Video Retrieval

Joemon Jose
Multimedia Information Retrieval Group
Department of Computing Science

CBVR = CBIR + CBAR??

- Many believe CBVR is just an extension of CBIR+CBAR
- In reality, Video have many additional factors, but of course, based on temporal data
- temporal data induces motion to objects
- if CBIR is retrieve images with object 'A', CBVR is retrieve videos with 'A' with certain behavior.
- Videos have a structured organisation, more hierarchical in nature

Multimedia Genres

- Television programs
  - News, sports, documentary, talk show, …
- Movies
  - Drama, comedy, mystery, …
- Meeting records
  - Conference, video teleconference, working group
- Others
  - Surveillance cameras, personal camcorders, …

A Real life situation!

- Companies have miles of video tape,
  - but if you don't know what's on it or how to find what you need, it is almost worthless
- Video retrieval tools aim to sort this problem out!
- Minutes after Princess Dianna's car accident, news organisations were scrambling to air timely retrospectives on the royal icon.
- Materials wasn't the problem but the amount!
- Production assistants spent hours combing through video tapes searching for the right segments

Corporate Video collections!

- Conference/board/staff meeting rooms are equipped with video cameras
- Formal meetings, presentations are video-taped, MPEG-encoded and available via company intranet – average 3hours/week
- Often difficult to find video file and a portion that is of interest
- ***Video is used more and more as a permanent record***
  - ***Hence important to find relevant passages***

In general..

- Data size of video collections has no comparison to text databases
- Retrieval from tera byte & peta byte collections is a greater challenge
  - Techniques from various disciplines
    - Database, information retrieval, computer vision, human computer interaction, …
- Storage of video in compressed formats calls for algorithms and techniques that deal with compressed domain

Video Structures

- Image structure
  - Absolute positioning, relative positioning
- Object motion
  - Translation, rotation
- Camera motion
  - Pan, zoom, perspective change

Video Structures

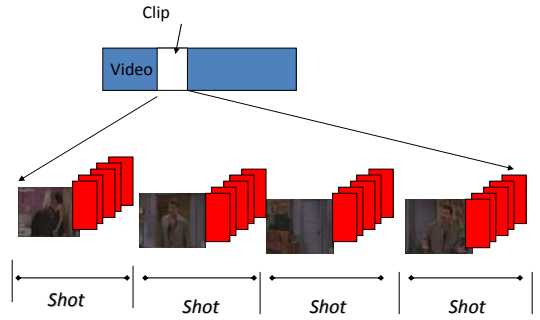
- **Frame level:** There is no (or little) temporal analysis at this level.
- **Shot-level:** A *shot* is a set of contiguous frames all acquired through a continuous camera recording.
- **Scene-level:** A *scene* is a set of contiguous shots having a common semantic significance.
- **Video-level:** The complete video object is treated as a whole.
- Clips
  - A set of frames with some meaning. Clip boundaries do not necessarily coincide with shot boundaries. A clip may correspond to several consecutive shots

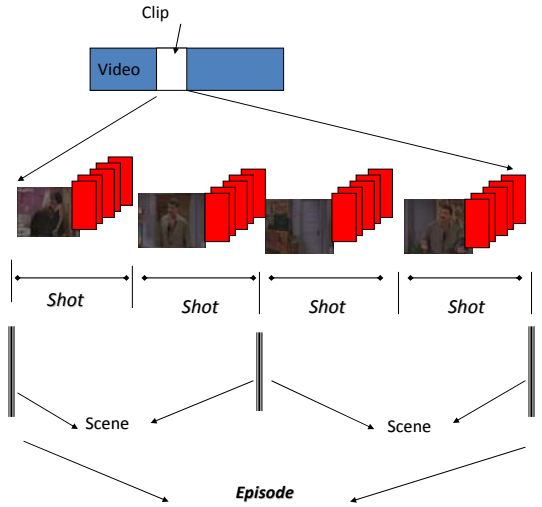
Video Structures

- *Cut*: A sharp boundary between shots. This generally implies a peak in the difference between colour or motion histograms corresponding to the two frames surrounding the cut.
- *Dissolve*: The content of last images of the first shots is continuously mixed with that of the first images of the second shot.
- *Fade-in and fade-out* effects are special cases of dissolve transitions where the first or the second scene, respectively is a dark frame.
- *Wipe*: The images of the second shot continuously cover or push out of the display (coming from a given direction) that of the first shot.

