avg 1.9 each)
- Requested shots with specific/generic:
  - People: George Washington; football players
  - Things: Golden Gate Bridge; sailboats
  - Locations: ---; overhead views of cities
  - Activities : ---; rocket taking off
  - Combinations of the above:
    - People spending leisure time at the beach
    - Locomotive approaching the viewer
    - Microscopic views of living cells

Interactive runs top 10 (of 13)



4.3 TRECVID2003

- Data is 120 hours of ABC & CNN news + 13 hours of CSPAN from 1998, with associated ASR (thanks LIMSI), and closed captions;
- Common shot boundaries and shipped on HDDs;
- Split 50:50 into training and test data;
- 4 tasks:
  - Shot boundary detection is as before;
  - Story bound segmentation is new, derived from news video, match a manual ground truth;
  - Feature extraction is as before, but new features;
  - Search is as before, but find clips, not news;

TRECVID2003 Features

1. Outdoors: segment contains a recognizably outdoor location, i.e., one outside of buildings. Should exclude all scenes that are indoors or are close-ups of objects (even if the objects are outdoors).
2. News subject face: segment contains the face of at least one human news subject. The face must be of someone who is not an anchor person, news reporter, correspondent, commentator, news analyst, nor other sort of news person.
3. People: segment contains at least THREE humans.
4. Building: segment contains a building. Buildings are walled structures with a roof.
5. Road: segment contains part of a road - any size, paved or not.
6. Vegetation: segment contains living vegetation in its natural environment
7. Animal: segment contains an animal other than a human
8. Female speech: segment contains a female human voice uttering words during and the speaker is visible.

TRECVID2003 Features

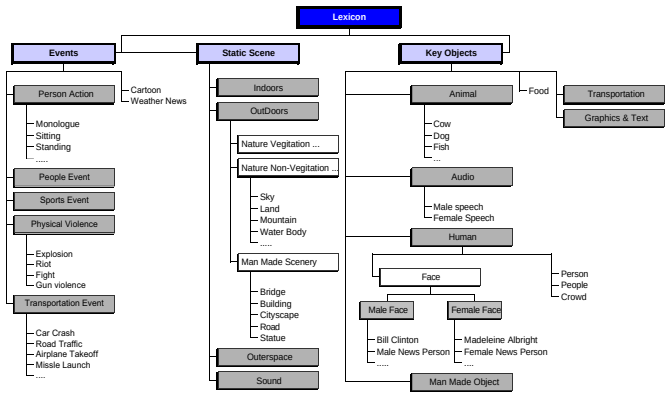
9. Car/truck/bus: segment contains at least one automobile, truck, or bus exterior.
10. Aircraft: segment contains at least one aircraft of any sort.
11. News subject monologue: segment contains an event in which a single person, a news subject not a news person, speaks for a long time without interruption by another speaker. Pauses are ok if short.
12. Non-studio setting: segment is not set in a TV broadcast studio
13. Sporting event: segment contains video of one or more organized sporting events
14. Weather news: segment reports on the weather
15. Zoom in: camera zooms in during the segment
16. Physical violence: segment contains violent interaction between people and/or objects
17. Person x: segment contains video of person x (x = Madeleine Albright)

TRECVID2003 Features

- In 2002 some groups made their feature extraction results available to others in MPEG-7 and the same is expected/hoped in 2003;
- For feature extraction, rather than have sites independently hand-label training data there is an annotation forum, steered by IBM;
- 21 sites, 100+ annotators, each manually annotating to "*develop a large video dataset with semantic labels by manually annotating Event descriptions, Static Scene descriptions, and Key Object descriptions associated to the shots and regions of these digital videos as test-bed for the entire research community.*"

TRECVID2003 Features

- Collaborative annotation tool, developed by IBM, called VideoAnnEx v2.1, used to annotate 63 hours;
- A pre-defined hierarchical lexicon of 133 labels sub-divided into Event , Static Scene and Key objects corresponding to the 17 features;

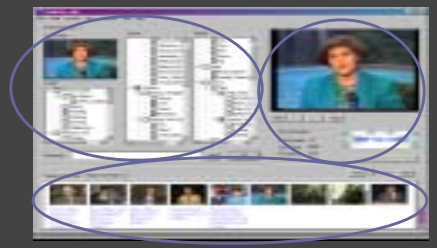


The VideoAnnEX Tool

- Below we see the VideoAnnEx used in the annotation process.
- We will follow the process of annotating one shot from a video sequence over the next few slides.

