18-796

Multimedia Communications:
Coding, Systems, and Networking

Prof. Tsuhan Chen
tsuhan@ece.cmu.edu

Multimedia Databases and MPEG-7

Problem Definition

- How to find the desired content in a multimedia database?
- Keywords
  - "NATO", "Bombing", "Yugoslavia"
- Semantic
  - "NATO intensified bombing Yugoslavia"
- Query by example
  - "Find objects like these…"
- Hierarchical approach
  - Low-level features, e.g., texture, color, motion, etc. and domain-specific high-level information

Video Categories

- Documentaries
  - Factual, historical, and biographical
  - Commonly used in multimedia database research
- Broadcast news
  - Pre-recorded vs. live footage
  - Commonly used in multimedia database research
- Sports
  - "plays" through the course of the event, e.g., football, basketball, baseball, soccer, and hockey
- Feature films

"Domain-Specific"

Analytical Features and Extraction

Analytical Video Features

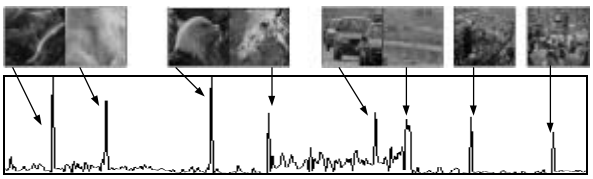
- Features that may be extracted from image or video without regard to content
  - e.g., scene changes, motion flow and video structure in the image domain, and sound in the audio domain
- Most often used in scene change detection, or scene segmentation
- Scene segmentation can be used to separate video content into semantic units

Types of Scene Changes

- Video cut
- Fast cut: A sequence of video cuts, each very short
- Distance cut: a cut from a wide shot to a close-up shot, or vice-versa
- Inter-cutting: scenes change back and forth
- Dissolves and Fades
- Wipes and Blends
  - A wipe is often used to convey a change in time or location

Scene Segmentation

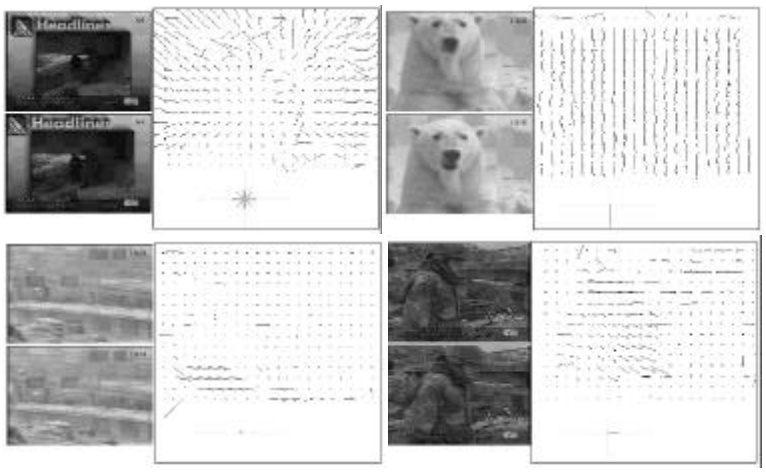
- Methods based on image difference
  - useful for detecting scene cuts, but susceptible to errors
- Edge based segmentation (temporal edges)
- DCT based segmentation
- Motion based segmentation
  - Based on high prediction error
- Color histogram difference



Other Analytical Features

- Motion (optical flow)
  - Camera motion
  - Object motion
- Texture
  - coarseness, contrast, directionality, and regularity [Tamura ‘78]
- Shape
  - e.g., head and shoulders
- Audio
  - Loud sounds, silence, and single frequency sound markers may be detected analytically

Motion Analysis



Video Structure

- News segments are typically 30 minutes in duration and follow a rigid pattern from day to day
- Commercials are of fixed duration, making detection less difficult
- Black frames
  - Between a transition of two segments
  - Between and story and a commercial in a news program

