10 Video Retrieval – Shot Detection

10.1 Video Abstraction

- Temporal and spatial structuring of the content of a video
- Important for questions related to temporal issues: “Find clips in which an object falls down!”
- Basically, two sub-domains
  - Video modeling and representation
  - Video segmentation and summarization

10.2 Shot Detection

10.3 Statistical Structure Models

10.4 Temporal Models

10.5 Shot Activity

10.1 Example

- News broadcast
  - Story unit:
    - War in Iraq
  - Structural units:
    - Introduction: “The fighting around the city ...”
    - Transmission: various scenes of war
    - Summary: “The reaction of the federal parliament ...”

- Shots
  - Anchorman in a studio
  - Pan across a desert landscape
  - Bombing of a city
  - Refugees
  - Anchorman in a studio
  - Speech in the parliament

- Typical frames for all shots
  - Usually represented by some key frame
10.1 Example

But how can shots be detected?

With the introduction of MPEG-7 shot detection is ready-made

- Metadata standard
  - The correct decomposition is already stored in the metadata
    - Camera information is easy to extract
  - But semantic annotation is unfortunately very expensive
  - Archive material still needs a lot of manual work

10.2 Shot Detection

A clip consists of many scenes
Images belonging to a scene are relatively similar to each other
  - Example: anchorman in the newsroom, desert landscape
For this reason, we do not have to index each individual frame to perform efficient video retrieval, but index only key frames

Problems in finding key frames
  - Detecting a scene transition with hard or soft transitions
    - A hard transition is called a "cut"
    - A soft transition "dissolve" (blending) or "fade in/out"
  - Selecting a representative image, either by random selection, or with regard to the camera movement or an image with average characteristic values, ...

For grouping of frames into shots each transition has to be recognized
  - With uncompressed videos
    - Information from each image is optimally used but the procedure is relatively inefficient
  - Or compressed videos
    - E.g., only data about the change is available

Shot detection in uncompressed videos
  - Template matching (Zhang and others, 1993)
    - Pixel wise comparison: For each pixel (x, y) in the image, the value of the color of the pixel in this frame is compared with the color value in a later frame
    - If the change between two frames is large enough (larger than a predefined threshold), a cut is assumed
    - This only works for hard transitions

10.2 Shot Detection
10.2 Template Matching

\[ D_{cut} = \sum_{x,y} |I(x, y, t) - I(x, y, t + 1)| \]

- It is impossible to distinguish small changes in a wide area of major changes in a small area
- Susceptible to noise, object movements and changes in camera angle

10.2 Histograms

- Histogram-based methods (Tonomura, 1991)
  - Assumption: frames containing identical foreground and background elements have a similar brightness distribution
  - Classification based on the brightness values
  - Histogram columns as the number of image pixels with a specified value

\[ D_{cut} = \sum_{j} |H(j, t) - H(j, t + 1)| \]

- Once again using a predefined threshold we can decide whether there is a cut or not

10.2 Histograms

- Histograms are invariant towards image rotation and change only slightly under
  - Object translation
  - Occlusions caused by moving objects
  - Slow camera movements
  - Zooming
- Significantly less error sensitive than template matching

10.2 Threshold

- Good choice of thresholds is important
  - Too low thresholds produce false cuts
  - Too high thresholds lead to missed cuts
- Selection depends on the type of videos (training)
- Choose the threshold such that as few cuts as possible are overlooked, but not too many false cuts are produced

10.2 Threshold

- Selection, e.g., using distribution functions
  - Differences within the sequences
  - Differences between sequences
  - Selection by minimal error rate
10.2 Twin-Thresholding

- For smooth transitions (dissolves, fades, ...) there are only small changes between consecutive transitions
  - Still, the differences between the middle frames of different shots, are large enough
- Idea: use two thresholds
  - One for the determination of hard cuts
  - And one for the soft cuts

10.2 Twin-Thresholding

- All differences of subsequent frames in the interval \([t + 1, t + n]\) are not computed regarding the direct predecessor, but the reference frame \(t\) (for some fixed \(n\))
  - Only if the difference rises above the threshold \(t_c\), there is a smooth cut, otherwise differences are simply re-formed between consecutive frames

10.2 Block based techniques

- Block-based techniques try to avoid the problem of noise and different camera settings (Idris and Panchanathan, 1996)
  - Each frame is divided into \(r\) blocks
  - Local characteristics are calculated for each block
  - Corresponding sub-frames are compared

