**12.0 Presentation of the results**

- **Text Retrieval**:
  - Summary (abstract) or
  - Parts of the text with the keywords of the query (snippets)
- **Image Retrieval**:
  - Images with reduced resolution, 'thumbnails'
- **Audio Retrieval**:
  - Short snippet, 'earcon'
- **Video Retrieval**:
  - ?

---

**12.2 Video Skimming**

- Summary (abstract) or
  - Textual summary is not enough especially if the query is based on visual characteristics
  - How is it in practice?
    - State of the art for results presentation in video search

---

**12.0 Presentation of the results**

- So... these existing solutions are not that great
- Are there any other approaches not yet implemented?
  - Yes... and that is video abstraction

---

**Multimedia Databases**

Wolf-Tilo Balk
Silviu Homoeanu
Institut für Informationssysteme
Technische Universität Braunschweig
http://www.ifis.cs.tu-bs.de

---

**12 Video Abstraction**

12.1 Video Summary
12.2 Video Skimming

---

**1.12 Movie Trailers**
12.0 Video Abstraction

- **Video abstraction** is essentially divided into two parts:
  - **Video summary** (still abstracts, storyboards) select frames from a video and put them in a sequence (with sentences extracted from the audio signal)
  - **Video skimming** (moving abstract, summary sequence, highlights) the multimedia presentation form remains the video, but shortens it's duration dramatically

12.0 Video Abstraction

- **Automatic video abstraction**
  - Simple determination of relevance for users
  - Saving transmission time and bandwidth
    - Videos are usually not on the local system of the user
    - Particularly in connection with access from mobile devices
  - Useful for advertising: e.g., movie trailers

12.1 Video Summary

- **Video structure:**
  - Summaries are usually based on shots

12.1 Video Summary

- The video summary should contain all the essential parts of the video
  - **Sampling** of the entire video
  - **Key frames** can be selected according to the shot detection

12.1 Key Frames

- Selection of key frames for each shot
  - First/middle/last frame of a shots
  - Frame, with average color, etc.
  - Randomly selected frame
  - ...

- Which key frames should be included in the video summary?
  - Important: Keep the order of the frames (for better understanding)
12.1 Key Frames

- **Difference-based selection** of the key frames for the video summary
  - Always choose the key frame of the first shot
  - Compare (color, object motion, etc.) the key frame of the next shot with the last chosen key frame for the summary
  - If the differences are “large enough” then pick it as the new key frame in the summary

12.1 Audio Information

- Often, the storyboards are backed with audio information
  - Combine each key frame with the most important sentence of the corresponding shots
  - Extraction of the most important sentence from a text e.g., through Latent Semantic Analysis (LSA)
  - Consider each sentence of a shot as a document and perform term frequency
  - Choose a LSA segmentation and for each shot the sentence with the highest value in the LS space

12.1 Example

- **Term-sentence matrix**
  - \( S_1 \) – \( S_6 \): topics “Road transport” and “Health”
  - \( S_1 \rightarrow S_5 \): key frame with the most important term “Driver” in sentence \( S_6 \)

- **Singular vector 1**
  - Choose sentence \( S_5 \) for the road transport shot and set \( S_3 \) for the health shot

12.2 Video Skimming

- **Summary sequences**: provides an overall impression of the entire video in significantly reduced time
- **Simplest approach**: uniform temporal scaling of the video
  - E.g., removal of every second frame
  - What happens to the audio?
12.2 Video Skimming

- Uniform scaling is only partially useful
  - Important vs. unimportant shots
  - The meaning of some complex shots can only be grasped either as unchanged or by watching more shots
  - It is better to drop whole shots than to make all shots incomprehensible
  - Leads to video highlights

12.2 Skimming vs. Highlights

- Video skimming

12.2 Skimming vs. Highlights

- Video highlight

12.2 Summary Sequences

- Two key questions:
  - How long will the audience need to capture the contents of a scene?
    - Close-up of a face vs. detailed scene
  - How does syntax affect the intelligibility of the whole video?
    - E.g., the order of scenes and type of each scene (dialogue, action, etc.)

12.2 Summary Sequences

- First automatic video summaries: Informedia project of the Carnegie-Mellon University
  - Analysis of the audio information, to find the important shots and omit the irrelevant ones
  - Automatic speech recognition and a few visual object detectors
  - Still relatively error-prone
  - www.informedia.cs.cmu.edu
12.2 Summary Sequences

- **Steps** to automatically derive a scenic sequence summary
  - Estimate the visual complexity of each shot
    - The minimum time period necessary to understand a shot (related to the visual complexity) can be estimated by means of psychological experiments (e.g., Sundaram and Chang, 2002)
  - Using insights from film theory one can select scenic structures important in understanding the film
  - Compression ratios of about 80% are possible
    - More compressed results through video highlighting

12.2 Video Highlighting

- Highlighting doesn’t attempt to summarize the entire video, but selectively chooses from **important scenes**
  - E.g., movie trailer
    - Scenes are not cut considering the minimum necessary cut to help the viewer understand
    - The atmosphere of the movie should be transmitted and the interest aroused
    - The end is usually not shown

12.2 Basic Procedure

- **Video segmentation and analysis**
  - Segmentation into shots and scenes
  - More detailed analysis of shots, which include text or effects e.g., explosions
  - Analysis of close-ups of faces

- **Clip selection**
  - Which shots should be included in the abstract?
  - Complete coverage of the whole video
  - Special effects

12.2 Basic Procedure

- **Combining the clips** (editing)
  - Maintain order?
  - Type of cuts between individual shots?
  - Appropriate audio track?