Semantic Level

- Episodes
  - A set of shots that are characterised by a specific sequence of shot types.
    - For example, a news episode can be composed of the anchorperson's introduction shot, the news shot, the reporter's shot, and so on.
- Scenes
  - A collection of consecutive shots that share the three properties of similarity in space, time and action. Scenes are related to stories and can be dynamic or static, depending on whether characters move or not.
    - For example, in movies, conversation scenes are static scenes.





Video Segmentation

- Breaking down video into its constituent elements, the shots, and their higher-level aggregates - scenes, episodes
- Traditional Approach
  - preview the whole video
  - manually identify the segments and boundaries
  - annotate them with texts
  - for 1 hour video it may take 10 hours.

Video Segmentation...

- Alternate way
  - Edit decision lists created by video producers during post-production
    - This list says details of scene change, shot change etc.
  - Problem?
    - Misalignment with video stream
    - A large part of existing video do not contain this
    - Applicable only to particular type of videos!
- Automatic methods
  - use image analysis techniques for the detection of shot boundaries
  - an active research area

Sharp Transition detection

- Cuts
  - The cut is defined as a sharp transition between a shot and the one following
  - It is obtained by simply joining two different shots without the insertion of any photographic effect
  - Cuts generally corresponds to an abrupt change in the brightness pattern for two consecutive images

Cut detection

- Cuts can be detected by sharp change of brightness
  - since two consecutive frames in a shot do not change significantly in their background and object content, their overall brightness distribution differs little

Two techniques...

Pairwise Pixel comparison differences between corresponding pixels in two consecutive frames

$$D_{cut} = \sum \; | I_{xy}(f_t) - I_{xy}(f_{t+1}) |$$

How do you extend it to colour frames?

Histogram Comparison Method

$$D_{cut} = \sum_{j=1}^{n} \; | H(f_t, j) - H(f_{t+1}, j) |$$

Cut detection-Problems

- In the presence of
  - continuos object motion,
  - camera movements
  - changes of illumination
  - the above approaches fail, because

it is difficult to understand
when the brightness changes

Gradual Transition Detection

- Fades and dissolves
  - obtained in the laboratory through optical process
  - They make the boundary between frames spread across a number of frames
- Fading
  - progressive darkening of a shot until the last frame becomes completely black (fade-out), or the opposite (fade-in) (black to white)

Dissolve
is a superimposition of fade-out and fade-in

Shot-to-Shot Structure Detection

- Create a color histogram for each image
- Segment at discontinuities (cuts)
  - Cuts are easy, other transitions are also detectable
- Cluster representative histograms for each shot
  - Identifies cuts back to a prior shot
- Build a time-labeled transition graph

Story Segmentation

- Video often lacks easily detected boundaries
  - Between programs, news stories, etc.
- Accurate segmentation improves utility
  - Too large hurts effectiveness, to small is unnatural
- Multiple segmentation cues are available
  - Genre shift in shot-to-shot structure
  - Vocabulary shift in closed captions
  - Intrusive on-screen text
  - Musical segues

Shot Classification

- Shot-to-shot structure correlates with genre
  - Reflects accepted editorial conventions
- Some substructures are informative
  - Frequent cuts to and from announcers
  - Periodic cuts between talk show participants
  - Wide-narrow cuts in sports programming
- Simple image features can reinforce this
  - Head-and-shoulders, object size, …