Compressed-Domain Features

Compressed-Domain Approach

- Compressed-domain approach for
  - Feature extraction, e.g., texture, motion, edge extraction
  - Object matching
- Advantages
  - Avoid unnecessary decoding
  - Avoid signal quality degradation in re-encoding
  - Avoid expensive processing in the uncompressed domain

Compressed Domain Techniques

- Scene change detection based on bitrates
  - Intra: bitrate peaks and changes
  - Inter: bitrate peaks
- Scene change detection based on motion vectors
  - P-frame: ratio of intra blocks
  - B-frame: ratio between forward and backward prediction
- Scene change detection based on DCT
  - All DCT coefficients or DC only
- Detection of gradual transition
  - Twin comparison [Zhang et al. ‘93]

Compressed Domain Techniques (cont.)

- Camera/object motion based on motion vectors
- Texture analysis and image matching based on DCT coefficients

- Challenges
  - The method depends on the compression standard
  - To identify the useful features in the the compressed domain
  - To develop new compression standard with content accessibility (e.g., MPEG-4 and MPEG-7)

Content-Based Features

Content-Based Features

- High-level understanding of the content
- The desired result has less to do with analytical features such as color, or texture, and more with the actual objects within the image or video
- In most cases, the query of interest is text-based, so the content is essential

Examples of "Content"

- Human faces
- Captions
  - e.g, video captions, logos, ticker-tape, character listings or credits
  - Horizontal rectangular structure of clustered sharp edges
- Graphics
  - Usually a recognizable symbol
  - Represent institutions, locations, and organizations
- Articulated objects
  - Animal objects, segmented objects, and rigid objects such as planes or automobiles
  - Discrimination of synthetic and natural backgrounds, or an animated or mechanical motion

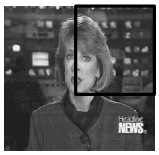
Examples of "Content" (cont.)

- Video structure from content
  - Visual effects introduced during video editing and creation may provide information for video content
    - e.g, scenes prior to the introduction of a person usually describe the person's accomplishments and often precede scenes with close-up views of the person's face; a person's name is generally spoken and then followed by supportive material and the person's face
- Audio and Language
  - Speech recognition and language understanding
- Closed-captions

Matching Techniques

Feature-Based Matching

- Global Image Matching
  - A histogram from the first video frame of each scene is stored and compared with that of video frames in subsequent scenes
- Sub-Image Matching
  - Icon or logo is often used to symbolize the subject of the video. Histogram differencing to a small region in the image we can detect changes in news icons.

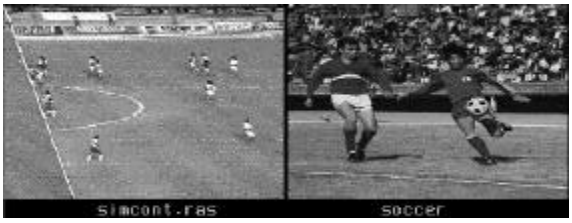



"Histogram"

Content-Based Matching

- Content matching attempts to correlate actual objects with a given query
- Name-It [Satoh '97]
  - Matching a human face to a name in news video
- Spot-it [Nakamura '97]
  - Identify known characteristics in news video for indexing and classification, such as interviews, group discussions, and conference room meetings.
- Pictorial Transcripts [Shahraray '95]
  - Video summarization when closed-captions are used with statistical visual attributes

