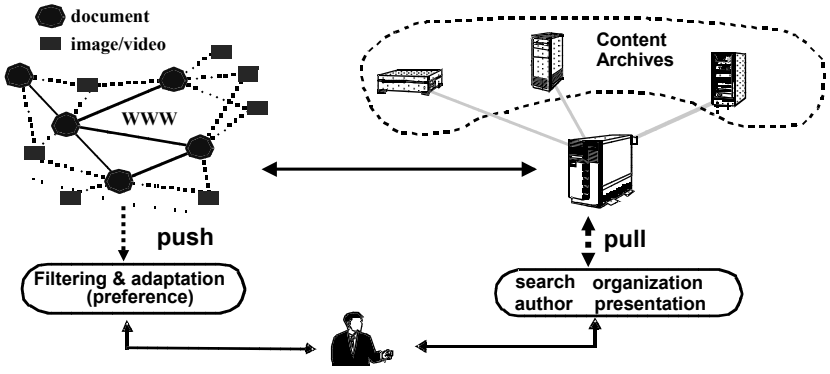
Audio-Visual Content Indexing, Filtering, and Adaptation

Shih-Fu Chang

Digital Video and Multimedia Group
ADVENT University-Industry Consortium
Columbia University
10/12/2001
http://www.ee.columbia.edu/dvmm

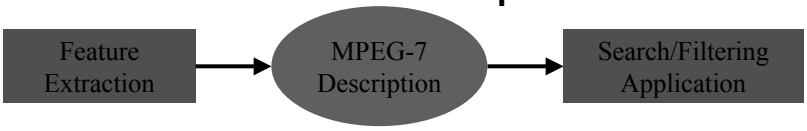
Challenge: tools/systems for searching and filtering



- **Important Issues:**
  - **— Information Overload, Limited User Attention, Time-Sensitive**
  - **— Heterogeneous Platforms**
- **Example Products:**
  - **AltaVista, Google, DoCoMo/IBM, NewsTake, TiVo/Philips, PC DVR**

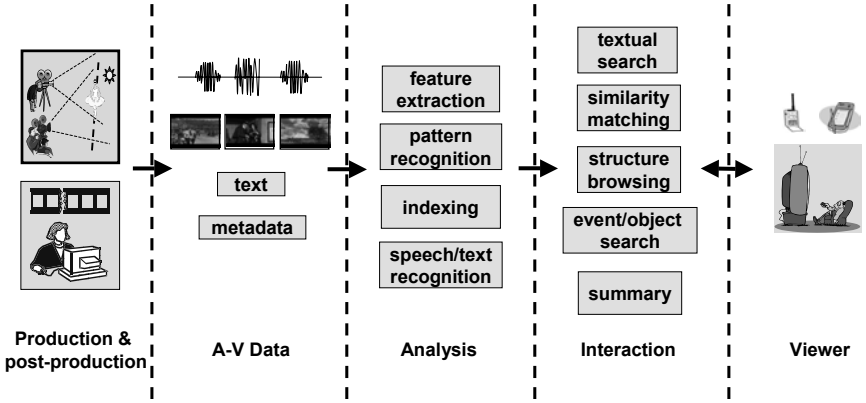
MPEG-7 and issues

- **ISO/IEC 15938**
  "Multimedia Content Description Interface"



- **Issues:**
  - **Active research in analysis and indexing**
  - **System issues: delivery, compression**
  - **What applications? …**

Re-examine Content Production/Usage Chain



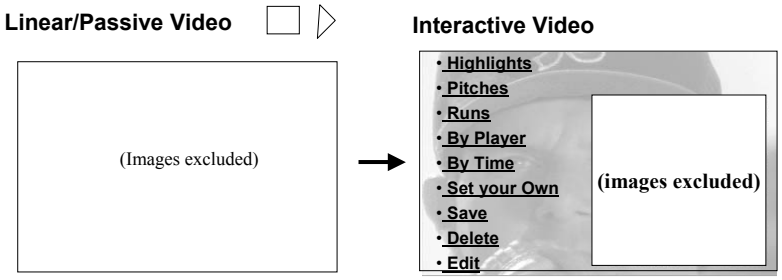
Types of Content Indexing

- Reverse engineering
  - Production structure parsing (shot, camera)
- Metadata preservation
  - Format, Production, Credits, etc.
- Feature measurement
  - audio-visual, spatio-temporal objects and features
- Content recognition and organization
  - Semantic meaning description
  - Scene grouping and skimming

When do we need automated tools

- Conditions for high impact
  - Metadata that's not available from production
  - Work that humans are not good at (e.g., low-level computation)
  - Large volume, low individual value
  - Time sensitive
  - Acceptable performance
- Less Promising Areas
  - Content from digital production tools with metadata
  - Prime content that can afford manual solutions

