Multimedia Information Systems

Samson Cheung

EE 639, Fall 2004
Lecture 11: Video boundary detection
Video characteristics

- Video is a rich information source
  - links between frames (cuts, fades, dissolves)
  - changes in color, shapes, motion of both camera and objects
  - video acquisition data (shot angles, camera movements)
  - each type of video has its own characteristics (commercials, news, movies, sport)
Video Structure

- **Frame:** typically 1/25 or 1/30 seconds
- **Shot:** sequence of similar frames
  - An unbroken sequence of frames from one camera
- **Scene:** A collection of consecutive shots focusing on
  - One or more objects of interest
  - One location
  - One thematic concept or event
- **Episode/Story:** consecutive scenes
  - intro, news, reporter, weather
Temporal segmentation

Identify the boundaries between shots and between scenes
Why shot/scene segmentation?

- Video Browsing and visualization
- Archival and indexing
- Non-linear editing
- Event detection
- Film Understanding
Production Models

- **Cut**
  - Abrupt shot change that typically occurs in one frame
- **Fade**
  - In or out between bright/dark and new shot
- **Dissolve**
  - Gradual transition between two shots
- **Wipe** and others
- **Motion**
  - Camera motion: pan and zoom
  - Object motion: not shot change
- Flashlight and aperture change
Program types

- News: lighting
- Sports: many cut changes
- Movies: motion, similar locations
- Television programs, e.g., music, documentary, sitcom etc.
- Commercials: short, fast changing
- Consumer videos: pan, zoom, noise, flash
- Meetings and presentations
Basic Detection Methods

- video
  - time scale: 33-66 ms

- audio
  - time scale: 100ms – few sec

**Representation** → **Feature** → **Detection**

- **Video**:
  - pixels
  - color
  - edge
  - motion
  - compression data
  - audio

- **Audio**:
  - pixels
  - histogram
  - DCT
  - motion vectors
  - audio features
  - audio types

- **Detection**:
  - difference thresholding
  - statistical
  - model based
Factors

- Accuracy (detection and timing)
- Speed
- Simplicity
- Generality

Typical problems:
- Camera movement
- Flash
Pair-wise pixel comparison

- Counts number of pixels changed from a frame to the next
- A shot boundary is found if more than $T_b$ pixels has changed
- Problem: sensitivity to camera/object motion and noise
  - many pixels change
- Trivial extension to block-based comparison

\[
DP_i = \begin{cases} 
1 & \text{if } |P_i[k, l] - P_{i+1}[k, l]| > t \\
0 & \text{otherwise}
\end{cases}
\]

\[
\frac{\sum_{k,l=1}^{M,N} DP_i[k, l]}{MN} \times 100 > T_b
\]
Histogram comparison

- Grey level histograms for three successive frames
- Frames 1 and 2 almost identical
- Camera break between frames 2 and 3
- Many techniques for computing differences between histograms.
- More robust against camera movement
An example

Histogram difference between successive frames in a documentary
Twin comparison approach

- Two thresholds
  - $T_b$ for camera break detection
  - $T_s < T_b$ for special effects like dissolves, motion
- Compare consecutive frames (e.g. histograms)
  - if difference exceeds $T_b$: cut
  - if difference exceeds $T_s$: potential graduate transition
    - start accumulating differences from that frame until the transition becomes lower than $T_s$
    - Boundary if accumulated difference becomes higher than $T_b$
Using twin comparison

camera break special effect

shot boundary
Edges before and after SC

Edges in frame 1

Edges in their sum

Edges in frame 2
Edge-based fade detection

- **Definitions**
  - $p_{in}$: number of entering edge pixels
  - $p_{out}$: number of exiting edge pixels

- **Observations**
  - **fade-in**: $p_{in}$ much larger than $p_{out}$
  - **fade-out**: $p_{out}$ much larger than $p_{in}$
  - **Dissolve**: a predominance of $p_{in}$ followed by a predominance of $p_{out}$ (assuming the new image is faded-in before the old one is faded-out)