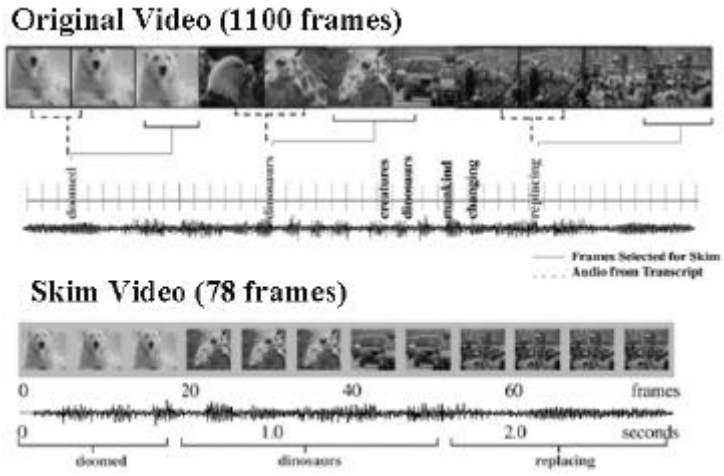
"Content"



Video Summary and Browsing

- Given a long video, how to allow users visualize rapidly?
- Increased playback speed vs. displaying only the video pertaining to a segment's content
- Short text titles and single thumbnail images
  - static nature ignores video's temporal dimension
- Use of multiple modalities
- Browsing through clustering
- High rate keyframe browsing

Scheme of Video Skimming



[Michael Smith]

Demo of Video Skimming

- "Hidden Fury"
  - Skimmed demo (25.33 sec)
  - Original demo (76 sec)
  - 3:1 compaction
- "Mass Extinction"
  - Skimmed demo (55.5 sec)
  - Original (9 min 15 sec)
  - 10:1 compaction

[Michael Smith]

MPEG-7

MPEG-7

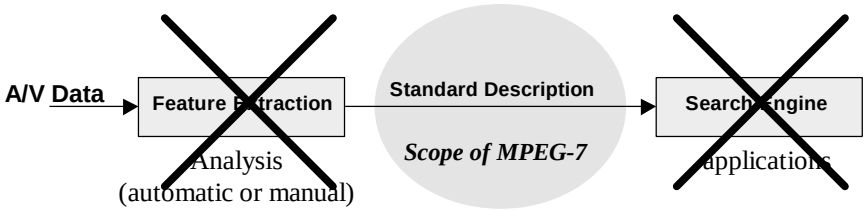
- Growth of digital audiovisual information
  - To find a video clip of Clinton's speech on Internet
  - To find a motorcycle like the one in Terminator II
  - To record TV programs that a viewer like
- "Multimedia Content Description Interface" to standardize the description of various types of multimedia content
  - Still pictures, graphics, 3D models, audio, speech, video, and composition information
  - Special cases: facial expressions, personal characteristics.

MPEG-7 (cont.)

- To enable fast and efficient search and retrieval
  - From text-based search (e.g., keywords) to content-based search (e.g., color, motion)
- cf. PDF or PostScript for drawings/documents

- MPEG-1/2/4 vs. MPEG-7
  - MPEG-1/2/4: Representation of data
  - MPEG-7: Representation of "metadata" (information about data)
  - MPEG-7 may use the shape descriptor in MPEG-4 or the motion vector field in MPEG-1/2

Scope of MPEG-7



Feature extraction is outside MPEG-7

Search and query are outside MPEG-7

Why?

Scope of MPEG-7 (cont.)

- "Standardize the minimum"
- Analysis should not be standardized
  - Can keep improving
  - Room for competition
  - Similar to encoding and segmentation (MPEG-4)
- Search engine should not be standardized
  - Application dependent
  - Room for competition
- Description for the same content may be different for different user domains and different applications

