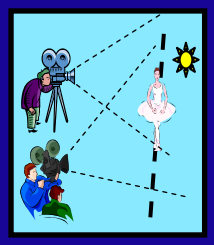
Video Indexing, Summarization, and Adaptation

Shih-Fu Chang

Digital Video and Multimedia Lab
ADVENT University-Industry Consortium
Columbia University
November 4th 2002
http://www.ee.columbia.edu/dvmm

Video Indexing --

Content Chain Principles



Understand Production Syntax

text

metadata

Integrate multimedia

feature extraction

speech/text recognition

Event Detection

Structure discovery

Knowledge mining

semantic search

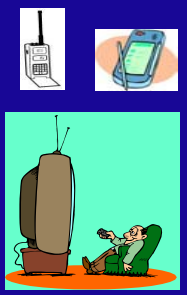
similarity matching

structure browsing

skimming

summary

Explore Multi-modality interaction



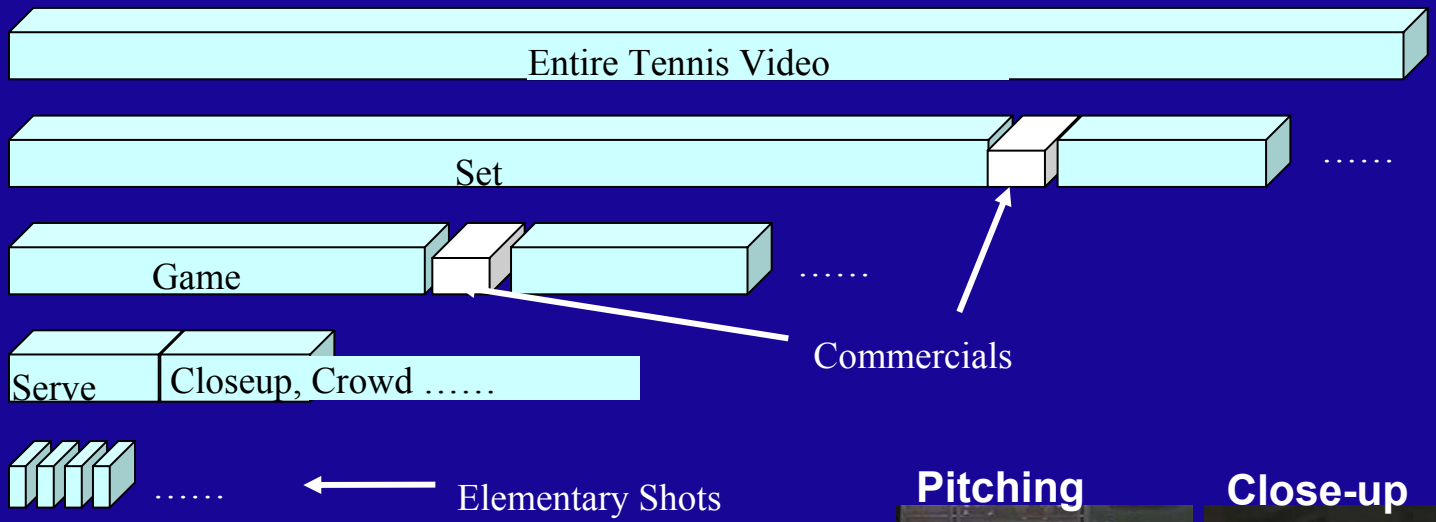
Understand Viewer Psychology

Outline

- Simple Model-Based Approach Detecting Regular Structures in Sports
- HMM Approaches Detecting Irregular Structures
- Application in Adaptive Streaming
- Generating Video Skims from Syntactic Structures and Utility Model
- Similar Structures in Medical Video

1. Detect regular structures

Regular Structure and Views in Sports Video



- **Production Syntax:**
  - canonical view ←→ recurrent semantic unit
  - view transition pattern ←→ types of events

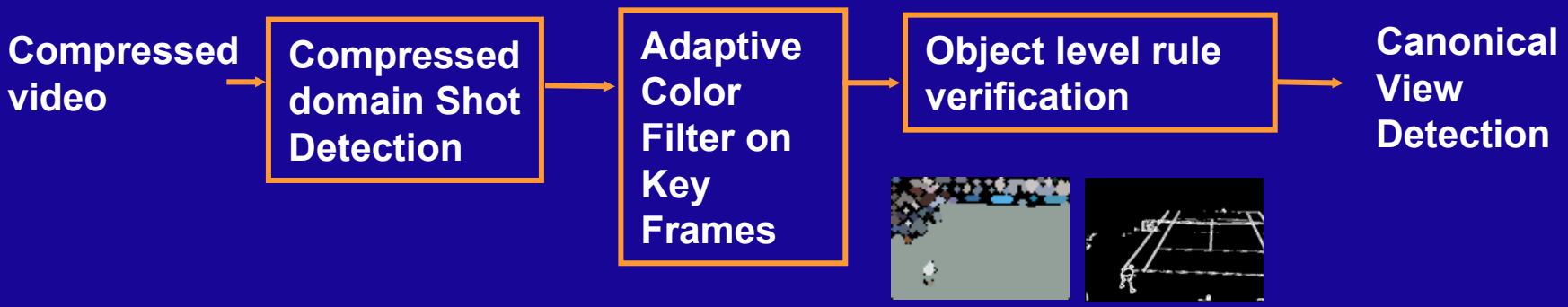
Detecting recurrent views

(Zhong & Chang '00)



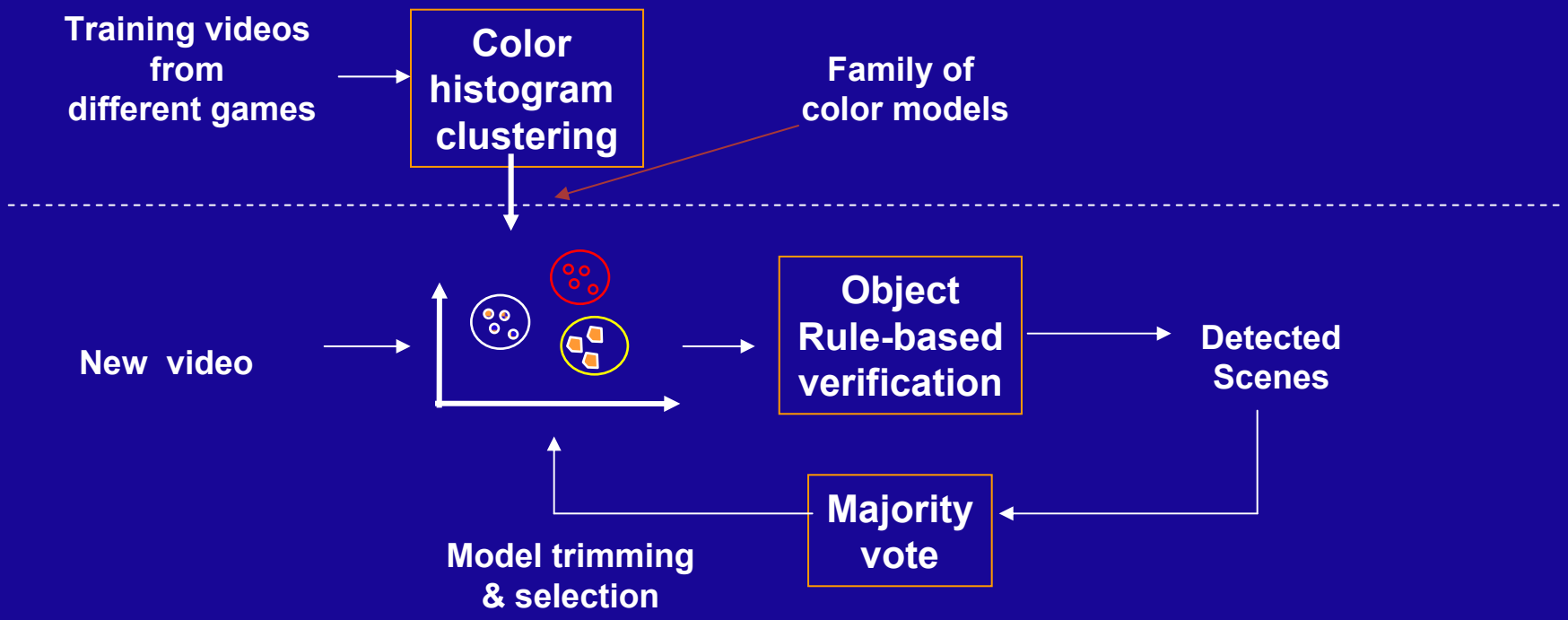



- Given known structure, supervised learning
- choose discriminative features: color, motion, object, layout
- real-time → maximize compressed domain processing
- accuracy → multi-stage coarse-to-fine verification