- **Statistics**: Edge-change Ratio

- **Lienhart paper comment**: not worth it!
Compressed-domain processing

- **DCT coefficients**
  - Fastest but coarse: DC term from I-frames
  - Fast but better: I-frames (DCT coefficient correlation)
  - Fast but best: DC images (need MV)
    - See Song & Yiu 99.
    - \[ \text{DC}(\text{P}_{\text{ref}}) \approx \sum_{i=0}^{3} \frac{h_i w_i}{64} \text{DC}(\text{P}_i) \]

- **See Koprinska, Carrato, 2001.**
  - Macroblock Coding Mode – many intra-blocks indicates a scene change
  - Motion Vectors – camera movement
  - Bit usage – scene change uses a lot of bits
Selection of threshold

- Most methods use a fixed global threshold.

- **Adaptive threshold**
  - **Sliding-window based method** [Yeo, Liu 95]
    - Look at a sliding window of histogram differences
    - Hard-cut if the maximum absolute difference is significantly larger than the second max absolute difference
  - **Shot-duration prior** [Vasconcelos, Lippman 00]
    - Likelihood testing
    - Threshold based on shot-duration modeling

- **Model-based : no threshold**
Shot-duration Prior

- Scene change:
  - SC = scene change, M = measurement (ex. Histogram difference, Δ = duration since last SC), SCD = scene change duration

\[
P(\text{SC} \mid M, \text{no SC for } \Delta) > P(\text{no SC} \mid M, \text{no SC for } \Delta)
\]

\[
\Rightarrow \quad \frac{P(M \mid \text{SC, no SC for } \Delta)}{P(M \mid \text{no SC, no SC for } \Delta)} > \frac{P(\text{no SC} \mid \text{no SC for } \Delta)}{P(\text{SC} \mid \text{no SC for } \Delta)}
\]

\[
\Rightarrow \quad \frac{P(M \mid \text{SC, no SC for } \Delta)}{P(M \mid \text{no SC, no SC for } \Delta)} > \frac{P(\text{no SC} \mid \text{no SC for } \Delta)}{P(\text{SC} \mid \text{no SC for } \Delta)}
\]

\[
= \frac{P(\text{SC} > \Delta + \delta)}{P(\Delta < \text{SCD} \leq \Delta + \delta)}
\]

\[
\Rightarrow \quad \text{Build statistical model of SCD based on genre}
\]

General approach for model based edit effect detection

- **model effect**
  - design a function $E(t)$ which describes the effect
  - transition can be modeled as a function $E(t)$ which can be written as a combination of the images $X$ and $Y$

- **design detector**
  - consider for example what remains constant in the effect which then forms the basis for detecting that a frame is part of a special effect
General model for fade-in/out and dissolve

T1 start of fade-out of X, T2 start of fade-in of Y
L1 length of the fade-out of X, L2 length of the fade-in of Y

\[ E(t) = \begin{cases} \frac{(t-T_2)}{L_2} Y & t \in [T_2, T_2+L_2] \\ 1 - \frac{(t-T_1)}{L_1} X & t \in [T_1, T_1+L_1] \end{cases} \]

\[ \frac{d}{dt} E(t) = \begin{cases} \frac{Y}{L_2} & t \in [T_2, T_2+L_2] \\ \frac{X}{L_1} & t \in [T_1, T_1+L_1] \end{cases} \]
Differentiation: take the pixelwise difference between the derivatives in subsequent frames

Observation: for the duration of the fade-in this should be constant

Detection: try different values for $L_2$ and see whether the formula below holds for the frame $Y$ at $t+L_2$

Drawback:
- too many types of transition to model
- Not robust against motion
Model-based technique

- **Features:**
  - Histogram differences
  - Audio cepstral difference
  - Motion vectors

- **Train HMM using label data, detect using viterbi algorithm.**

- **No threshold**

- **Two cut state as a sharp cut may last for two frames in interlace video**

- See Boreczky, Wilcox. 98. A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features
Scene and Story Segmentation