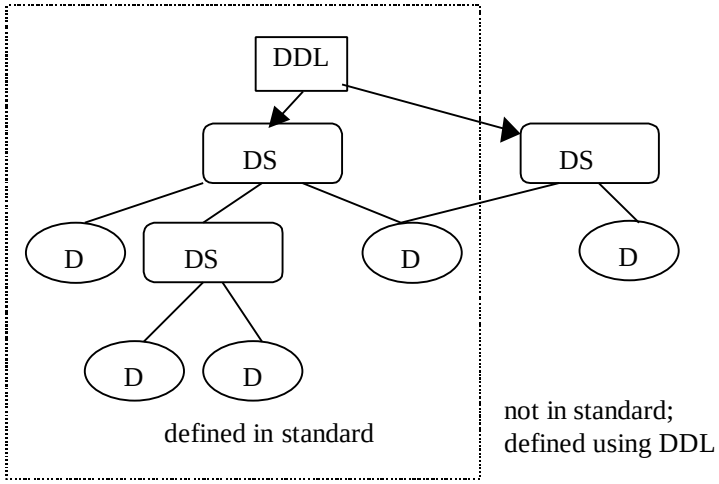
Concepts in MPEG-7

- Data
- Feature: e.g., color, motion
- Descriptor: e.g., histogram, motion vectors
  - Mapping between representation values and the feature
  - Basic unit of a description scheme
- Description scheme (DS)
  - A framework that defines the descriptors and their relationships
- Description
  - An instantiation of a DS
  - Combination of descriptors and DS's

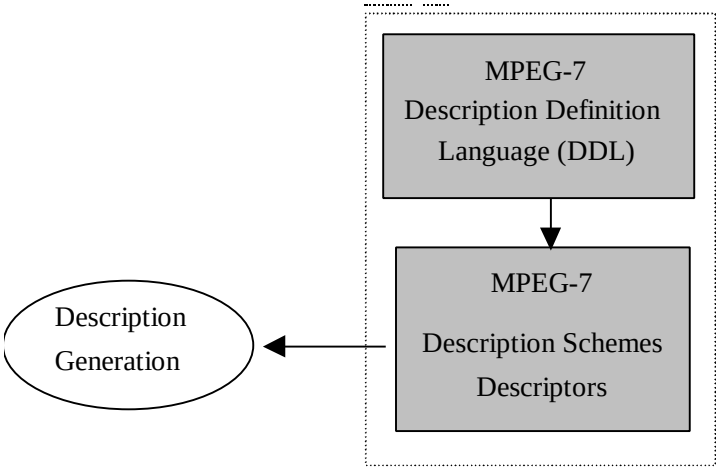
Concepts in MPEG-7 (cont.)

- Coded description
  - Compressed version of the description
- Description Definition Language (DDL)
  - A language to define, modify, and combine DS's
- So, MPEG-7 will standardize
  - A set of descriptors and DS's
  - DDL
  - A scheme for coding the descriptions

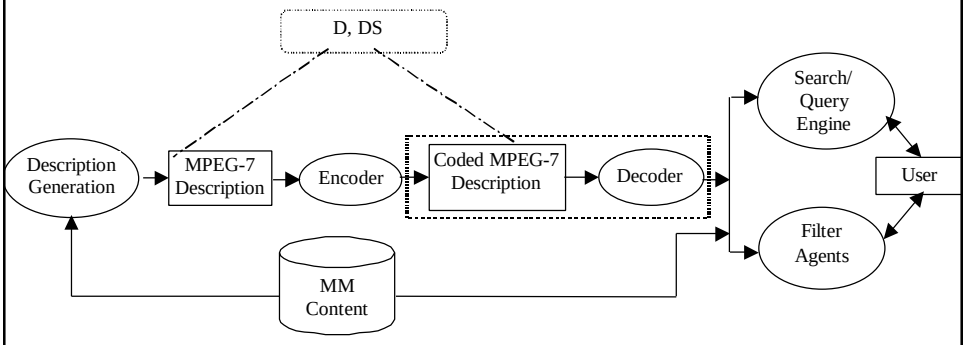
Example Relations Between D's and DS's



Role of D's and DS's



Example Application of MPEG-7



"Pull" and "Push"

- Pull scenario
  - Multimedia databases
    - e.g., video footage libraries
  - Indexing and retrieval
  - Request for inclusion of information
- Push scenario
  - Broadcasting and webcasting
  - Selection and filtering
  - Request of exclusion of information

Example Applications

- Digital libraries
  - e.g., image/video catalog, musical dictionary
- Multimedia directory services
  - e.g., yellow pages
- Broadcast media selection
  - e.g., radio channels, TV channels
- Multimedia authoring
  - e.g., personalized news service, digital photo/video albums

Example Uses

- Music
  - Play a few notes on a keyboard or whistle a melody
  - Retrieve a list of musical pieces that are similar
- Graphics
  - Draw a few lines on a screen
  - Retrieve a set of images containing similar graphics
- Images
  - Define objects, including color patches or textures
  - Retrieve examples among which you select the interesting objects to compose your image

Example Uses (cont.)