92%-98% detection accuracy for baseball/tennis

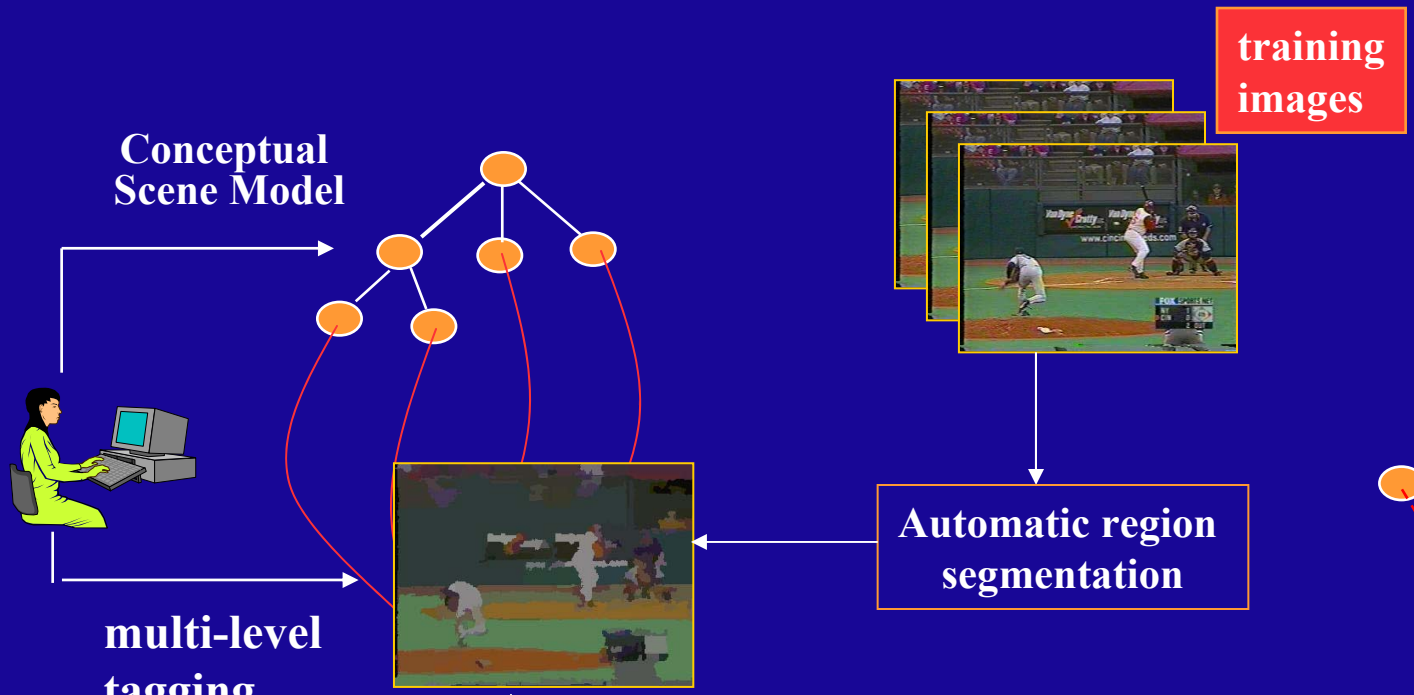
Adaptive Model Filtering



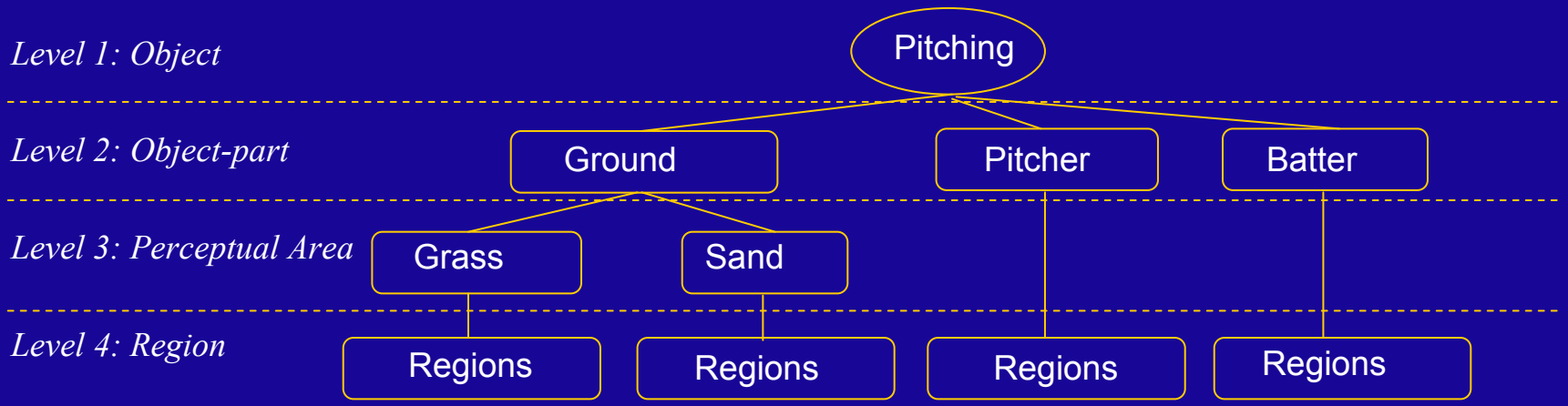
- Models are refined and selected using initial test data
- Improve robustness over source variations
- Reduce model complexity and improve speed

Learning Object-Level Rules

Approach: Multi-level Supervised Learning: *Visual Apprentice, Jaimes and Chang '99*)