10.2 Block based techniques

- Advantages
  - We can detect and ignore effects occurring in only part of the picture through block-wise comparison
    - E.g., movement of the anchorman's head
  - If a high number of the \(r\) blocks are the same in a sequence of two consecutive frames, this is an indication of the frames belonging to the same shot
10.2 Model based procedure

- There are only a small amount of possible transitions between two shots
  - **Idea:** model the transitions as mathematical operations
  - Characteristic temporal patterns in video streams can be detected
  - **Advantage:** this doesn't only recognize transitions, but also their type

- E.g., a temporal model for fades
  - When fading out the pictures of the first shot become darker. The brightness histogram is **compressed** in the x direction
  - Then there are some (almost) black frames
  - When fading in, the images of the second shot become brighter. The histogram is **stretched** in the x direction

10.2 Model based procedure

- This behavior can be interpreted as the application of mathematical operations on the histogram and observed on a stream of frames
  - Defining the start and end of the fade out/in process delivers the shot boundaries
- **Similar models** can be set for other transitions (e.g., dissolve)

10.2 SD in Compressed Videos

- Shot detection in compressed videos is needed due to the size of video data
  - Pixel-based methods for shot detection use uncompressed videos and are therefore usually very computationally intensive

<table>
<thead>
<tr>
<th>Format</th>
<th>Resolution</th>
<th>Frame/ second</th>
<th>Total amount (compressed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264</td>
<td>720p</td>
<td>25</td>
<td>15 MB</td>
</tr>
<tr>
<td>H.265</td>
<td>1080p</td>
<td>25</td>
<td>5 MB</td>
</tr>
<tr>
<td>VP9</td>
<td>720p</td>
<td>30</td>
<td>10 MB</td>
</tr>
</tbody>
</table>

- Shot detection is possible also on the compressed data however trading between efficiency and accuracy
- **Approaches:**
  - MPEG compression information
  - Cosine transformation coefficients
  - I P B frame structure
  - Motion vectors information
10.2 Cosine Transformation

- Video compression is based often on **discrete cosine transform (DCT)**
  - E.g., MPEG, H.264, MotionJPEG, ...
  - The DCT coefficients have correspondents in the real space of the input signal
  - The oscillation of the coefficients can be used for shot detection

- In compression techniques, the image is **divided into blocks** (e.g., 8x8 pixels in JPEG). Each block is separately transformed using DCT
  - The first coefficient (DC) of the DCT is the average intensity of the block
  - A DC-frame is created by using only the DCs of all the blocks and ignore all the higher coefficients

- A sequence of DC frames is called **DC sequence**. DC sequences abstract video clips without having to decode them
- If features are extracted from DC frames we can form traces of sequences, such as the generalized traces (Taskiran and Delp, 1998)
- From these traces it is possible to calculate the probability of cuts for each frame

- Compression based on the encoding of changes between frames
  - I-frames are independently coded (I: independent)
  - P-frames are encoded with change information from preceding I or P-frames (P: predicted)
  - B-frames are interpolations between two P- or I- and P frame (bi-directional)
  - B-frames can thus be calculated both from the preceding, and from the subsequent frame (depending on the encoder)

- A shot is thus a chain of I-, P- and B-frames:
  - IBPBPBPBP ...
- The video stream is *rearranged* for transmission:
  - IPBPBPBPBP ...

- I-frames are **independently** encoded
  - Direct access to the DC component to measure differences between two consecutive I-frames
  - Recognition method with DC-frames are directly applicable
  - Accuracy: between two I-frames there usually are about 15 B- and P-frames
10.2 Motion Vectors

- **(Block) motion vectors** can be extracted directly from an MPEG bitstream.
- They **tend** to change **continuously** within a scene.
  - The number of motion vectors in consecutive frames belonging to the same shot is similar.

10.2 Motion Vectors

- Example of **shot detection** (Zhang et al., 1993)
  - Determine the number of motion vectors in the P-frames.
  - For B-frames count the number of forward and backward movements.
  - Let $M$ be the smaller of two numbers.
  - If $M$ is smaller than a specified threshold, then it probably represents a shot boundary.

10.2 Hybrid Approaches

- Procedures for the use of DCT coefficients and motion vectors can be combined:
  - Increase the recognition accuracy.
  - Utilization of various frame types in MPEG.
  - E.g., Meng and others, 1995.

10.2 Shot Detection

- Shot detection at work with MSU Video tool.
- Shot detection algorithms:
  - Pixelwise comparison
  - Global histogram
  - Block based histogram
  - Motion based detection

10.2 Shot Detection

- E.g., shot detection on Avatar movie trailer
  - Pixelwise (1) and motion based (4) produced 16 cuts.
  - Global Histogram (2) produced 18 cuts.