- Movement
  - On a given set of objects, describe movements and relations between objects
  - Retrieve a list of animations fulfilling the described temporal and spatial relations
- Scenario
  - On a given content, describe actions and get a list of scenarios where similar actions happen
- Voice
  - Using an excerpt of Pavarotti's voice to retrieve a list of Pavarotti's records or video clips

Work Plan

- Oct 1996-date: Defining the requirements
- Competitive: Call for proposals and evaluation
- Collaborative: Developing the standard
- Time table

Call for Test Material	Mar 1998
Call for Proposals	Oct 1998
Proposals due	Feb 1999
1st Experiment Model (XM)	Mar 1999
Working Draft (WD)	Dec 1999
Committee Draft (CD)	Oct 2000
Final Committee Draft (FCD)	Feb 2001
Draft International Standard (DIS)	July 2001
International Standard (IS)	Sep 2001

Call for Proposals

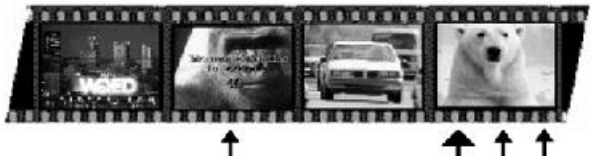
- Seek proposals in
  - Descriptors and DS's
  - DDL
  - Coding methods for compact representation of Descriptions
  - Tools for creating and interpreting DS's and Descriptors
  - Tools for evaluation (A difficult task!!!)

CMU Informedia Project

- Search and retrieval of multimedia content
  - Automated information extraction from video
  - Search and retrieval of spoken language documents
  - Integration of speech, image, and natural language understanding for library creation and exploration
  - Validation through user testbeds

The Scheme

- Decompose video into shots
- Compute representative frame from each shot
- Locate query scoring words



- Use frame from highest scoring shot

Technologies

- Speech recognition for index generation
- Topic detection and tracking
- Commercial detection
- Face detection and matching
- VideoOCR
- Name-It: Name-face correlation
- Video skimming
- Corpus
  - CNN: ~1000 hours + 12 hours/wk
  - Documentary corpus: 400 hours

Levels of Modeling

MODELS	CODED INFORMATION	EXAMPLES
Pixels	Color of pixels	PCM
Statistically dependent pixels	Prediction error or transform coeffs	Predictive Coding Transform Coding
Moving blocks	Motion vectors and prediction error	Block-based coding H.261/263, MPEG-1/2
Moving regions	Shapes, motion, and colors of regions	Region-based coding H.263+, MPEG-4
Moving objects	Shapes, motion, and colors of objects	Model-based coding MPEG-4
Facial models	Action units	MPEG-4
A/V objects	Descriptive languages	MPEG-7

Levels of Modeling (cont.)

- Better image understanding implies
  - Higher compression
  - More content accessibility
  - More complexity
  - Less error resilience
- Currently
  - Block-based: H.261, H.263
  - 2D region-based: H.263+, MPEG-4 Video
  - Model-based: MPEG-4 SNHC
  - High-level descriptive language: MPEG-7

References

- MPEG-7 Call for Proposals
  http://drogo.cselt.it/mpeg/public/mpeg-7_cfp.htm
- "MPEG-7 Context and Objectives"
  http://drogo.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm
- "MPEG-7 Requirements"
  http://drogo.cselt.it/mpeg/public/w2461.html