Prof. Dr.-Ing. Stefan Deßloch
AG Heterogene Informationssysteme
Geb. 36, Raum 329
Tel. 0631/205 3275
dessloch@informatik.uni-kl.de

Technische Universität Kaiserslautern

Chapter 6 - Video

Video Multimedia Object

- Combination of image (raster/vector) and audio
- Raw data:
  - enormous data volume:
    - 25 images/s, 250KB/image
    - audio with 11 bit, 16 kHz
    - results in 6250 KB + 22 KB ≈ 6,3 MB per second
  - initially required special storage devices
    - video tape/recorder (VCR), analog picture disc ("laser disc")
- Registration data
  - recording format (VHS, Beta, U-Matic, …) or recorder/player to be used (controlled by computer)
  - time codes
  - file format (MPEG, …)
- Description data
  - structure (scenes):
    - individual scenes/shots (first frame, length)
    - type of shot: panorama, wide shot, figure shot, close-up, pan, zoom

JPEG

- "Joint Photographic Expert Group"
  - joint activity of ISO/IEC JTC1/SC2/WG10 and Q.16 committee of CCITT SGVIII
  - ISO (international) standard since 1992
- Standard format for raster images
  - support for high compression rates
  - as motion-JPEG used for video, *foundation for MPEG*
- Configuration
  - user can decide about quality of the picture, duration of compression, size of the compressed image
  - compression modes
    - lossy, sequential, DCT-based: baseline mode
    - lossy, extended, DCT-based: set of alternatives to base mode
      - allows progressive mode (image constructed non-sequentially, from blurry to sharp)
    - lossless: low compression rate, no advantage over other formats
    - hierarchical: image stored with different resolutions, each using one of the modes above
- Methods – see literature for details
  - steps: create 8x8 blocks, discrete cosine transformation (DCT), quantization, encoding

H.261 (p x 64)

- Standard for transmission of moving images over ISDN
  - symmetric method for video phone, video conferencing
  - narrow-band ISDN connection: two B-channels (64 kbit/s each),
- Image/frame size
  - 288 lines of 352 pixels (3:4 ratio) for luminance, 144x176 for chroma (Common Intermediate Format – CIF, for videoconferencing)
    - i.e., only 1 color pixel for 4 brightness pixels
  - support for half resolution (QCIF) for video telephony
  - compression rate 47:1 (for QCIF, 10 fps, 64kbit/s)
- Two steps of compression
  - intra-frame: compresses single-frame data (like JPEG)
  - inter-frame: considers previous frame, identifies similar blocks, stores only difference and motion vectors
  - resulting data stream: compressed images, error correction information, frame numbers (5 bits), command for "freezing" the last displayed frame

MPEG

- "Moving Picture Expert Group"
  - initially a sub-group of ISO/IEC JTC1/SC2/WG8, now WG11 in SC29
- Video and Audio
  - constant bitrate of up to 1.856.000 bit/s (also suitable for CD-ROM)
  - incorporates JPEG, sequence of still images supported
- Asymmetric compression
  - encoding effort may be way more expensive than decoding
  - max. frame size: 768 x 576 Pixel
- I-frames (intra coded pictures): independent of other frames (like JPEG)
- P-frames (predictive coded pictures): requires previous frame
- B-frames (bi-directionally predictive coded pictures): requires previous and following (I- or P-) frames
- D-frames (DC coded pictures): independent frames, low quality, for fast forward

MPEG (2)

- Stored image sequence
  - may differ from presentation frame sequence due to B-frames!
- Choosing I-, P-, or B-frames
  - application-dependent
  - heuristic: IBBPBBPBBIBBPBBPBBI ....
  - resulting granularity for random access is 9 frames (330 ms), very good compression rate
- Audio: like Audio-CD or DAT
- MPEG-2:
  - 4–100 Mbit/s,
  - allows for scalability in terms of resolution, bitrates, etc.
  - core standard for DVDs, digital TV