10.3 Statistical Structural Models

- Idea: decomposition of a video in **semantic units** (shots)
  - Previously: low level primitives (brightness, color information, movements, ...)
  - Now: perceptual features (e.g., visual structure of the whole video).

- Film theory: stylistic elements
  - Montage: temporal structure, editing, ...
  - Mis-en-scene: spatial structure, scenery, lighting, camera position, ...
10.3 Statistical Structural Models

- Goal: build models of stylistic elements
  - Allows the extraction semantic features for the characterization and classification
  - Provides background information for the use of low level features to shot boundary detection

10.3 Example

- Trailer for movie arranged according to average shot length (montage) and activity during shots (Mis-en-scene)

10.3 Example

- Shot duration and shot activity are very rough categories, but have equivalents in movie directing
  - Basic trend: the shorter the shot, the higher the action (and vice versa)
  - If we widely divide the movies into categories action film, comedy and love movies, then we can cluster according to these categories

10.3 Example

- Clusters can be explained through film theory
  - If emotions have to be transferred then long passages of text and detailed facial expressions (a long close-up) are required
  - The development of a character and his connection with the audience takes time
  - Charles Chaplin: “Tragedy is a close-up, comedy a long shot.”

10.3 Example

- For action or suspense, rhythmic patterns are used (e.g., “Psycho” or “Birds” by Hitchcock)
  - Fast cuts require a continuous adaptation of the viewer and create confusion
  - Long dialogues are unnecessary, people express themselves through acts
10.3 Video Structure

• Semantic **structure** assists in categorizing
  – Either based on film theory
  – Or learned from a sample collection
• From high-level structure patterns emerge “more” semantics than from low level features
  – Statistical inference

10.3 Assumption

• The more a video is **structured**, the more semantic information can be derived from it
  – News programs are highly structured and relatively easy to fragment
  – Home made videos are mostly unstructured and almost impossible to fragment

10.3 Classical Elements

• The classical element of the movie direction is the **shot duration**
• Classic elements of the mis-en-scene are more difficult to capture
  – **Activity** in scenes is important
    • Not only between actors (explosions,...)
    • Often correlates to violence
  – But also **mood** (e.g., brightness, colors)

10.4 Temporal Models

• Temporal video structure: **shot boundaries** can be modeled as a series of events occurring in succession
  – Queuing theory: arrivals of persons
  – Modeling through a **Poisson process**
    • Number of events in a fixed time interval follows a Poisson distribution
    • Temporal distance between two successive events is exponentially distributed

10.4 Temporal Video Structure

• **Problem 1**: exponential distribution leads to many short, but very few long shots
• **Problem 2**: exponential distribution has no memory, i.e., the probability that within the next \( t>0 \) time units a shot change will happen, is independent of \( t \)

• **Alternative models**: shot durations are not exponentially distributed, but follow distributions like
  – Erlang distribution
  – Weibull distribution
• **Objective**: estimate the model parameters from a training collection, were the shot boundary is manually determined
  – Maximum likelihood estimate
  – This knowledge can then assist in the detection of shot boundary of unknown videos
10.4 Erlang Model

• Consider shot durations are **Erlang distributed**
  - The length $\tau$ of a (fixed) shot has probability density
    $$P(\tau) = \frac{\lambda^r e^{-\lambda \tau}}{(r-1)!}$$
  - Generalization of the exponential distribution ($r = 1$)
  - Expected value (average shot duration): $\frac{r}{\lambda}$
  - The sum of $r$ independent random variables exponentially distributed with parameter $\lambda$ is $(r, \lambda)$-Erlang distributed

10.4 Erlang Model

• The sum of $r$ independent random variables exponentially distributed with parameter $\lambda$ is $(r, \lambda)$-Erlang distributed
  - It represents a Poisson process since only exactly each $r$-th event is counted
  - $r = 2$: structure of the context of the whole image, followed by a zoom on the essential details
  - $r = 3$: emotional development, followed by an action, followed by the result of this action

10.4 Erlang Model

• **Likelihood function** for a single Erlang-distributed random variable:
  $$L(r, \lambda) = \frac{\lambda^r e^{-\lambda \tau}}{(r-1)!}$$
• Corresponding log-likelihood function:
  $$\log L(r, \lambda) = r \log \lambda + (r-1) \log \tau - \lambda \tau - \log((r-1)!)$$
• Choose the optimal parameters $r$ and $\lambda$ for a sample $T_1, \ldots, T_N$ of $N$ independent and identically Erlang distributed random variables:
  $$\{r^*, \lambda^*\} = \arg \max_{r, \lambda} \sum_{i=1}^{N} \log L(r, \lambda(T_i))$$