12.2 Basic Procedure

- Schematic representation

- **High-quality abstracts according to the film theory contain the following components:**
  - **Relevant entities and individuals:** mostly in scenes with high contrast
  - **Actions:** scenes with strong movement
  - **Mood:** scenes for which the color distribution varies slightly from the average color distribution of the film
  - **Dialogue:** scenes with close-ups of faces and corresponding audio
• Shot detection leads to individual shots, which have to be **heuristically grouped into scenes**
  - E.g., grouping by **common background**
    - Sequential shots with similar color distributions
    - Background color doesn’t usually change too much with camera movements
    - Most significant change between scenes

• **Alignement** of video and audio-cuts
  - A video cut without audio change is usually not a change of scene

• **Dialogues** are only present within a scene
  - If successive shots can be classified together with the audio as a dialogue, then they belong to a scene

• **Classification of effects** within of scenes (Pfeiffer et al, 2001)
  - **Faces of the actors** are often essential for a highlight
    - Algorithms for face recognition (e.g., Rowley and others, 1995) have detection rates of about 90% with very few false positives

• **Close-ups** of faces are often associated with dialogs (remove shots where the face occupies less than 30% of the image)
  - **Important people** appear often in the video (remove all the shots of faces, which occur rarely in the movie)
  - **Alternating shots** of faces of different people represent a dialog

• **The basic idea** of most algorithms for face recognition is the training of neural networks with pictures on which the position of the eyes and nose are manually selected (salient points)
  - Accelerate the recognition through **color filters**, which filter frames with dominant skin color
  - **Visually similar faces** are grouped (e.g., using the distance between the eyes)
12.2 Video Analysis

- **Text extraction** from the title should occur in the abstract
  - Segment text regions in frames at the beginning of the video (high contrast, contiguous region with low color variance, often moved linearly over several frames)
  - Clustering rows using vertical and horizontal size and distance
  - The title usually has the largest font size

12.2 Video Analysis

- Select the contents of the cluster with the largest surface area per row as a bitmap of the title and use OCR to convert it to text
- The procedure works reasonably well for stationary or linearly moving text

12.2 Video Analysis

- Recognition of special effects such as explosions, gunfire, etc.
  - Integrating such scenes in the trailer arouses interest
  - Detection especially in the audio track: calculate loudness, frequencies, pitch, etc., in small time windows, and recognize acoustic events

12.2 Abstract Generation

- Setting a goal length controls the selection criteria, otherwise thresholds must be set for the criterias
- Choose text sequences, dialogues, special effects as potential scenes
- Choose a suitable shot from each scene (usually too long)
  - For text sequences always pick the shot with the (probably) title text

12.2 Abstract Generation

- Choose the action-rich shots (determined from the motion vectors) from a scene, since they offer a lot of action in a short time
- Choose the shots whose color distributions best correspond to the average of the video
- Automatic genre detection can select appropriate shots using typical parameters for the genre

12.2 Abstract Generation

- Choose dialogues and special effects
  - Distribute the selection of shot for dialogue and special effects as evenly as possible over the whole video
  - For movies: avoid shots from the end
- Fill the highlight with shots belonging also to other types of scenes
  - Distribute the fill-up scenes for the highlight as well as possible on the video taking into account already chosen shots
12.2 Editing

- When editing the possible operations are establishing the **order** of frames and the **type of a cut** between the shots
  - The order of the frames greatly influenced the **understanding** of the audience
  - If one does not keep the **original** order, an order should at least be kept in the frames belonging to the same class e.g., dialog, special effects, fill-up scenes and text

12.2 Editing

- **Hard or soft cuts** (transitions, etc.) can be used between shots
  - Special effects and action-rich scenes should always be separated from the environment through hard cuts
  - Text, dialogue and fill-up scenes can also be separated by smooth transitions

12.2 Editing

- **Simple principles for scene transitions** in highlights

<table>
<thead>
<tr>
<th>Scene Type</th>
<th>Hard Cut</th>
<th>Editing Cut</th>
<th>Other Cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene Clips</td>
<td>hard cut</td>
<td>hard cut</td>
<td>hard cut</td>
</tr>
<tr>
<td>Editing Clips</td>
<td>hard cut, solo</td>
<td>hard cut, solo, drift</td>
<td>hard cut, solo, drift</td>
</tr>
<tr>
<td>Other Clips</td>
<td>hard cut, solo</td>
<td>hard cut, solo, drift</td>
<td>hard cut, solo, drift</td>
</tr>
</tbody>
</table>

12.2 Audio Editing

- **Audio editing**: more difficult, since the content of the **audio track** can't be automatically abbreviated
  - It is impossible to copy the original sound of each selected frame and integrate it in the summary
  - Audio segments of the **special effects** should be kept if possible
  - In **dialogs**, audio cuts must have priority over shot boundaries

12.2 Audio Editing

- The audio tracks of the **filler scenes** are ignored
- **Fading transitions** should be used between audio tracks
- It helps to use a background audio track (e.g., the soundtrack, if available)
  - For dialogues and special effects, the music is reduced in volume
12.2 Highlight: Groundhog Day

- Groundhog Day

12.2 Groundhog Day: Trailer

- Original trailer:

Next lecture

- Indexes for Multimedia Data
  - Tree Indexes: R-Trees, M-Trees