MPEG-4

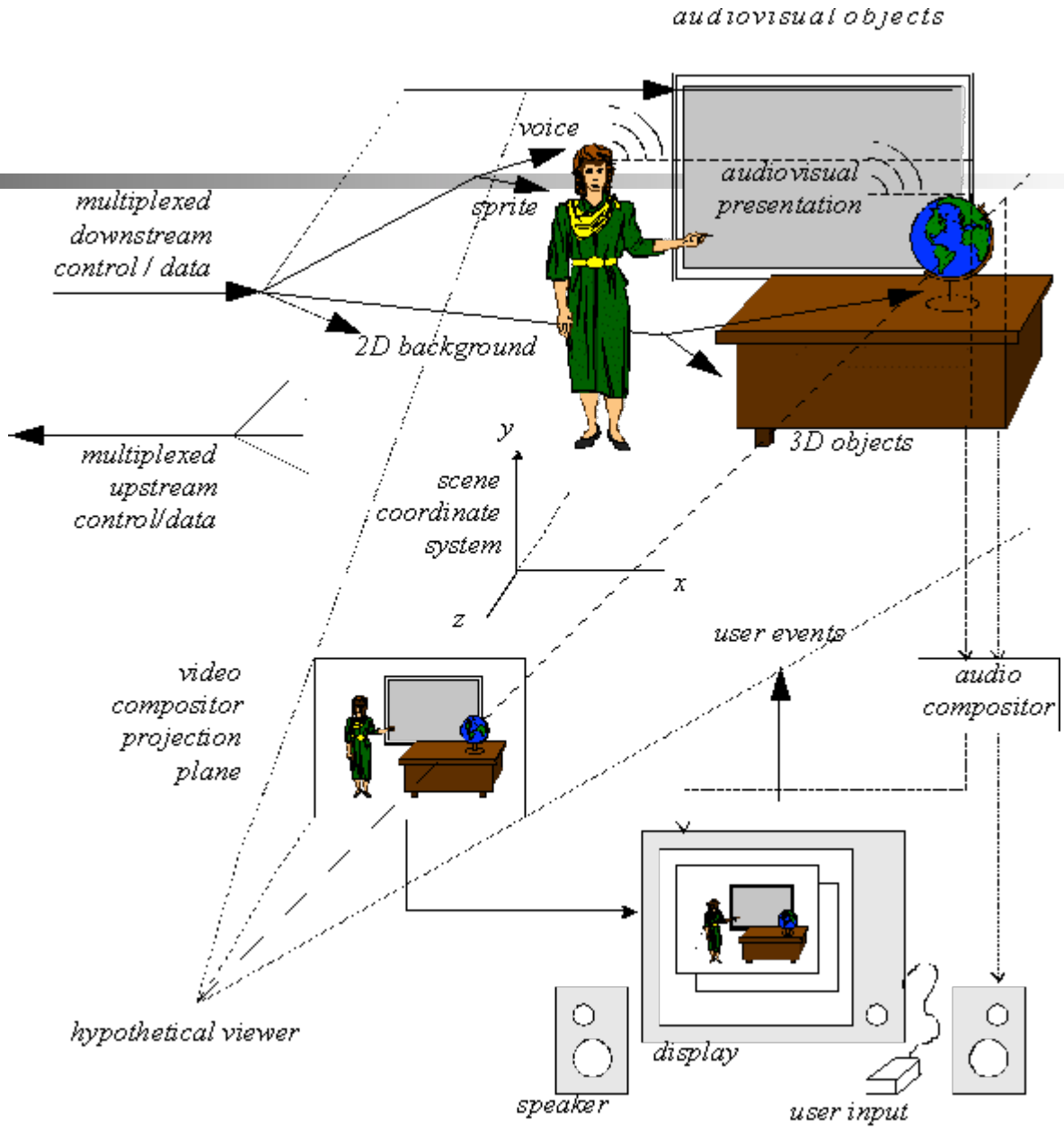
- ISO/IEC international standard 14496
  - defines a multimedia system for interoperable communication of complex scenes that may contain audio, video, synthetic/structured audio (MIDI) and graphics
  - started in 1993, Committee Draft in 1997, International Standard in 1999
- Goals
  - for authors: increased flexibility, reuse
  - for providers: generic QoS-descriptors
  - for end users: more interaction
- Provides standardization for
  - encoding of media objects (recorded or synthetic)
  - composition of media objects resulting in scenes
  - multiplexer and synchronizer for transfer
  - interaction

MPEG-4 (2)

- Parts of the standard
  - systems, video, audio, conformance, reference software, delivery multimedia integration framework (DMIF)
- System
  - framework for the integration of components into scenes
  - hierarchical structure (graph)
  - uses Virtual Reality Modeling Language (VRML)
- Composition
  - frames for audio and video
  - but also **objects**, which make up a scene
    - video objects in different 2D shapes
    - audio objects, possibly associated with video objects
  - description of scenes
    - text, editable or binary (Binary Format for Scene Description, BIFS)

MPEG-4 (3)

- Composition of a scene
  - arbitrary placement in a coordinate system
  - grouping (e.g., voice/sprite)
  - interactive choice of viewer perspective, position
  - information is preserved in the encoding