- Much harder problems
- Domain information can help a great deal
- Graph-based technique
  - Scene Transition Graph
- More next lecture
Video Scene Layout

Let the use of capitals with a number indicate that the shots are equivalent. E.g. below B, C could denote two views of people in an interview, where D shows both of them in one view.
Scene Transition Graph

- Indicate shot changes by arrows
- Loosely defined: scenes correspond to groups of shots that are “highly” connected
1. Detect shots using any algorithm
2. Time-constrained clustering: group shots together if
   - they are visually similar (ex. Using Hausdorff distance)
   - time difference less than T
3. Create the graph
   - connect two clusters by an arrow if they contain consecutive shots
4. Partition the graph based on “cut-edges” – edges whose removal lead to disconnected components
Scene transition graph

See Yeung, Yeo, Liu 96
Problem:
- Rapidly growing quantities of digital video
- Increasing research in content-based retrieval from digital video
- But no common basis for evaluation/comparison of approaches

Approach:
- Find as much video data as possible and make it available to the community of researchers
- Use the data to build an open, metrics-based evaluation

History
- TREC = Text REtrieval Conference, sponsored by NIST and DoD, started in 1992 to study text IR and speech recognition
- TREC Video Retrieval is a separate activity started in 2003.
TRECVID 2003

- Data: 133 hours (1998 ABC/CNN news + C-SPAN)
- Four tasks:
  - Shot-boundary determination
  - High-level feature extraction (17)
  - Story segmentation and classification
  - Search (manual and interactive)
- Participating groups: 24
Shot-boundary Detection

Results for cut detection

Results for gradual transition detection

Performance of individual groups: best appear to come from IBM (yellow dots) and CLIPS-IMAG (green triangle)
Story segmentation & classification

- **Segmentation task**
  - Identify story boundaries in CNN and ABC new shows
  - Evaluation based on precision & recall, boundaries have to be within +/- 5 seconds interval around ground truth boundaries

- **News classification task**
  - Annotate stories as either news or non-news
  - Evaluation based on percentage of correctly identified news story footage

- **Three categories:**
  - Audio-Visual (features provided by CMU)
  - AV + Automatic Speech Recognizer output (LIMI)
  - ASR only
Results for segmentation

Performance of individual groups: best come from National University Singapore
Feature Extraction

- Identify the following high-level “features” in the video:

11. Indoors
12. News subject face – not anchor person
13. People – at least three humans
14. Building – walled structure with roof
15. Road
16. Vegetation in its natural environment
17. Animal
18. Female speaking (visible, audible)
19. Car/truck/bus – exterior of ....
20. Aircraft
21. News subject monologue
22. Non-studio setting
23. Sporting event
24. Weather news
25. Zoom in
26. Physical violence
27. Madeleine Albright – visible
Results
Search Task

- TRECVID systems, including a human in the loop, are presented with a topic and are to return up to 1,000 shots which meet the need;
- Note the unit of retrieval is the shot, not the news story
- Two search modes: manual and interactive
- Require a user interface from participant
Search Types

MANUAL:

ICONIC

TOPIC

HUMAN

QUERY

SYSTEM

RESULT

Human formulates query based on topic and query interface, not on knowledge of collection or search results.

System takes query as input and produces result without further human intervention.

INTERACTIVE:

TOPIC

HUMAN

QUERY

SYSTEM

RESULT

Human (re)formulates query based on topic, query, and/or results.

System takes query as input and produces result without further human intervention on this invocation.
25 Topics  [total relevant found]

100. Find shots with aerial views containing both one or more buildings and one or more roads [87]
101. Find shots of a basket being made - the basketball passes down through the hoop and net [104]
102. Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at [183]
103. Find shots of Yasser Arafat [33]
104. Find shots of an airplane taking off [44]
105. Find shots of a helicopter in flight or on the ground [52]
106. Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery [31]
107. Find shots of a rocket or missile taking off. Simulations are acceptable [62]
108. Find shots of the Mercedes logo (star) [34]

....
117. Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)