Example of Hierarchical Scene Model



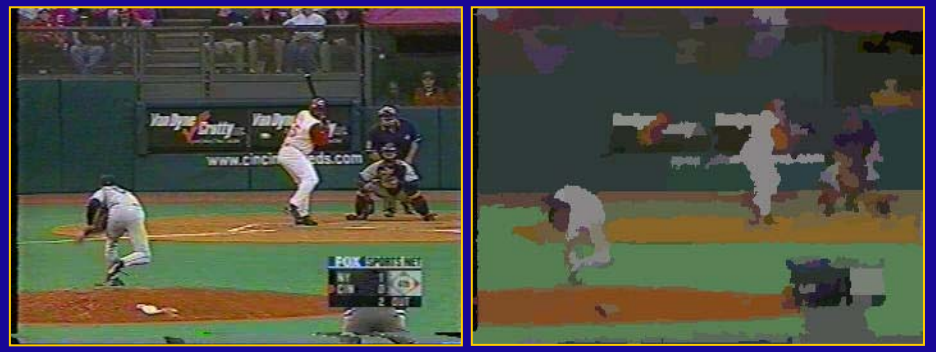
Conceptual model

↓

Interactive Tagging

↓

Multi-level rules and classifiers



2. Detect irregular structures

Detecting Irregular Temporal Pattern

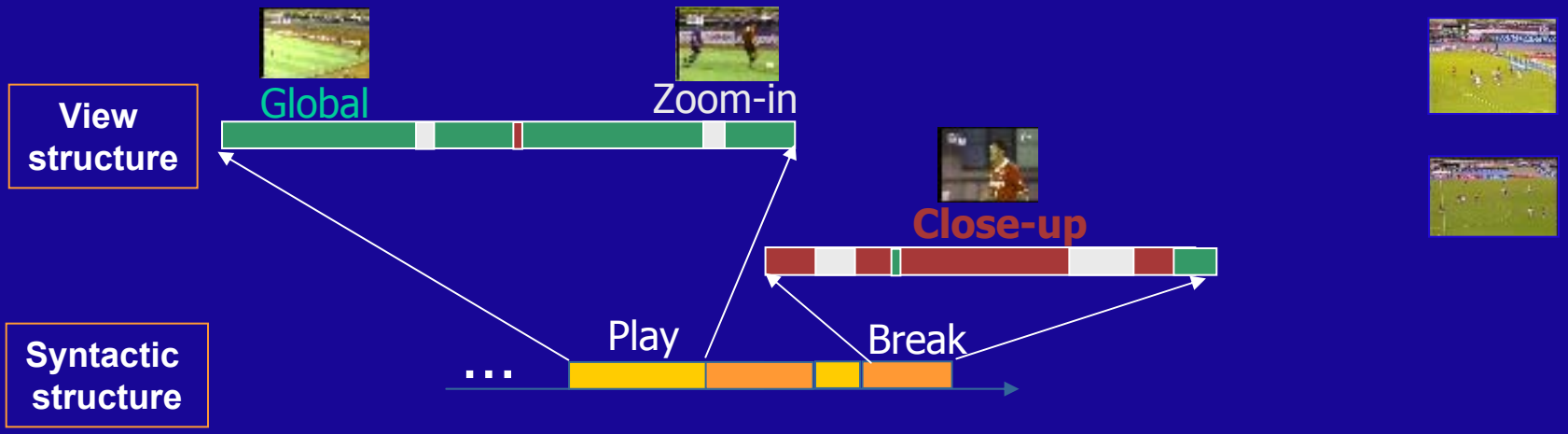
(Xie, Chang, Divakaran, Sun, '01, '02))

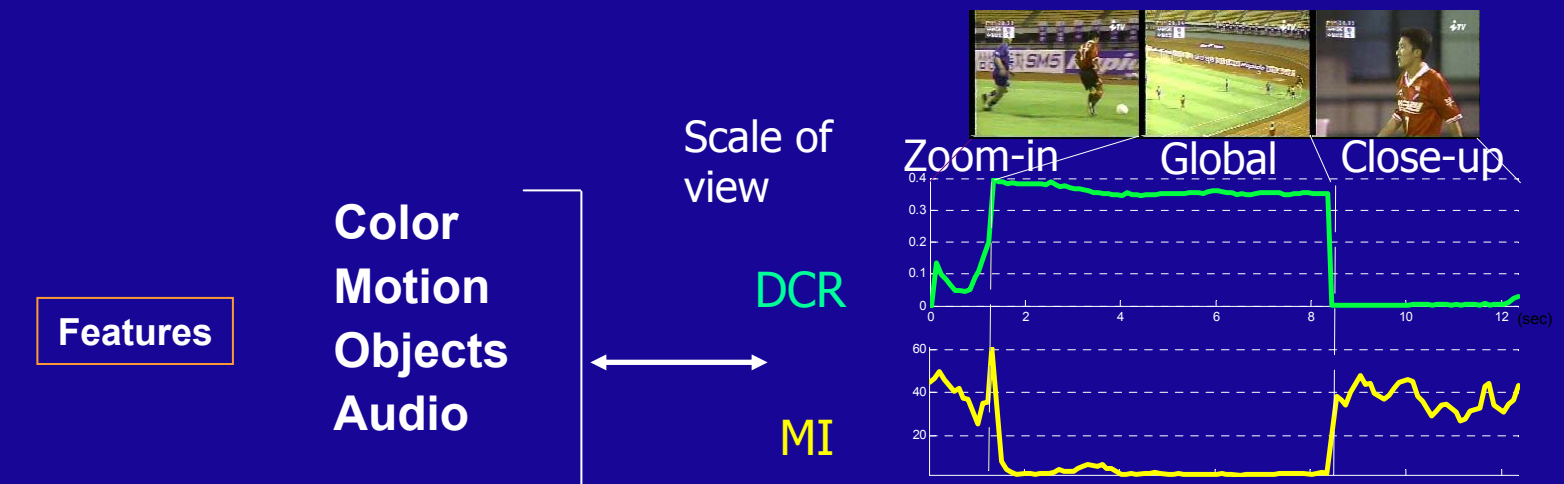
- Irregular Soccer Structure → a sequence of play and break segments
- Play/break → No canonical views/events
- Knowing structures help detection of sporadic events (goal, shoot etc)
- Similar domains: surveillance, meeting