Case 1: Live Sports Video Filtering and Navigation



- Time sensitive interest
- Massive production and audience
- Time compressibility
- Temporal structure and production rules

Beyond Ringer and Clip Download

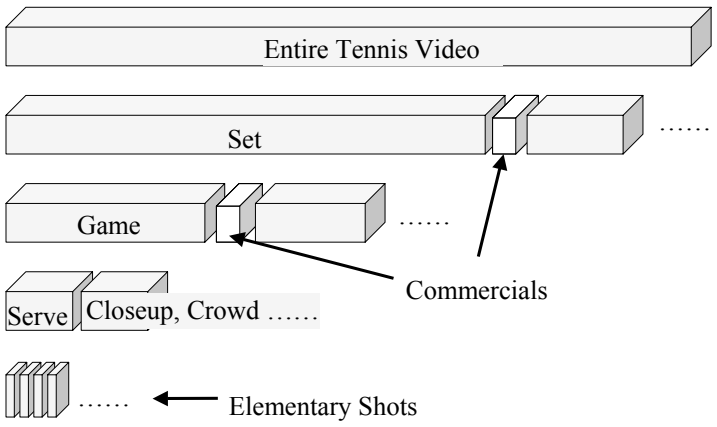
Services:

- Messaging
- Localized information
- Media/game download
- Multi-person games
- TV phone
- On-site purchase

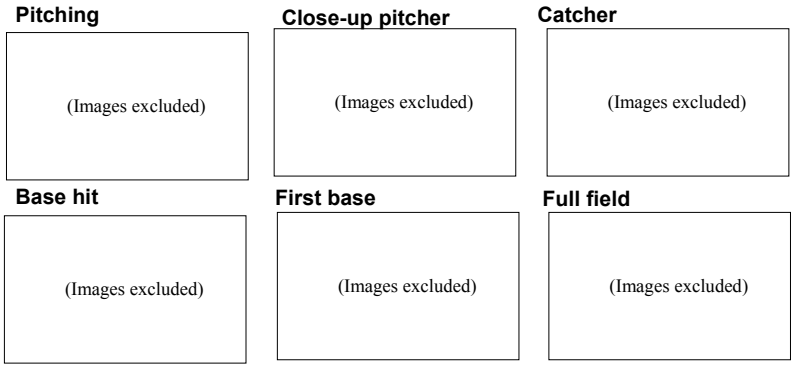
Time-sensitive short video messages suiting personal needs

Image showing sports video on mobile handset excluded

Regular Structure and Views in Sports Video



Regular set of views



- **Important issues:**
  canonical view ←→ beginning of recurrent semantic unit
  view transition pattern ←→ types of events

Real-Time Sports Video Parsing

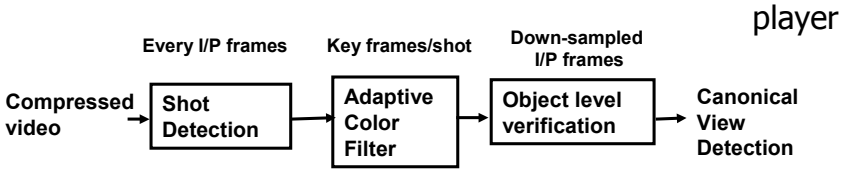
- Detect and classify recurrent semantic units
- Real-time processing
  - simple features
  - compressed-domain processing
- Combine global-level and object-level classification

(Chang, Zhong, Kumar '01)

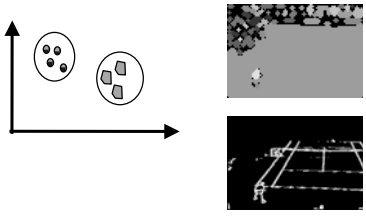
Detecting recurrent semantic units: Serve

Images showing different views of tennis serve excluded

Color, lines, player



(Chang, Zhong, Kumar '00)