Video Operations

- Play/view
  - on a separate monitor or in a separate window
  - separate process, which needs to allow control by the user (stop, pause, resume, …)
  - still image (perhaps import into program as a raster image)
  - slow motion, time-lapse
  - possibly other kinds of electronic manipulation (e.g., overlay, bluebox/bluescreen, …)
- Edit, copy, concatenate
  - problems with lossy compression techniques: decompression/re-compression before/after manipulation results in additional loss of quality
- Resynchronization (replace audio track)

Video Search

- Metadata-based
  - title, author, producer, director, cast/actors, production date, type etc.
- Text-based
  - subtitles, captions
- Audio-based
  - audio track
  - speech or music segment
- Content-based
  - images (frames)
  - all, or in a particular group (scene/shot, see subsequent charts)
- Combination
  - multiple of the above techniques used together
- Goal: Search for complete video and for a part
  - user is only interested in a specific scene of the movie, or a part of the news clip

Video Query

- Combined approach proposed by [Bolle+1998]
- Stages of video query
  - Navigation: use metadata to direct the search to specific
    - interval of time
    - topic
    - category or genre
    - video server
  - Searching
    - first based on text (filtering)
      - metadata
      - transcribed audio, captions
    - visual aspects (see most of the following discussion)
  - Browsing
    - inspect high-level overviews/summaries
  - Viewing
    - view result object in its entirety
      - play, pause, fast-forward, reverse, …

Content-based Video Retrieval

- Prerequisite: Segmentation
- Structure
  - Shots
    - filmed with a single camera
    - problem: fading between shots
  - Scenes
    - a series of shots
    - associated with the same situation, part of the film action (i.e., continuous regarding time)
    - e.g., a single dialog
    - harder to identify
    - facilitated (if available) by storyboards, screenplay
- Key frames
  - represent a scene
  - searchable using image retrieval

Segmentation

- Difference between two consecutive frames
  - quantitative aspect: metric
  - threshold
- Simple metric: sum of pixel differences of two consecutive frames
  - not effective; too many false positives
  - fast motion of big objects result in big differences
- Sum of histogram differences
  - distributions remains similar also with motion

$$SD_i = \sum_j | H_i(j) - H_{i+1}(j) |$$

Segmentation (2)

- Threshold
  - critical!
  - approach: average distance of consecutive pictures, plus some small tolerance
- Not applicable for gradual shot changes
  - dissolve, wipe, fade-in, fade-out
  - differences are bigger compared to frames within a shot, but smaller compared to "cuts"
- Idea: use two threshold values
  - difference bigger than T_b: "cut"
  - difference smaller than T_b, but bigger than T_s: maybe a gradual change
  - then add all consecutive differences > T_s and compare with T_b again: if bigger, then the frame sequence is a gradual shot change
  - still low recognition rate: < 16%

Segmentation (3)

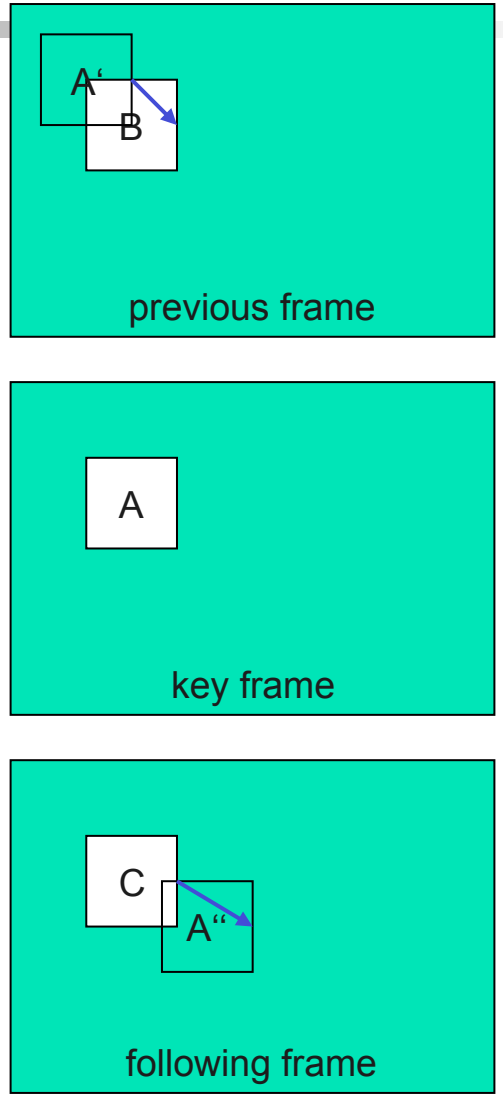
- Recognition errors caused by
  - panning and zooming
    - use motion recognition
  - changes in lighting conditions (lamps, clouds, reflections)
    - normalization before computing differences
- Other approaches
  - motion filter before difference computation
  - edge detection
    - count number of edges that (dis-)appear
    - threshold
  - use information automatically recorded by modern cameras
    - position, time, orientation