Production syntax?
- camera views
- transition rules
- features

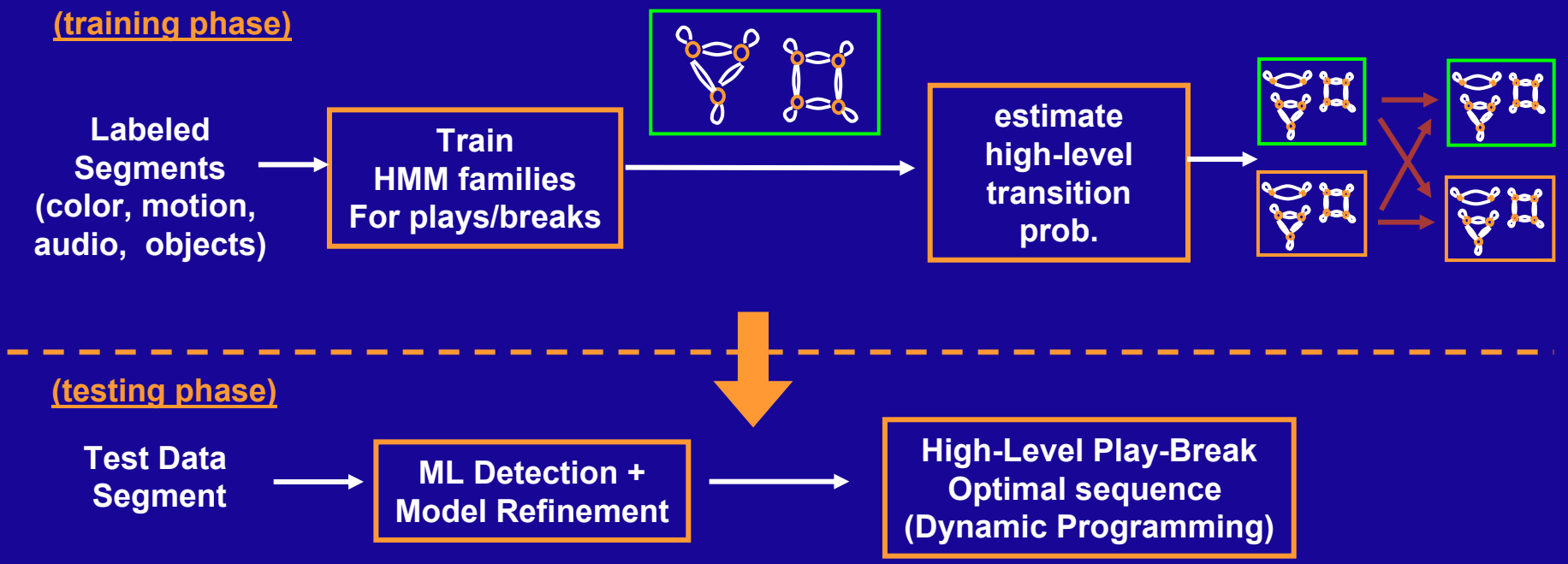
global zoom-in close-up

Production Syntax: Structures and Features

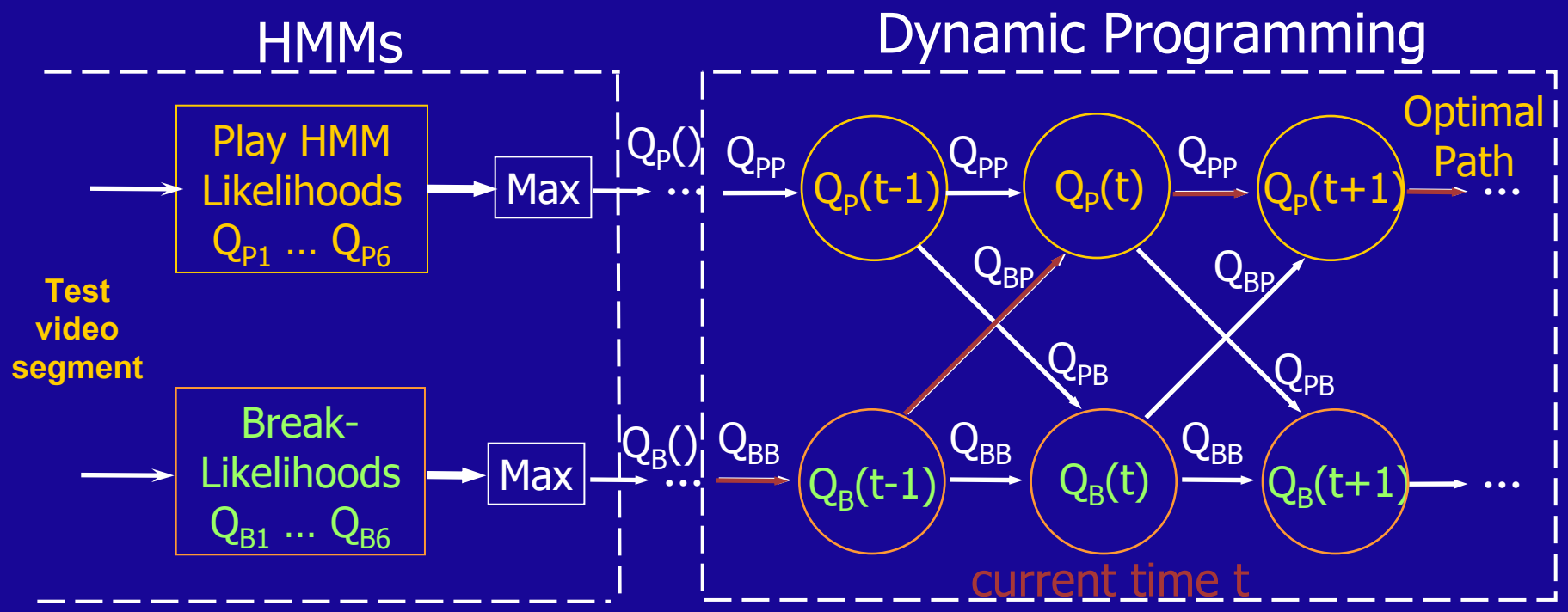




Statistic Models of Temporal Transition Patterns



Temporal Path Optimization



DP recursion:

$$\sigma_P(t) = (1-\lambda)Q_P(t) + \max\{\lambda Q_{PP} + \sigma_P(t-1),\ \lambda Q_{BP} + \sigma_B(t-1)\}$$
$$\sigma_B(t) = (1-\lambda)Q_B(t) + \max\{\lambda Q_{PB} + \sigma_P(t-1),\ \lambda Q_{BB} + \sigma_B(t-1)\}$$

DP result : accuracy +2.2%, t-test stat. confidence 99.5%

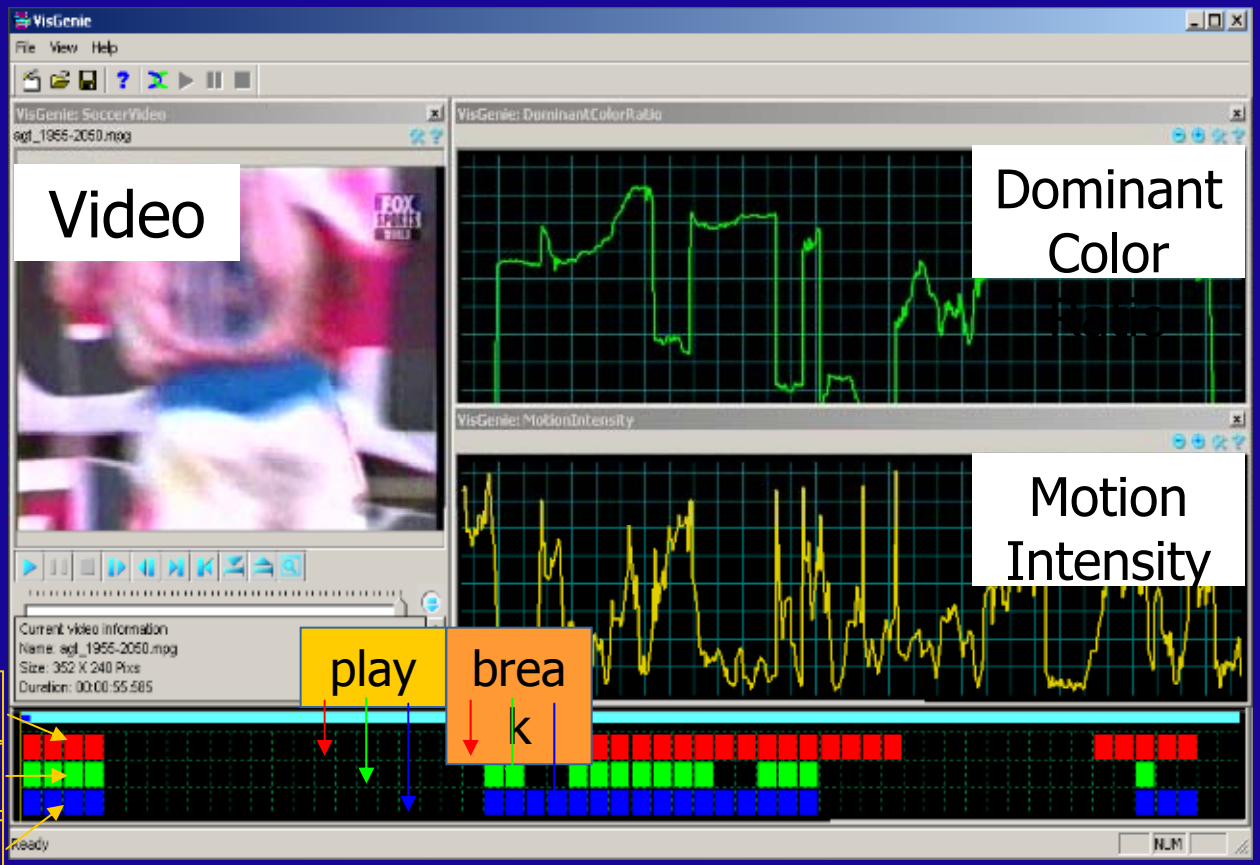
Soccer Structure Decoding Experiment

- 4 test clips, 15~25 min., various source
- Cross Validation

Test set	Training Set			
	Argentina	KoreaA	KoreaB	Espana
Argentina	87.2%	82.5%	82.5%	80.6%
KoreaA	78.1%	84.3%	84.3%	79.8%
KoreaB	79.9%	85.3%	85.3%	89.6%
Espana	79.9%	89.6%	89.6%	81.7%