Learning spatio-temporal rules

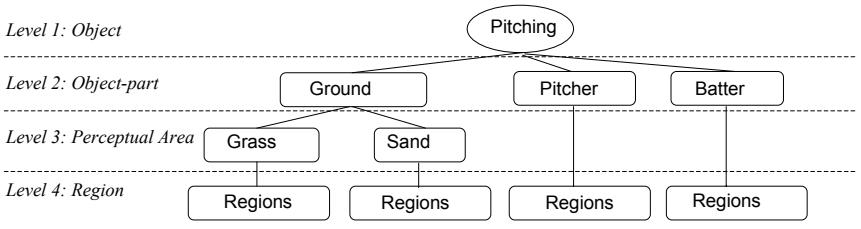
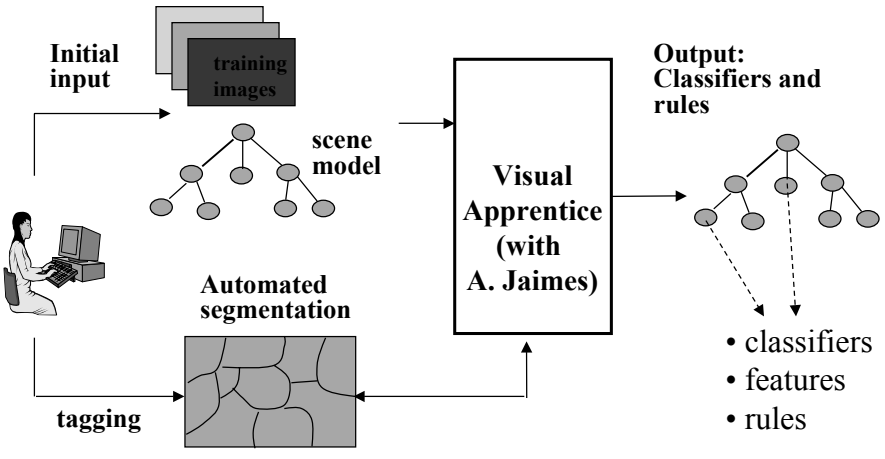


Image showing baseball pitch excluded

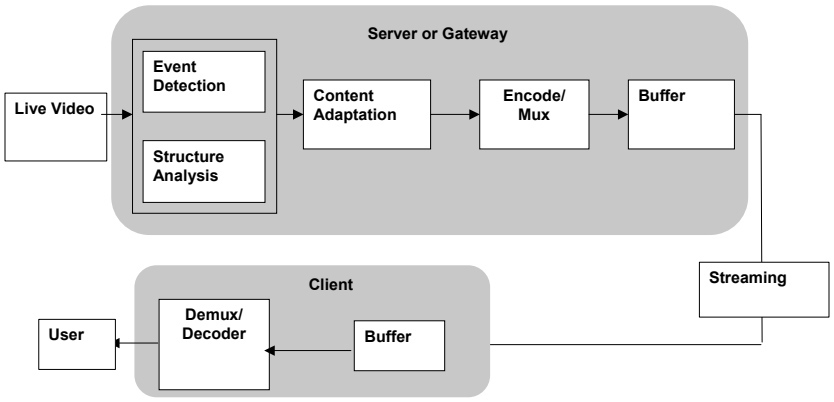
Systematic learning of object rules



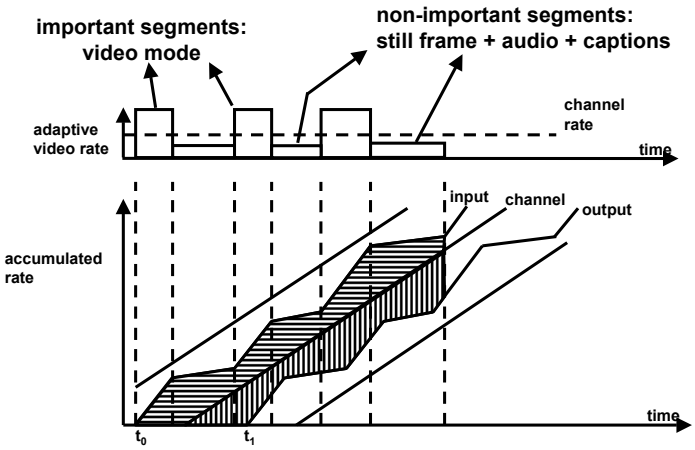
(Jaimes and Chang '99)

Application: Time-Sensitive Mobile Video

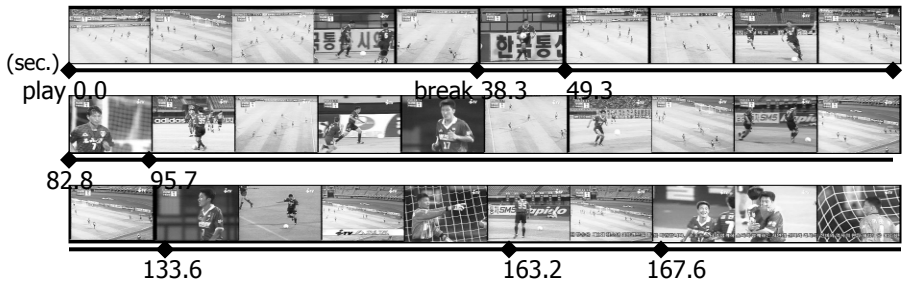
Content Adaptive Streaming



Content Adaptive Streaming



Recurrent semantic units in soccer video?



