EE 6850 Lecture #4  (Sept. 25, 2002)

- Syntactic-level indexing
  - Shot segmentation

- References
  - Irena Koprinska , Sergio Carrato, "Temporal Video Segmentation: A Survey," Institute for Information Technologies, Department of Electrical Engineering, Acad. G. Bonchev Str., Bl. 29A, 1113 Sofia, Bulgaria, contact e-mail: irena@iinf.bas.bg.
Definition

- **Shot**
  - 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
Why shot/scene segmentation?

- Browsing and visualization
- Non-linear editing
- Event detection
- Demos
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
Program types

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

representation

video

audio

representation → feature → detection

- pixels
- color
- edge
- compression data
- audio

- pixels
- histogram
- DCT
- motion vectors
- audio cepstral
- audio types

- thresholding
- statistical
- model driven

time scale
33 ms
100 ms
few seconds
Factors

- Accuracy
- Speed
- Synaptic
- Generality
Pixel Difference

\[ D(x, y) = \frac{\sum_{i=1}^{N} |P_i(x, y) - P_{i+1}(x, y)|}{XY} \]

Variations: (tolerant to motion and noise)

1. Local Adaptive Threshold
2. Chromatic Image: \( \sum_{x,y} \left( \frac{P_i - P_{i+1}}{P_{i+1}(x, y)} \right) \)
3. Block-based statistical measure

\[ \lambda_k = \frac{\left( \frac{\sigma_{k,i} + \sigma_{k,i+1}}{2} + \left[ \frac{\mu_{k,i} - \mu_{k,i+1}}{2} \right]^2 \right)}{\sigma_{k,i} \cdot \sigma_{k,i+1}} \]

4. Motion-Compensated Difference
Histogram

Count

![Bar graph with bins](image)

Discrete representation of variable

e.g. Quantize pixel value to 4x4x4 in RGB

e.g. HSV

Value → 36 → 16 bins
Saturation → 4 bins
Value → 4 bins

* Choice of color space YIQ, L*a*b*, Munsell color coordinates
  ⇒ better accuracy
RGB, HSI, YIQ

* Non-uniform quantization
  ⇒ use value only at low value and low saturation
  unstable hue
Other Histogram Metrics

\[ D_1(x, x+1) = \sum_i \frac{H_i}{\hat{H}_i + \hat{H}_{i+1}} \]

\[ D_2(x, x+1) = \sum_i \hat{H}_i \]

Histogram Intersection

\[ D_1 = \sum_i \frac{\min(H_i(\xi), H_{i+1}(\xi))}{\min(\sum_{\xi} H_i(\xi), \sum_{\xi} H_{i+1}(\xi))} \]

Quadratic Distance

\[ D_2 = \sum_{i1} \sum_{i2} \left( H_i(\xi) H_{i+1}(\xi) \right) \alpha(\xi_1, \xi_2, \xi) \frac{H_i(\xi_2)}{H_{i+1}(\xi_2)} \]

\[ \alpha(\xi_1, \xi_2) \text{ correlation between colors } \xi_1, \xi_2 \]

\[ \text{e.g. } d_{\xi1, \xi2} \]

\[ \text{e.g. } \quad \begin{array}{c}
\uparrow  \\
\text{red} \\
\end{array} \rightarrow 
\begin{array}{c}
\uparrow  \\
\text{orange} \\
\end{array} \rightarrow 
\begin{array}{c}
\uparrow  \\
\text{blue} \\
\end{array} \]

Other histograms

* Edge histograms + total count
* Texture
* Issue quality of edge, texture extraction, lighting (dark frame)
Compression Data

DCT coefficients

DC
c

horizontal energy

t
vertical

motion vectors

macroblock type: intra mvf mvb
magnitude: direction errors

1. candidate boundary on P frame
2. " B

\[
R_p = \frac{\# \text{intra}}{\# \text{mvf}} \\
R_b = \frac{\# \text{mvb}}{\# \text{mvf}} \\
R_f = \frac{\# \text{mvf}}{\# \text{mvb}}
\]

peak \( R_p \)
peak \( R_b \)
peak \( R_f \) on previous B
peak \( 146^2 \) on I
Detecting Gradual Transitions

Twin Threshold

FD

Cumulative (Running)