- Avg. Play-Break Classification Rate: 83.5%
- Boundary timing accuracy: 62% within 3 seconds

Demo: Visualization using *VisGenie*



Ground Truth

HMM

HMM+DP

Current Work:
Unsupervised Discovery and Inferencing of Structure w/o predefined interest

3. Application in Adaptive Streaming

Application of Structure Discovery/Detection: Live Sports Video Filtering

Interactive Video

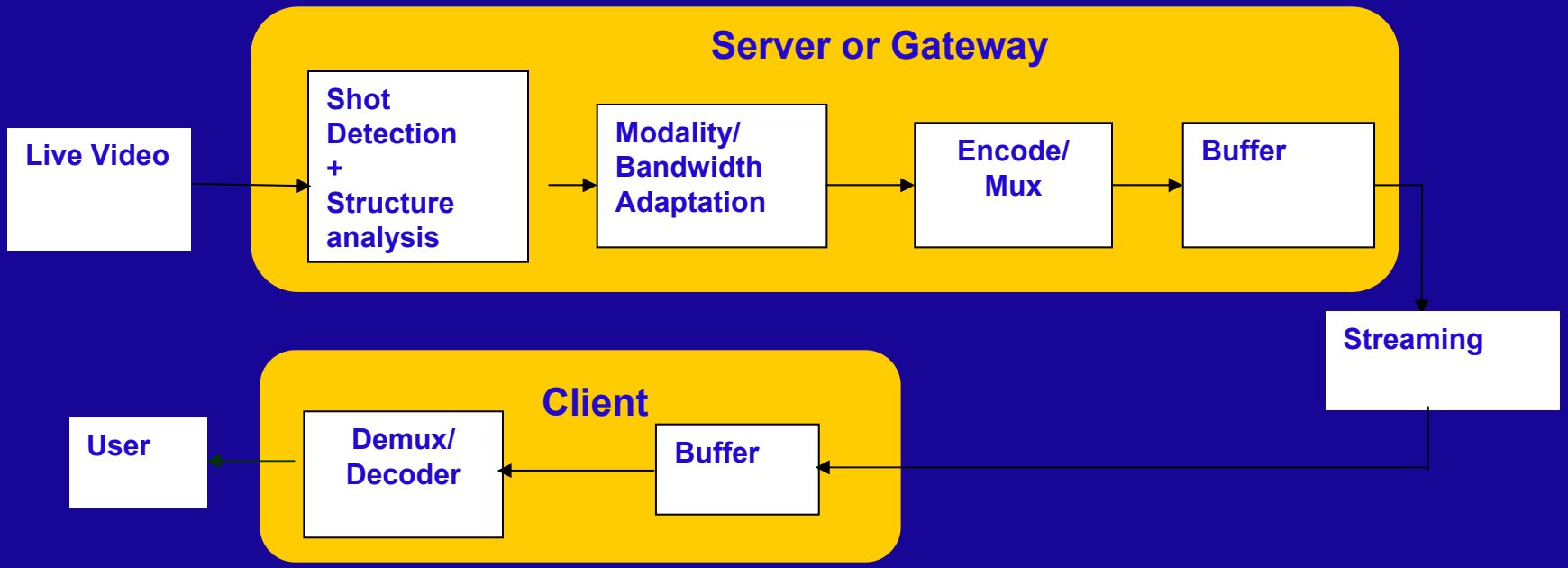
Video highlight streaming

- Temporal structure and production rules
- Massive production and audience
- Time sensitive and compressibility
- Synergy with universal media access/streaming

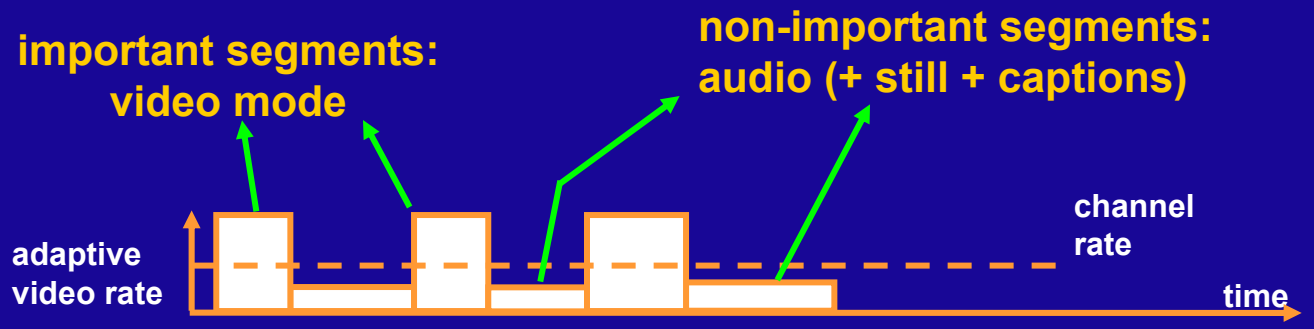
Application: Content-Adaptive Streaming

(Chang, Zhong, Kumar, 2000))

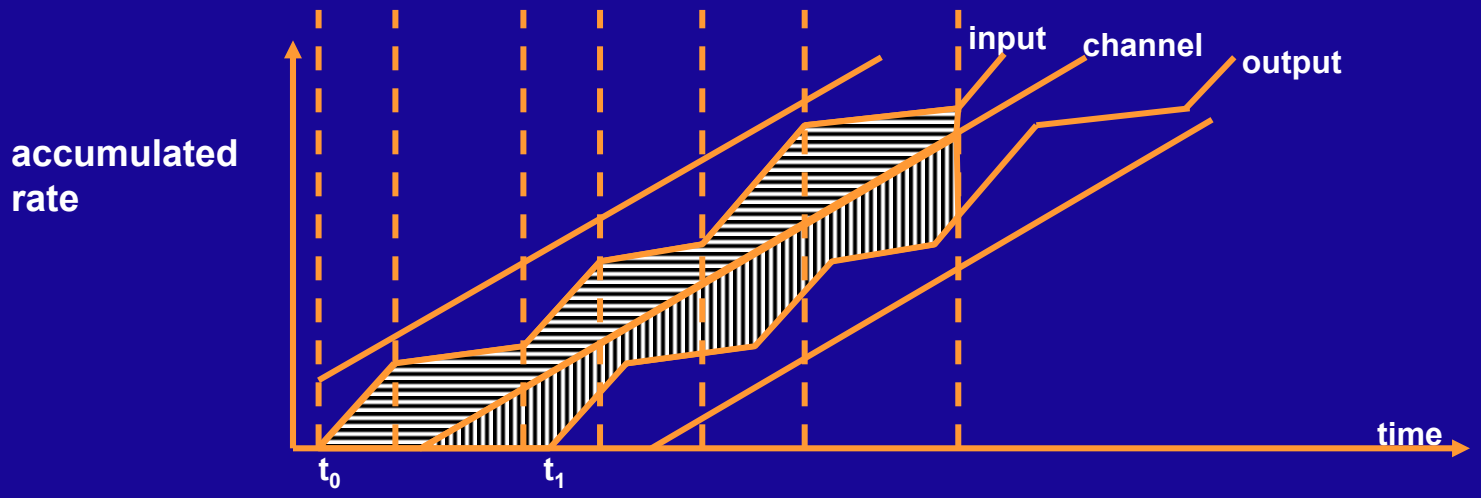
Live bandwidth adaptation based on content