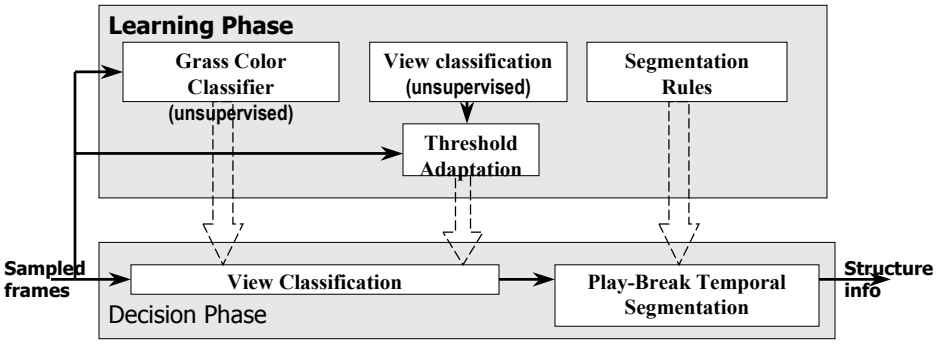
- Game → a sequence of play and break segments
- Play/break → No canonical views/events
- Sporadic events (start and end)
- Shot boundary ≠ play boundary

Heuristic Model: Features→ Views → Structure

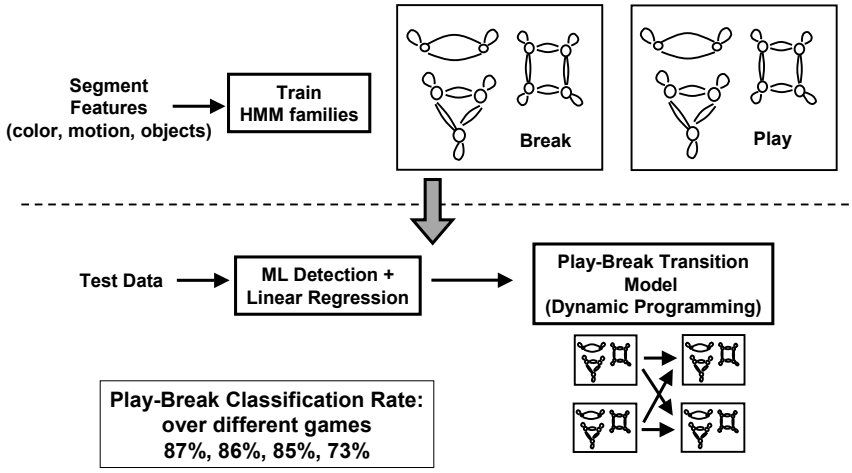

global


zoom-in

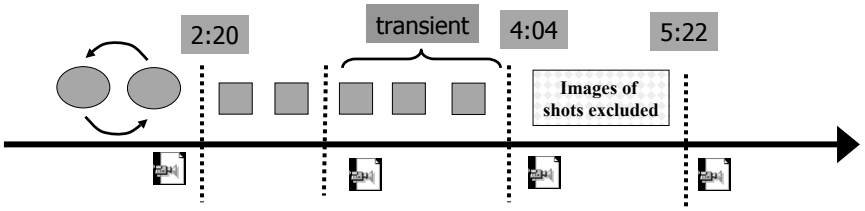
(Image Excluded)
close-up



Modeling Transitions with HMM



Case 2: Long Programs with Loose Structures



- Goal:
  - Scene/Topic Segmentation, Browsing, Preview
- Challenge:
  - Diverse production styles and views

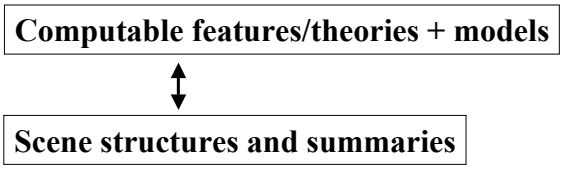
Domain Consideration

- **Film:**
  - **Does not satisfy impact conditions**
  - **Not much value except**
    - archived collection in libraries
    - less popular selection in PVR/STB
  - **Challenging, rich styles and media**
  - **Fun**

Problems and Approaches

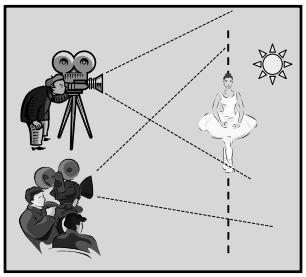
- **Explore production models**
- **Explore multimedia integration**
- **Incorporate structural and special cues**
- **Explore viewer's perceptual models**

- **Goal:**



Production Constraints → Video Scenes

- **Camera placement [180^o rule]**



overlapping background causes chromatic consistency

- **Lighting/chromaticity continuity**

Audition Psychology → Audio Scenes

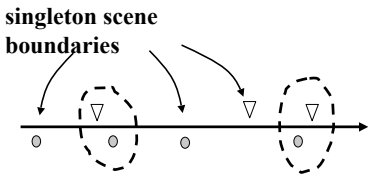
A. Bregman '90: Auditory Scene Analysis

- **Unrelated sounds seldom begin/end at the same time.**
- **Sounds from the same source change properties smoothly and gradually over time.**
- **Changes in acoustic states affect all components of the sound (e.g., walking away from a ringing bell)**
- **Human perception uses long term grouping (e.g., a series of footsteps).**