$T_c$

2 thresholds: $T_H, T_L$

Model Based

$P_i = \alpha_i \cdot P_{in}$

$+(1-\alpha_i) \cdot P_{out}$

Approximation:

$Var_i = \alpha_i^2 \cdot Var_{in}$

$+(1-\alpha_i)^2 \cdot Var_{out}$
HMM Model Based (Boreczky & Wilcox 98)

self loop model duration
2 cuts model cut between 2 frames
training: av features of each frame label (state of each frame)
testing: Viterbi decoding generates state sequence

video features
1. histogram difference
2. pan estimate zoom estimate
   eq  \( \text{Avg of mv magnitude} = \frac{\sum \text{mv magnitude}}{\text{frame count}} \)

Audio feature likelihood of audio change
Result *HMM adds flexibility higher accuracy
* additional features did
Audio change likelihood

\[
\begin{align*}
2 \text{ sec} & \quad 2 \text{ sec} \\
\downarrow & \\
200 \text{ ms} & \Rightarrow \text{compute 12-dim cepstral vector} \\
\Rightarrow 100 \text{ vectors} & \quad X_1 = (\underline{\bar{v}}_1 \quad \underline{\bar{v}}_2 \quad \bar{v}_{110}) \\
 & \quad X_2 = (\underline{\bar{v}}_1 \quad \underline{\bar{v}}_2 \quad \bar{v}_{110}) \\
\quad \text{every time shift intervals by 200 ms} & \\
\Rightarrow \text{model 100 vectors in each interval} & \\
& \quad \text{by Gaussian Mixture Model (16 Gaussians)}
\end{align*}
\]

Likelihood that \(X_1\) is produced by model \(\Theta X_1\)
\[
L(X_1, \Theta X_1)
\]

Likelihood of different models
\[
L(X_1, \Theta X_1) L(X_2, \Theta X_2)
\]

Likelihood of same model
\[
L_0 = L(Z, \mathcal{Z}) \quad Z \quad \text{combine 4 sec}
\]

Likelihood ratio distance
\[
d(X_1, X_2) = -\log (X_1) = -\log (\frac{L_0}{L_1}) - \log (L_1) - \log (L_0)
\]
Evaluation

Precision/Recall

\[ \text{Precision} = \frac{a}{a+b} \rightarrow \text{false positive} \]

\[ \text{Recall} = \frac{a}{a+c} \rightarrow \text{miss} \]

Confusion Matrix

\[
\begin{array}{c|cc}
\text{Ground Truth} & \text{shot} & \text{non-shot} \\
\hline
\text{Detection} & a & b \\
& c & d \\
\end{array}
\]

Example

\[
\text{Pre} \uparrow \\
\text{Recall} \downarrow
\]
Comparison

- Boreczky and Rowe (1996)
- Training (for setting the thresholds)
  - Some movie clips and a computer animation sequence (44 cuts, 2 dissolves, some pans and zooms)
  - 320x240, 30fps, motion JPEG, 25:1 compression
- Testing
  - 74 mins TV programs (series), 45 mins News (national and local), 79 mins movies, 28 mins commercials
  - Total frames: 419,745, cuts: 2507, gradual: 506
Methods compared

- Histogram difference, single threshold
- Region histogram
  - 16 blocks, 64 gray-scale histograms, difference threshold for each block, and count threshold for changed blocks
- Running histogram
  - 64 gray-scale histogram for each frame
  - Twin thresholds
  - Compute motion vectors, if excessive motion, reject gradual changes
- Motion compensated pixel difference
  - 12 blocks per frame, motion vector for each block
  - Compute average residual errors, if larger than high TH → cut
  - Use cumulative errors to detect gradual changes (similar to above)
  - Use motion vectors to reject false gradual changes
- DCT difference
  - Concatenate 15 coefficients of same locations from different blocks to form a vector
  - Compute (1 – inner product of two vectors from consecutive frames)
FIGURE 1. Recall vs. precision for the Television Program data set

FIGURE 2. Recall vs. precision for the News Program data set
FIGURE 3. Recall vs. precision for the Movie data set

FIGURE 4. Recall vs. precision for the Television Commercial data set
FIGURE 5. Recall vs. precision for the Full data set

FIGURE 6. Recall vs. precision for the Full data set - Cuts only

Hist: consistent accuracy, simple
Region: best, but sometimes cannot achieve high recall
Running: best when recall is very high