Content Adaptive Streaming



Use buffering and startup delay to smooth traffic for constant rate channels

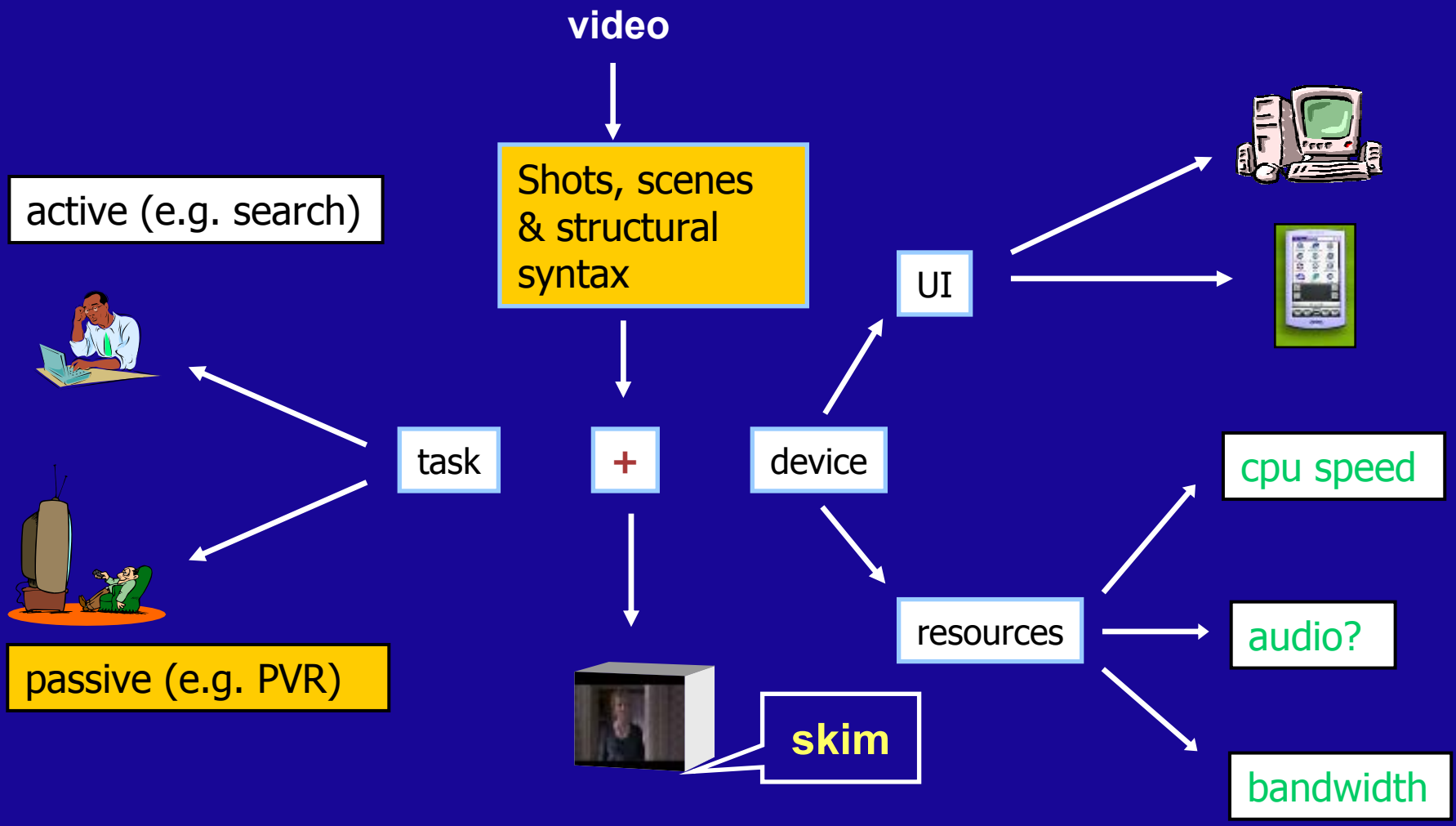


Demo

4. Generating Video Skims from Syntactic Structures and Utility Model

A case for passive skims

(Sundaram, Chang, '01 '02]



Video Skimming

Film:

Original-1 ⇢ 10:3 time reduction ⇢ Skim-1

Original-2 ⇢ 5:1 time reduction ⇢ Skim-2

News:

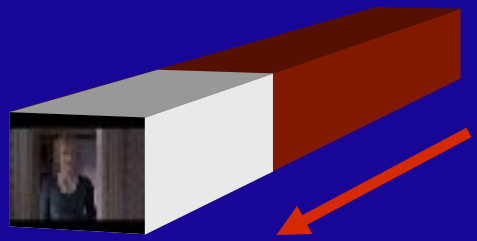
Original ⇢ 5:1 time reduction ⇢ Skim

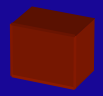
Video Skim Generation

Skim: Drastically condensed audio-video clips

News story
100 sec → 16 sec

Action scene
190 sec → 38 sec
proportional





dropped frames

1. What's the right level of entity for manipulation – shot, syntax, scene?
2. Possible operations: shot dropping and trimming.
3. How will skimming affect audio-video relations?
4. How is the "quality" affected? Aesthetic affects, information comprehension

→ **Utility Framework**

→ utility framework to model relation between operations and user comprehension
→ optimal skim generation

Modeling Utility of Shots

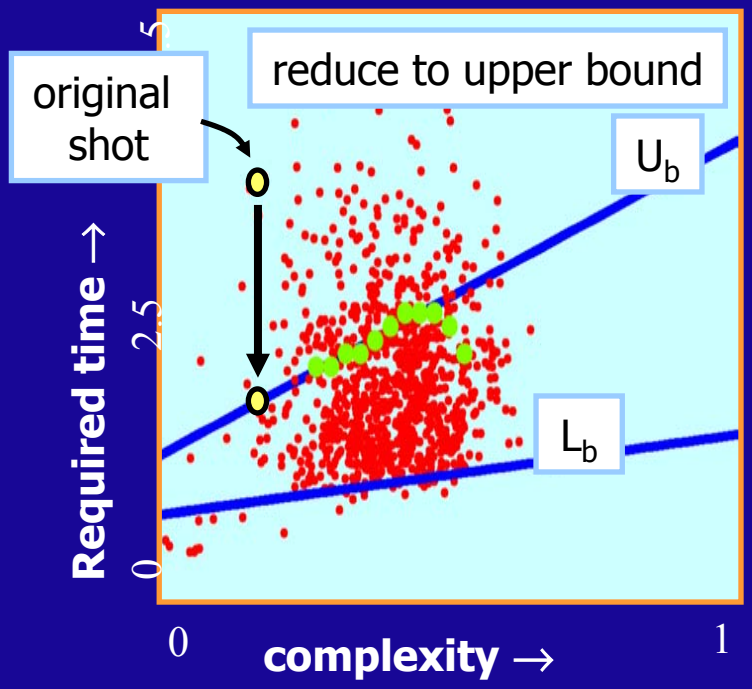
- **Viewer Psychology Principle: Explore Viewer Perceptual Model**

(a)

(b)

- **How much time is required for generic comprehension (who, what, where, when)?**
- **Not *why* → unknown in passive tasks and uncomputable**
- **Is comprehension time related to the visual spatio-temporal complexity of the content ?**

Estimate Utility Function from Subjective Study



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

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

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

Shot utility function

$$S(t,c) = \beta c(1-c) \bullet (1-\exp(-\alpha t))$$

t: duration, c: complexity

Utility of shot sequence

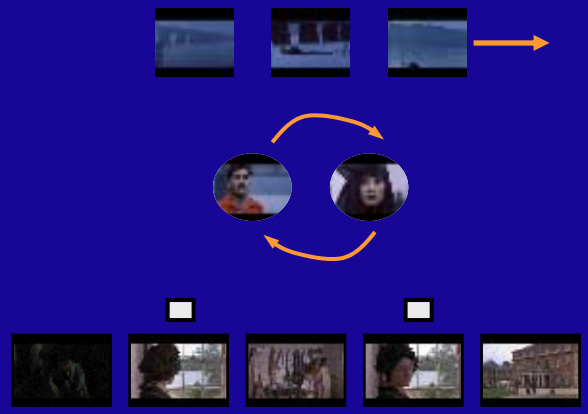
$$U(\vec{t},\vec{c},\phi) = \frac{1}{N_\phi} \sum_{i:\phi(i)=1} S(t_i,c_i)$$

$\phi(i)$: selection indicator

Film syntax is important

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 (e.g., dialog, point of view, closeup-long-closeup)
- The specific duration of the shots, to direct viewer attention (e.g., long-short-short …)
- Changing the scale of the shots

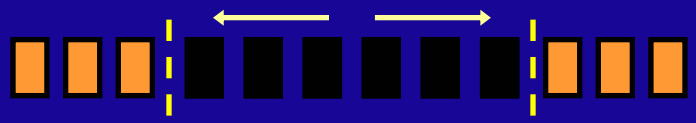


Film makers think in terms of syntax of shots and not individual shots.