- **An audio scene change is said to occur when majority of the sound sources change.**

Explore Audio-visual association

- Complementary audio-visual structures
- first hour of Blade Runner
  - video: 28 scenes
  - audio only: 33 scenes
- audio/video agreement: 24
  - video change only: silent or transient scenes
  - audio change only: mood/state change

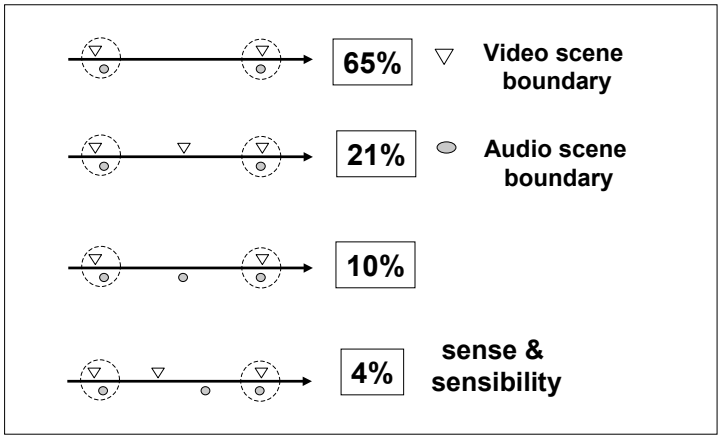


▽ video scene boundary

○ audio scene boundary

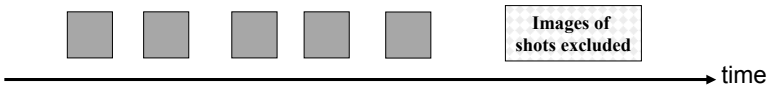
Sundaram & Chang ICASSP '00

Mixing audio-visual scenes



Film: Common Scene Structures

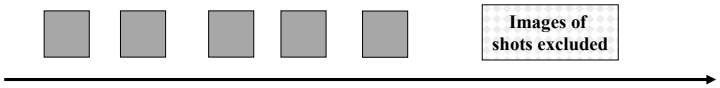
- 4 common scene types in film
  - ***Progressive or Coherent***:
    same location, similar camera takes, consistent audio



(Demo)

Common Scene Types (2)

- ***Transient*** :
  different locations, dissimilar camera takes, evolving audio

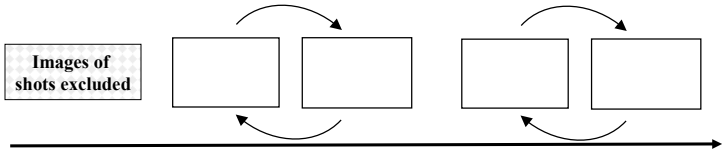


  - ***MTV-type*****: fast, dissimilar shots, consistent audio**

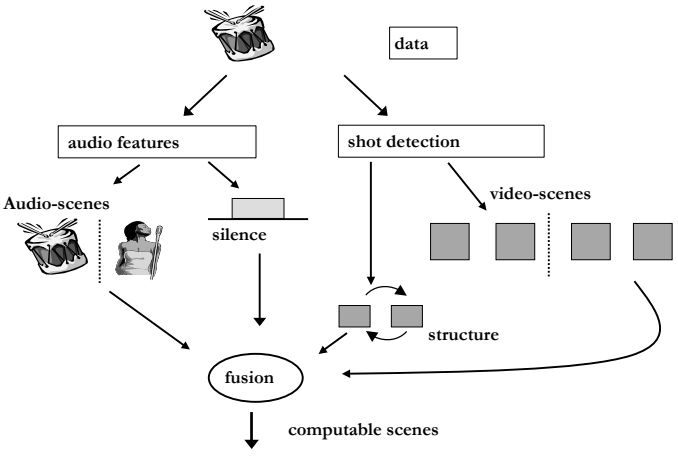
(Demo)

Common Scene Types (3)

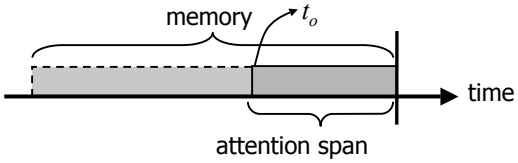
- ***Dialog*: repeating structures, consistent audio**