The VideoAnnEx Interface

- The VideoAnnEx tool is divided up into four different regions (three below):
  - Shot annotation & keyframe
  - The view panel
  - Video playback & Shot details



The View Panel

The View Panel contains two tabs

'Shots in the Video'

- **Displays the keyframes of all shots in the video**
- **Below each shot's keyframe is a list of annotation descriptions, if provided**

'Frames in the Shot'

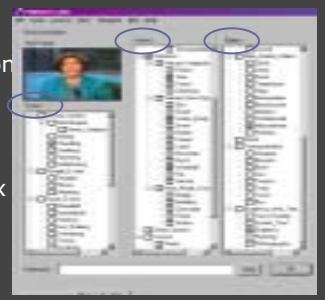
- **Displays all the I-frames of the currently selected shot**
- **The keyframe for the shot can be manually selected by double clicking on a particular I-frame**

Video PlayBack & Shot Details

- This region plays the shot that is currently being annotated (a)
- The details of the current shot are displayed below the playback (b)

Shot Annotation & Keyframe

- The three lexicons, combined, contain 133 labels, subdivided into three groups:
  - event
  - scene
  - key object description
- Each label has a corresponding check box for the user to select if a relevant feature is in the shot.



Region Annotation

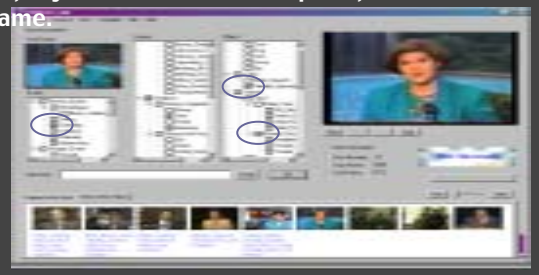
- Once the user has finished checking the labels appropriate to the current shot they must then allocate a region to each of the annotations.
- This is achieved by dragging a box over the region in the shot associated with any given label.

Annotating a Shot

- The user chooses a shot to be annotated by clicking on the relevant frame in the "Shot in the Video" tab (a). The keyframe for that shot is then displayed in the top left window (b).

Annotating a Shot (cont.)

The user annotates the chosen shot by checking the boxes corresponding to the events, objects and scene description, evident on the keyframe.



Assigning Shot labels

- For example this shot has been annotated with the following labels:

- Standing
- Outdoors
- Trees
- Greenery
- Water body
- Waterfall
- Microphone
- Female Speech
- Female Face

Selecting the Region

It is necessary to select a region on the keyframe representing the shot, for each label chosen, prior to the completion of the annotation process for any shot.

On Completion of a Shot Annotation

Once the Regional annotation has been completed, the entire set of annotations given for that shot are displayed in the view panel underneath the keyframe in question;

The user is automatically presented with the next shot in the video sequence to annotate.

TRECVID2003

- The annotation process was finished early July and released to participants on July 14;
- The annotation will be made available to the research community after TRECVID2003;
- As we speak it is being used by some groups to train their feature detection classifiers;
- As we speak, NIST are preparing (25) topics for distribution in August;
- As we speak, sites are running their shot bound detection, and story bound detection, for submission shortly;

4.4 Conclusions on TRECVID (1)

- TRECVID has grown significantly
  … data, groups, tasks, measures, complexity
- It will continue into 2004 and maybe beyond;
- In the usual TREC philosophy, data (video, annotations, topics, assessments, submissions) will be made available, subject to licence and media;
- TRECVID is now a separate workshop to TREC at NIST;
- The search task is becoming increasingly interactive, as we'll see this year
- Evaluation framework has settled down – should be repeated on new data with only minor adjustments … 2003 has been data-traumatic !

Conclusions on TRECVID (2)

- Donated features enable many sites to take part and greatly enrich the progress .. this cannot be overstated … TRECVID is very collegiate and beneficial all-round … very unlike other TRECs
- But, there are issues … we don't really have an established benchmark yet, just a better understanding of some of the issues related to building one, and they are complex, but we've got data;
- There are also issues related to the construction of the topics, namely, what is it we are searching for … video clips, or news ? Do people have sample videos/images when they search ? How to capture and evaluate the interactive experience ?

5. Overall Conclusions

- Standards and technologies are now fixed but little work to date on content access to video archives;
- Video navigation is search, local browse and collection-wide link traversal, mostly built around old text search technology;
- Local browsing is OK, but could be much better through better shot summaries;
- Video summarisation at program level is a current hot topic and there is work in sports summarisation, object detection and tracking, object-based hyperlinking across videos and analysis and browsing of consumer (home) movies;

Overall Conclusions

- Search in video is currently text but could be much more, multiple facet & feature combinations;
- Collection-wide link traversal is text, but could be much more in both similarity computation (see search) and also dynamic, personalised, user-driven, contextual and transmedia;
- We're still far short of the hundreds of thousands of hours in TV archives, and our retrieval quality is far short of text-based IR, but the problems are different, and we're getting better;