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.

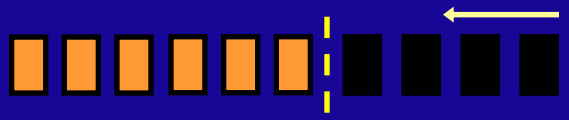


Maximal shot removal:
eliminate all the dark shots.

Structure (dialog)

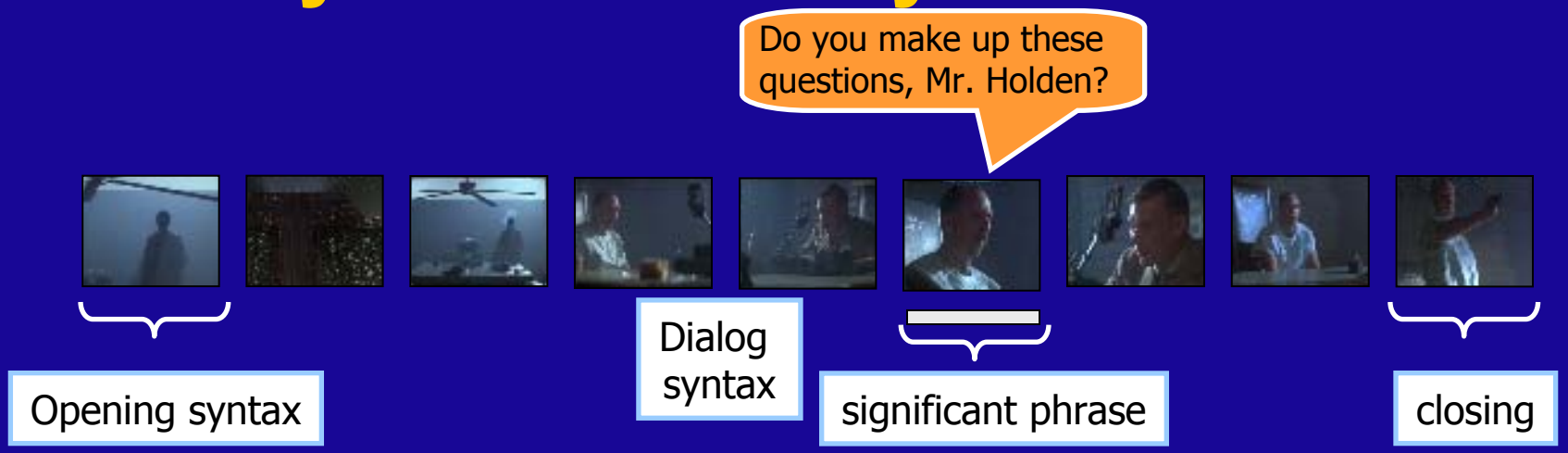
"Depicting a conversation between m people requires 3m shots." [sharff 82].

Hence, a dialog must at least have six shots



Maximal adaptation:
eliminate all the dark shots.

A-V Synchronous Syntax



- Synchronous segments:
  - Include all significant speech phrases, opening and ending syntactic segments
  - Audio and video boundaries are fully synchronized
  - Not condensed or de-synchronized
  - Such tied segments allow viewers to "*catch up*" when viewing skims
- Untied segments:
  - Audio-video can be dropped, condensed, reduced
  - Audio-video segments do not have to synchronize

The Constrained Search Problem

$$(\vec{t}_a^{\,*}, \vec{t}_v^{\,*}, \vec{\xi}^{\,*}, n_c^{\,*}) = \arg\min_{\vec{t}_a, \vec{t}_v, \vec{\xi}, n_c} O_f(\vec{t}_a, \vec{t}_v, \vec{\xi}, n_c)$$

subject to:

duration constraints

$$\begin{cases} t_{L_b,i,v} \le t_{i,v} \le t_{o,i,v}, & i : \phi_v(i) = 1, \\ T(k_i) \le t_{i,a} \le t_{o,i,a}, & N_{\phi,v} \ge N_{\min}, \end{cases}$$

target time constraints

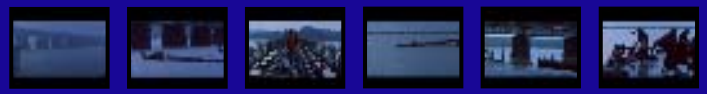
$$\begin{cases} \sum_{i:\phi_v(i)=1} t_{i,v} = T_f, \\ \sum_j t_{i,a} + \xi_i = T_f, \end{cases}$$

Multimedia tie constraints

$$\left\{ \sum_{i=1}^{N_{l,v}} t_{v,i} = \sum_{j=1}^{N_{l,a}} t_{a,j} + \xi_j, \qquad l : 1 \ldots n_c \right.$$

Skim generation framework

Entity analysis:
video shot/
syntax detection /
auditory analysis



video utility model

audio utility model

objective function

audio / video duration constraints

target skim time

tie constraints

visual syntax constraints

constraints

iterative maximization

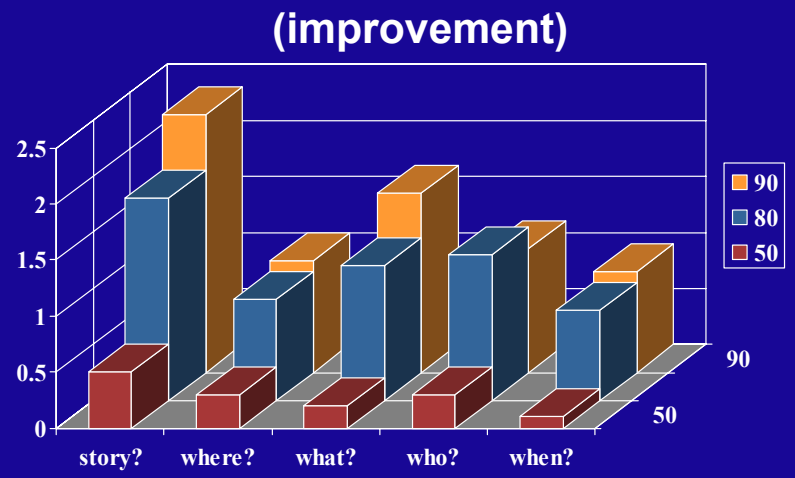
optimal

proportional

skim generation

Experimental results

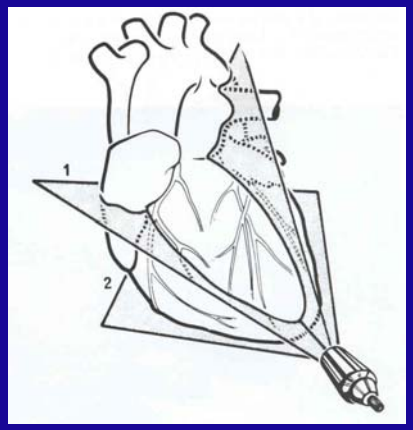
- a pilot user study to validate the utility maximization approach:
  - 12 users
  - three skim generation mechanisms
  - three compression rates (90%, 80%, 50%)
- The user study indicates:
  - the optimal skim, has a superior raw score, in all cases.
  - the optimal skim is perceptually superior, in a statistically significant sense, at the high rates.
  - The lack of a statistically significant improvement at the 50% rate, is due to the rhythm entity.