10.4 Erlang Model

• Optimization problem over a discrete variable ($r$) and a continuous variable ($\lambda$)
• Film theory: $r$ is small
• Brute-force solution:
  - Test all $r = 1, \ldots, 10$ and compute the optimal $\lambda$
  - Choose the pair $(r, \lambda)$ that maximizes the above expression

10.4 Erlang Model

• If $r$ is known then the determination is simplified
  $$\lambda^* = \frac{N \bar{T}}{\sum_{i} T_i}$$
• Derivative with respect to $\lambda$ and zero values returns:
  $$\lambda^* = \frac{N \bar{T}}{\sum_{i} T_i}$$
10.4 Erlang Model

- Estimation of the parameters $r$ and $\lambda$ from a training collection:

Erlang distribution solves the first problem (distribution of shot durations)

- Problem 2, however, remains
  - The Erlang distribution itself has memory but the exponentially distributed random variables underlying each shot have no memory
  - Solution: Weibull distribution (a generalization of the exponential distribution)

10.5 SD through Shot Activity

- To assess the activity within one shot, we can again rely on low level features
  - One possibility: the difference of color histograms of two consecutive frames
    \[ X(a_n b_n) = \sum_{m=1}^{n} |a_m - b_m| \]
  - Goal: determine a statistical model for the activity within one shots with the help of histograms

Film theory: continuity in editing

- In order not to confuse the audience, the frames separated through cuts should differ clearly
  - Segment the video into regular frames (state $S = 0$) and shot boundary ($S = 1$)
  - Attempts to classify each frame either as regular frame or short-boundary
  - Additionally use low level features such as color histograms

10.5 Shot Activity

- Experience:
  - Training data for shot activity can not be approximated good enough by means of “standard deviation”
    - Therefore use several different distribution components (Vasconcelos and Lippman, 2000)

- Activity within shots ($S = 0$)
10.5 Shot Activity

- Activity in shot transitions ($S = 1$)

[Image showing a mixture of two random variables: a normal, and a uniform distribution]

10.5 Shot Boundary Detection

- Application of statistics:
  - Given: two frames, there are two hypotheses:
    - $H_0$: there is no cut in between ($S = 0$)
    - $H_1$: there is a cut in between ($S = 1$)
  - Likelihood ratio test: choose $H_1$ if
    $$P(D|S = 1) > P(D|S = 0)$$
    (or equivalently: $\frac{P(D|S = 1)}{P(D|S = 0)} > 0$)
  - and $H_0$ otherwise ($D$ is the measured distance between the two frames)

- The likelihood ratio test uses no knowledge about "typical" shot duration
- However, we know the a-priori distribution of the shot duration (or we can at least estimate it)
- Therefore, we now use Bayesian statistics to test the two hypotheses
- We obtain in this way a generalization of the basic thresholding method for histogram differences

- Hypothesis $H_1$ (there is a shot change) is valid, if
  $$P(S_{t+\delta} = 1|S_t = 0, D_{t+\delta}) > P(S_{t+\delta} = 0|S_t = 0, D_{t+\delta})$$
  - Equivalent formulation:
    $$\log \frac{P(S_{t+\delta} = 1|S_t = 0, D_{t+\delta})}{P(S_{t+\delta} = 0|S_t = 0, D_{t+\delta})} > 0$$

- If there was a cut at time $t$, and none in the interval $[t, t + \tau]$, then the probability for a cut in the interval $[t + \tau, t + \tau + \delta]$ according to Bayes, is:
  $$P(S_{t+\delta} = 1|S_t = 0, D_{t+\delta}) = \gamma P(D_{t+\delta}|\gamma - 0, \gamma_{t+\delta} - 1) P(\gamma_{t+\delta} - 1|\gamma - 0)$$
  - $\gamma$ is a normalization constant
  - On the other hand, the probability that there is no cut, is:
    $$P(S_{t+\delta} = 0|S_t = 0, D_{t+\delta}) = \gamma P(D_{t+\delta}|\gamma - 0) P(S_{t+\delta} = 0|\gamma - 0)$$
10.5 Shot Boundary Detection