Computable Scene Detection Architecture



The Viewer/Listener Model

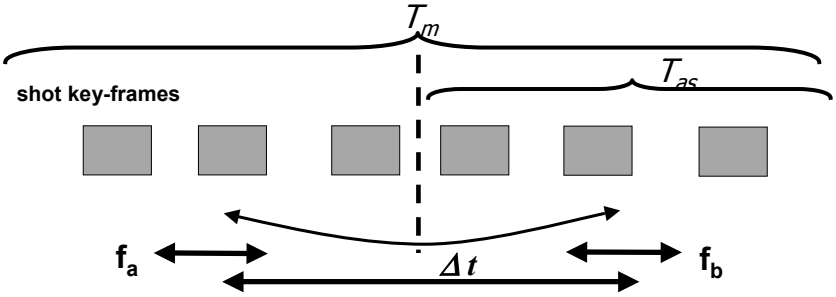


t_o: decision instant

- Memory (M): Net amount of data remaining in viwer's memory, e.g., 32 sec.
- Attention span (AS): The most recent data occupying user's attention, e.g., 16 sec.
- Measure the group, long-term coherence between M and AS

(Kender and Yeo 98, Sundaram and Chang 00)

Video Scene Segmentation

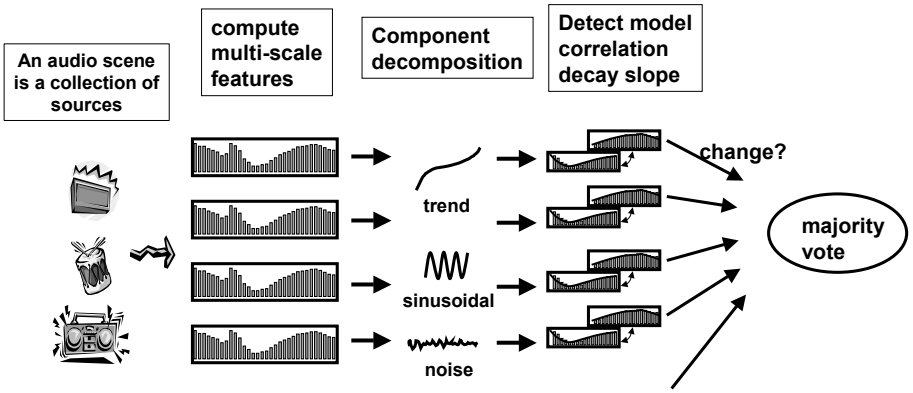


- Group coherence based on viewer memory model

$$R(a,b) = (1 - d(a,b)) \bullet f_a \bullet f_b \bullet (1 - \Delta t / T_m)$$

$$C(t_o) = \left(\sum_{a \in T_{as}} \sum_{b \in \{T_m \setminus T_{as}\}} R(a,b) \right) \Big/ C_{\max}(t_o)$$

Audio Scene Segmentation

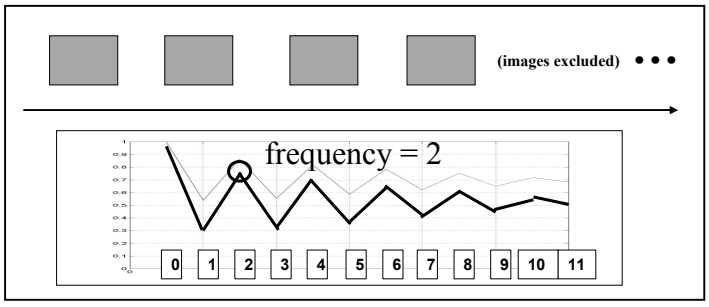


(sundaram and chang '00)

Detecting Dialog Structure

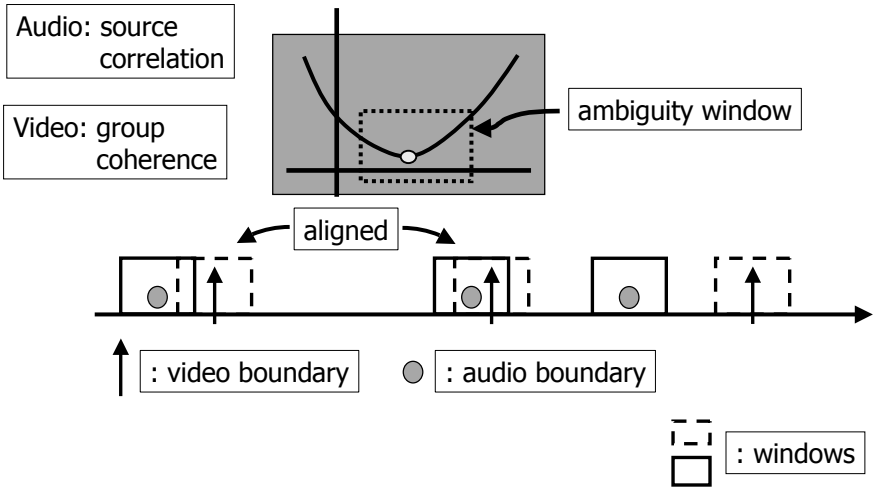
cyclic autocorrelation

$$\Delta(n) \equiv \frac{1}{N}\sum_{i=0}^{N-1} d(o_i, o_{\mathrm{mod}(i+n,N)})$$



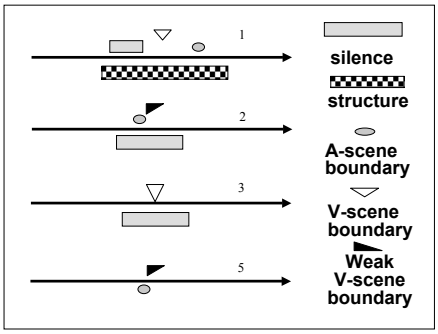
Accuracy: 80%/94% - 91%/100% precision recall