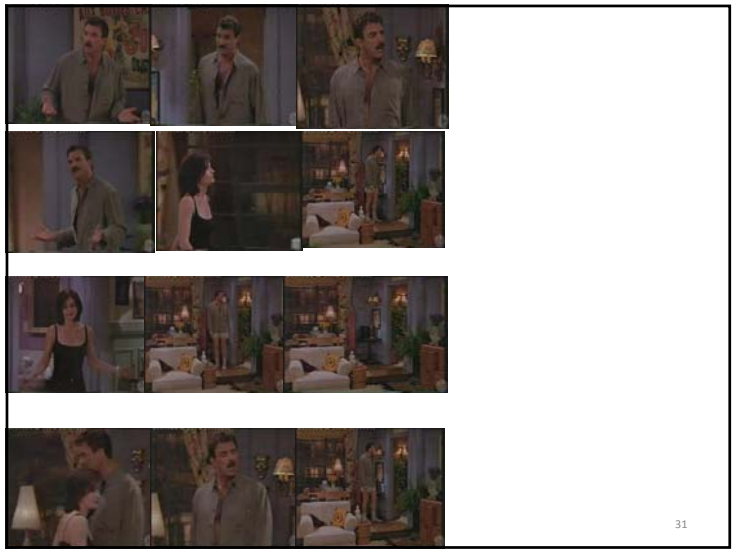
Key Frames?

- When searching collections of videos users often interested an overview of the document
- Key-frames can be used to distinguish videos from each other, to summarise videos and to provide access points into them
- To distinguish shots/scenes from each other
- To summarise shots
- To provide access points to them
- Well chosen key frames
  - Help video selection
  - More visually appealing

Key Frame Extraction

- First frame of a shot is easy to select
  - But it may not be the best choice
- Genre-specific cues may be helpful
  - Minimum optical flow for director's emphasis
  - Face detection for interviews
  - Presence of on-screen captions
- This may produce too many frames
  - Color histogram clusters can reveal duplicates

Key frame selection – A comparison of methods



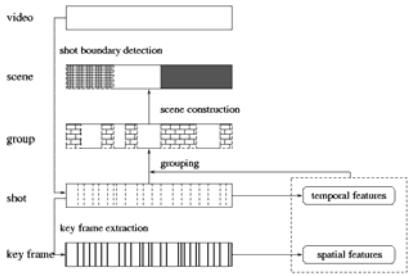
Video retrieval

- At the Frame level
  - Lighting conditions/Composition features
  - Perceptual properties like colour and texture
  - Object identification and location
- At the shot level
  - Select a key frame from a shot
    - Browse & navigate

Video Browsing!

Components of a Video retrieval System!

- A subsystem for parsing video into individual shots
  - Detect shot boundaries automatically.
  - Select a representative frame from the set.
- A subsystem for the extraction of perceptual features of visual data from frames.
  - Image Processing & pattern analysis techniques are used in order to detect salient visual clues



Efficient access of digital video

- Finding desired materials in the database
  - Selection of information –
    - indexing and retrieval techniques?
    - Data organisation & storage
    - Volume of data
- Assimilating and using the found materials
  - Process of information gathering, browsing, understanding, and extraction of meaning from retrieved set
    - Video composition
- Delivering the materials
  - Delivery across networks, hence bandwidth constraints
  - Relation to query reformulation

Video

- Video conveys informative messages through multiple planes of communication.
  - These include the way in which frames are linked together by using editing effects
    - Cuts, fades, dissolves, mattes and so on
  - And what is in the frames
    - The characters, the story content, the story message
- Changes in colour, texture, shape and motion (of both camera and characters/objects) observed in multiple frames, are more important than information embedded in single frames
  - Techniques employed to obtain video (shot angles, camera movements) are also very important and contribute to the information content of the video stream.
  - Each type of video has its own peculiar characteristics
  - These are reflected in the way in which video units are extracted, organised in knowledge structures, indexed and accessed by users

Video Abstractions

- Summaries of original data intend to communicate to the user the salient content of a source video within a short time
- Summaries should be short as possible and present the gist of a video clip in a simple way without losing any valuable piece of information.
- Summaries should present all the key points of the original data in a natural and fluid manner
- Effective abstraction tools allow users to browse and filter segments of video sequences.
- Aid users in finding pieces of information that is closer to their need without forcing them to view the entire video

Exploiting Multiple Modalities

- Video rarely appears in isolation
  - Sound track, closed captions, on-screen captions
- This provides synergy, not just redundancy
  - Some information appears in only one modality
- Image analysis complements video analysis
  - Face detection, video OCR