• Thus:

\[
P(D_{t+t+\delta} | S_{t}, t+\delta) = \frac{P\left(\neg D_{t}, t+\delta \mid S_{t}, t+\delta\right)}{P\left(\neg D_{t}, t+\delta\right)} = \frac{P\left(D_{t}, t+\delta \mid S_{t}, t+\delta\right)}{P\left(D_{t}, t+\delta\right)}
\]

\[
= \frac{P\left(D_{t}, t+\delta \mid S_{t}, t+\delta\right)P\left(S_{t}, t+\delta\right)}{P\left(D_{t}, t+\delta\right)P\left(S_{t}, t+\delta\right)}
\]

\[
= \frac{P\left(D_{t}, t+\delta \mid S_{t}, t+\delta\right)}{P\left(D_{t}, t+\delta\right)}
\]

• Supposition: \(D_{t+t+\delta}\) is conditionally independent (with \(S_{t+t+\delta}\)) from all other \(D\) and \(S\)

10.5 Hypothesis Verification

\[
P(D_{t+t+\delta} | S_{t}, t+\delta) = \frac{P\left(\neg D_{t}, t+\delta \mid S_{t}, t+\delta\right)}{P\left(\neg D_{t}, t+\delta\right)} = \frac{P\left(D_{t}, t+\delta \mid S_{t}, t+\delta\right)}{P\left(D_{t}, t+\delta\right)}
\]

Behavior of conditional probabilities for activity (is estimated from the training collection, shot activity)

Behavior of the probabilities for cuts (estimated from the training collection, distribution of shot duration)

• So hypothesis \(H_1\) is valid if the logarithm of the above expression is positive

10.5 Hypothesis Verification

• Intuitive interpretation

– The left side uses information about the “normal” frame distances within shots and shot transitions

– The right part uses knowledge regarding the “normal” distribution of the shot duration (a priori probability)

10.5 Hypothesis Verification

• According to our initial Bayesian approach, we can decide whether there is a shot transition at point \(\Delta_\tau = [t, t+\tau+\delta]\) or not, by using the following threshold based estimation

– If the last cut took place at time \(t\), and we now observe \(D_\Delta\), then and only then there is a new cut, if applicable:

\[
\log \frac{P(D_\Delta, S_\Delta = 1)}{P(D_\Delta, S_\Delta = 0)} \geq \int_{\tau}^{\tau+\delta} p(\alpha) d\alpha = T(\tau)
\]

– This means: with the introduction of a priori probability, the verification of our hypotheses doesn’t depend anymore from a fixed threshold

– The threshold changes dynamically with the time elapsed since the last cut

– The density \(p(\tau)\) can be assumed to be an Erlang or Weibull distribution density
10.5 Erlang Model

- Density function of the Erlang distribution:
  \[ e_{\lambda}(\tau) = \frac{\lambda^{\tau} \cdot \tau!}{(\tau-1)!} \]

- For the Erlang model, the following threshold function results:
  \[ T_{\lambda}(\tau) = \log \left( \frac{\sum_{i=1}^{n} e_{\lambda}(\tau + \delta)}{\sum_{i=1}^{n} [e_{\lambda}(\tau) - e_{\lambda}(\tau + \delta)]} \right) \]

10.5 Erlang Model

- Typical time distribution of thresholds:

10.5 Erlang Model

- Initially, the threshold is high
  - Cuts are unlikely
  - Cuts are therefore accepted only if the frame differences are very large
- Then, the threshold drops
  - Cuts are accepted for clearly less changes to the features
- Problem is the asymptotic convergence to a positive value
  - Constant level for several consecutive soft cuts

10.5 Erlang Model

- For all Erlang Thresholds we have:
  \[ \lim_{\tau \to \infty} T_{\lambda}(\tau) = \frac{e^{-\lambda \delta}}{1 - e^{-\lambda \delta}} > 0 \]
  and thus there is always such a boundary line

  Threshold
  - The problem comes from the assumption of the underlying exponential distribution in the Erlang model
  - Also here is the solution the Weibull distribution

10.5 Experimental Verification

- Experimental verification (Vasconcelos and Lippman, 2000)
  - Test within a collection cinema trailers
  - Training (determination of model parameters) with the objects from the collection
  - Task: segmentation of a new trailer (“Blankman”)

10.5 Experimental Verification

- Trailer for “Blankman”
10.5 Experimental Verification

• For each trailer simple color histogram distances were used for determining the selected activity.
• The fixed threshold was chosen as good as possible (through tests).
• “O”: Missed cut
• “*”: False estimated cut

10.5 Experimental Verification

• Fixed threshold:

10.5 Experimental Verification

• Weibull threshold:

10.5 Experimental Verification

• Direct comparison of two samples

10.5 Experimental Verification

• Total number of errors:

10.5 Experimental Verification

• Video Signatures
  – Intuitive Video Similarity
  – Voronoi Video Similarity

Next lecture