Aligning Scene Boundaries



Fusion Rules

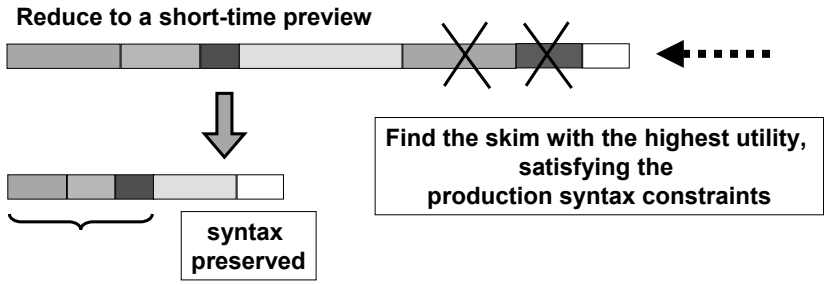
- **Synchronized a-scene and v-scene → scene**
- **Exceptions:**
  - **Ignore scenes within structures**
  - **Ignore (weak v-scene, weak a-scene)**
  - **Add (strong v-scene, silence)**
  - **Require tighter synchronization if (weak v, normal a)**



Open Issues

- **Video**
  - **Use higher-level cognitive models**
    - Currently focus on group coherence between groups
    - Syntax structure within scenes not considered
    - Potential use of information theory and temporal modeling
- **Audio**
  - **Source separation (music, speech, noise, etc)**
- **Scene-level organization**
  - **Visualization, summarization, matching**

Scene Skimming



Utility of a shot

- Explore Perceptual Model and Scene Syntax

(a) (b)

- Is comprehension time related to the visual spatio-temporal complexity of the shot ?
- The presence of detail robs a shot of its screen time.

Shot complexity and comprehension time

- Represent shot by its key-frame

- A shot is selected at random
- The subject was asked to *correctly* answer four questions in minimum time:
  - Who
  - What
  - When
  - Where

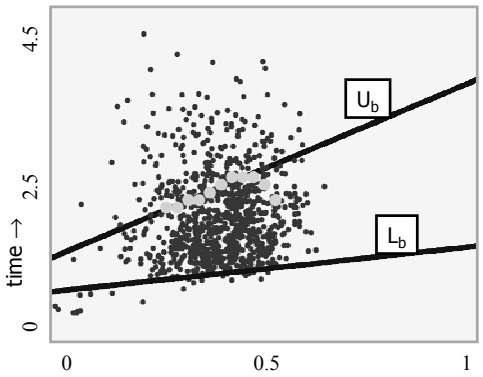
"Why" was not asked

Time-complexity relationship

- **Plot of average time vs. complexity shows two bounds**

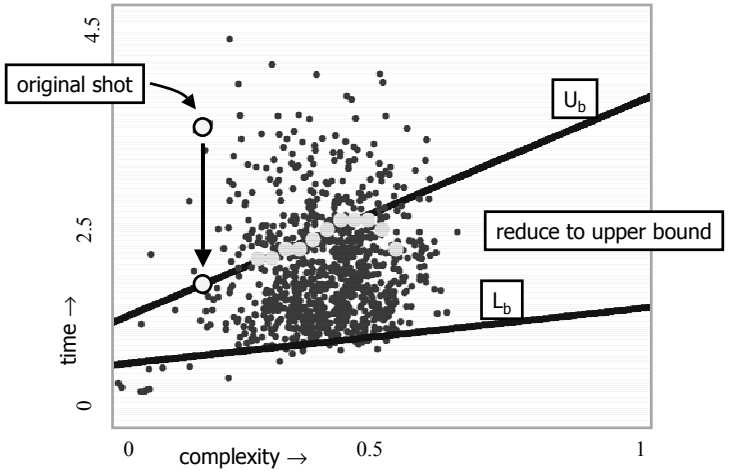
$$U_b(c) = 2.40c + 1.11$$

$$L_b(c) = 0.61c + 0.68$$



comprehension time increases with complexity

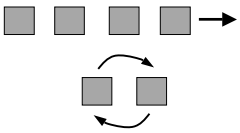
Shot condensation



Film syntax

The specific arrangement of shots so as to bring out their mutual relationship. [sharff 82].