Key Frames

- Key frames or representative frames (r frames)
- How many per shot?
  - exactly one
  - proportional to the length, e.g., one per second
  - dependent on content (motion, …)
- Which frames?
  - depending on the number of frames; "segment" is either the whole shot, one second, or anything in between
  - "average picture": take every pixel in the pixel-by-pixel intersection of the frames, then determine the most similar frame
  - use histograms instead of pixels
  - separate foreground from background; compile artificial picture

Motion Information

- Complementing the key frames
- Derive from motion vectors
- Parameter
  - moving content
    - complete motion within shot
  - motion continuity
  - horizontal pan
  - vertical pan
- For complete video, each shot, each key frame



Scenes

- Time-constrained clustering of shots
  - Determine key frames of all shots
  - Compute similarity "classes" of shots
    - based on the visual characteristics
    - constrained by the temporal location of the shot in the video
      i.e., shots that are similar but far apart don't end up in the same group
- Results in a sequence of "class labels": e.g., A, B, A, C, D, F, C, G, D, F …
  - first scene includes shot 1, the last shot with the same label ("A") and all the intermediate shots
  - for each intermediate shot, the scene has to include the first and last shot with the label as well, …
  - here: scene 1 (A, B, A), scene 2 (C, D, F, C, G, D, F)
- Exploits the fact that there is "discontinuity" between the scenes (e.g., at different locations)

Scene Types

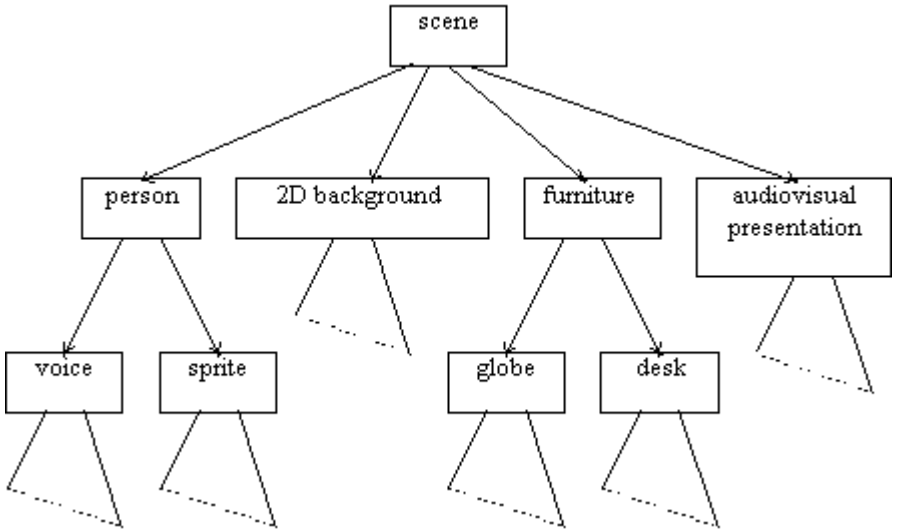
- Films are made using "a system"
  - film language
  - famous book: Daniel Arijon: Grammar of the film language. Hastings House : New York, 1976
  - e.g., dialog:
    - the person speaking is visible in the shot
    - camera "jumps" to various angles/positions
- Idea:
  - consider the shot labels of each scene
  - pattern: ABABAB …
    - includes timing: interval
  - classify based on production "stereotypes", here: dialog
- More general notion of stereotypes
  - consider lack of repetition, average shot length, …
  - example: fast action scene

Visual Summaries for Browsing Results

- Based on techniques discussed above
  - key frames
  - groups/clusters of shots
  - scenes
- Pictorial summary
  - sequence of representative images in temporal order
  - representative image may contain sub-images (e.g., key frames of shot clusters)
- Scene-transition graph (STG)
  - nodes are groups of similar key frames
  - directed edge connects nodes, if one of the shots in the group of the source node directly precedes one of the shots in the group of the target node

Other Options

- Search over objects
  - MPEG-4
- Search over metadata
- Search over annotations
  - MPEG-7
- Combination of the above



Summary

- Video multimedia objects
- Formats and encoding
  - JPEG, H.261, MPEG 1, 2, 4
- Video search
  - meta-data, text, audio, visual content
- Content-based video retrieval
  - segmentation
    - shot detection
    - key frames
    - scene detection
    - scene types
  - visual summaries
  - other options