- **Minimum number of shots in a scene**
- **The particular ordering of the shots**
- **The specific duration of the shots, to direct viewer attention**
- **Changing the scale of the shots**



Film makers think in terms of phrases of shots and not individual shots.

The progressive phrase

"Two well chosen shots will create expectations of the development of narrative; the third well-chosen shot will resolve those expectations."
[sharff 82].

Hence, a phrase (a group of shots) must at least have three shots.

Maximal compression: eliminate all the dark shots.

The dialog

"Depicting a conversation between m people requires 3m shots." **[sharff 82].**

Hence, a dialog must at least have six shots

Maximal compression: eliminate all the dark shots.

Utility Optimization Framework

Objective function for each skim

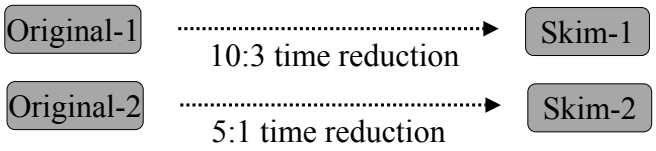
$$O(\vec{t},\vec{c},\phi) \equiv 1 - U(\vec{t},\vec{c},\phi) + \gamma_o R(\vec{t},\phi) + \gamma_l P(\phi)$$

Utility of Skim | **Rhythm Penalty** | **Dropped shot penalty**

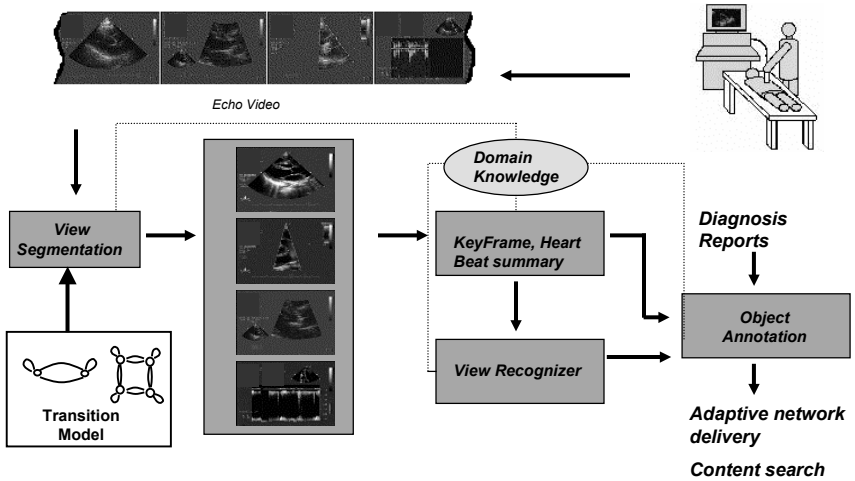
Find the skim with the minimal cost, satisfying the syntax constraints

syntax preserved

Scene Skimming

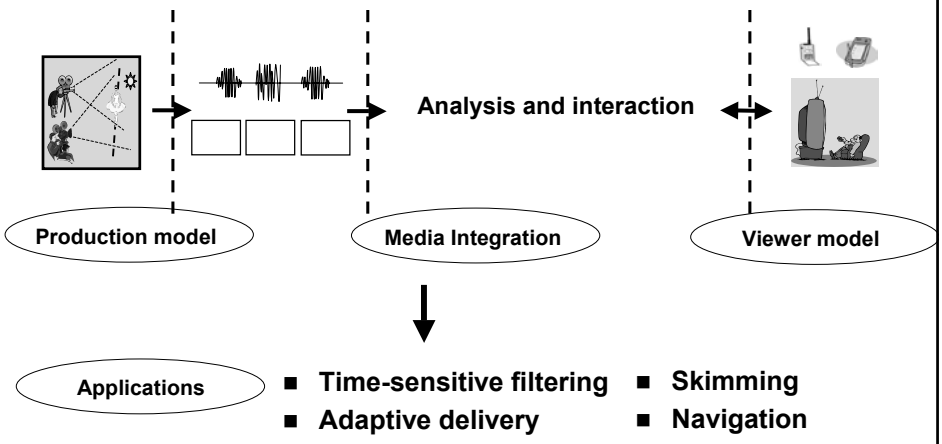


Echo Video Digital Library & Remote Medicine



(Ebadollahi and Chang '00)

Conclusions



Acknowledgements

- Live Sports Filtering:
  Di Zhong and Raj Kumar
- Soccer Video Parsing:
  Lexing Xie, Peng Xu, Ajay Divakaran, Anthony Vetro, Huifang Sun
- Scene Segmentation and Skimming:
  Hari Sundaram
- Medical Video:
  Shahram Ebadollahi

More Information

- **Columbia Digital Video/Multimedia Group**

  http://www.ee.columbia.edu/dvmm

- **ADVENT Industry/University Consortium**
  http://www.ee.columbia.edu/advent

- **http://www.ee.columbia.edu/~sfchang**