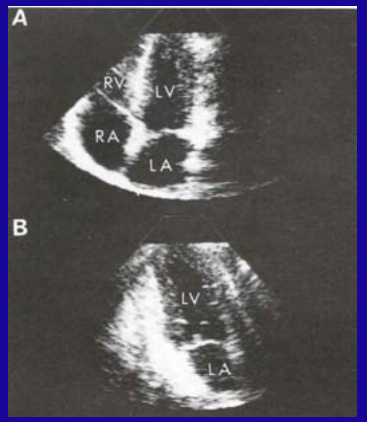


5. Medical Video Structuring

Syntactic Analysis of Echocardiogram Video

(Ebadollahi, Chang, & Wu '01 '02]

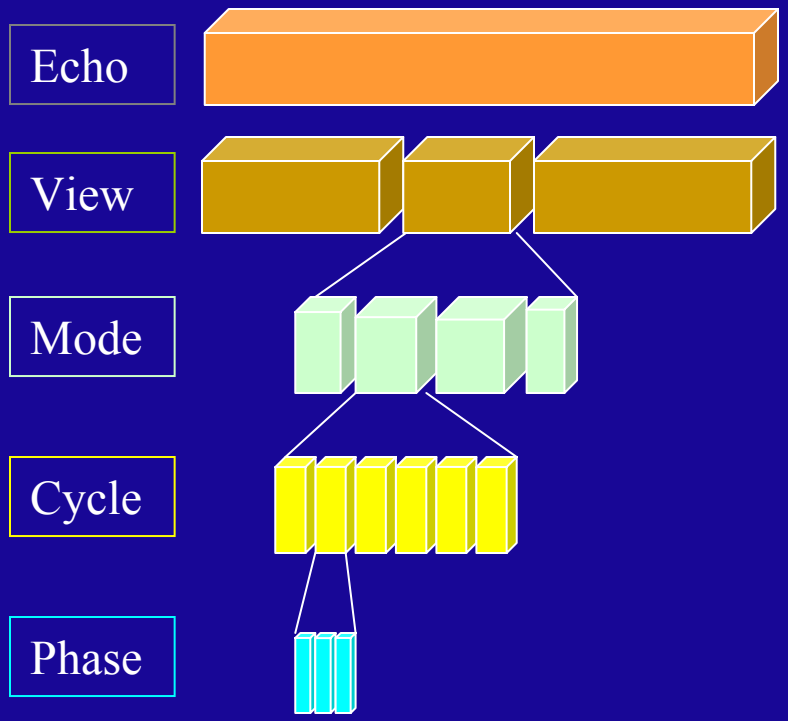




(@1994 from Echocardiography by Harvey Feigenbaum. Reproduced by permission of Lippincot Williams & Wilkins, Inc.)

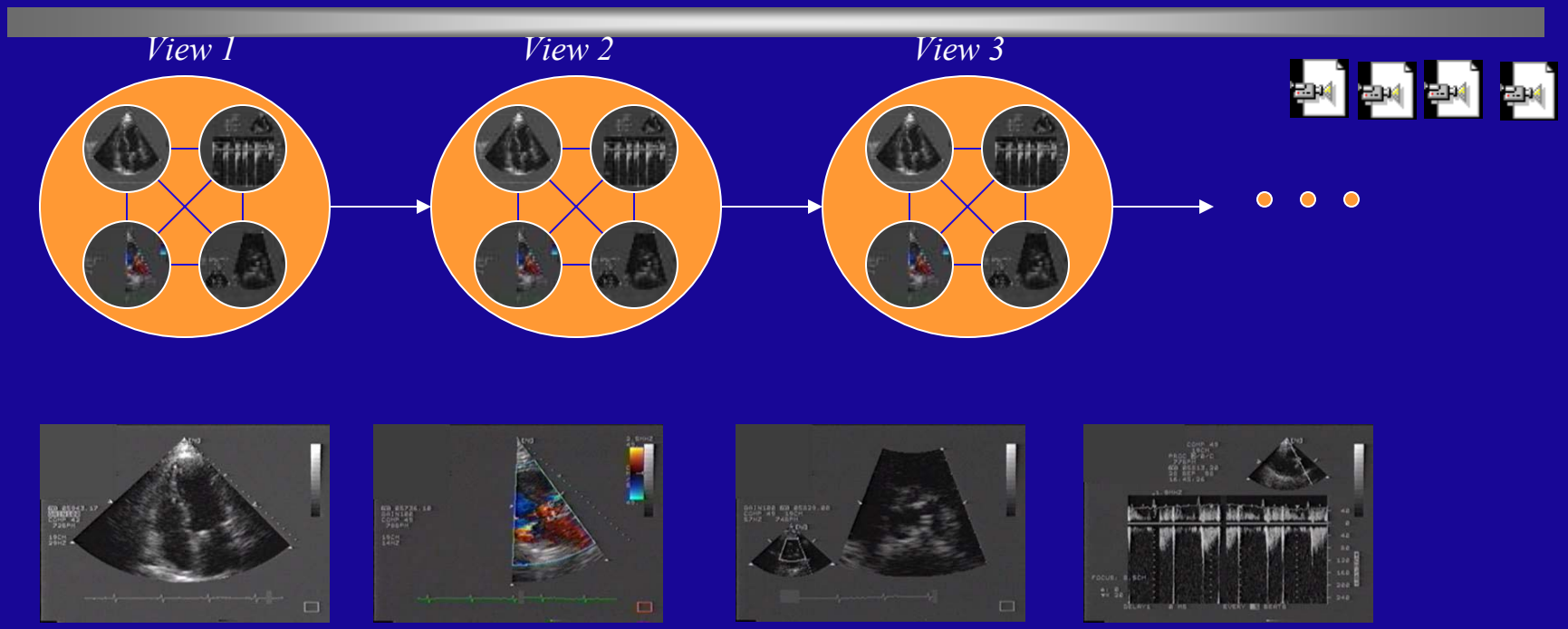
Important syntactic structures exist

Understanding Echo Video Temporal Structure



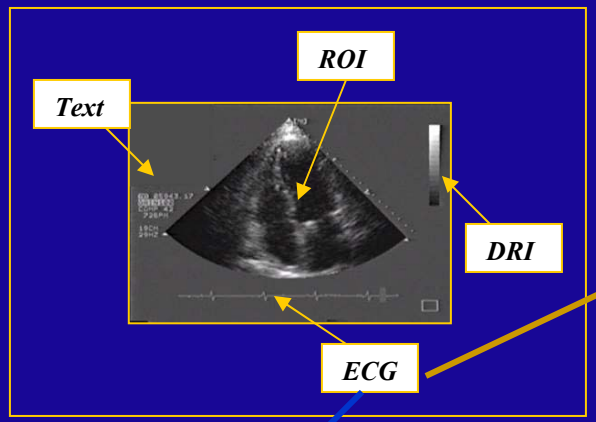
- Echo video consists of a hierarchical temporal structure.
- A *view* corresponds to a transducer position and a scanning section of the heart.
- *Modes* are used to study different features and scales.
  - Regular 2D: spatio-temporal structure of the corresponding cardiac section
  - Color Doppler: overview of pattern, direction, and velocity of the blood flow
  - Zoom-in: details of region of interest
  - Doppler: measure blood flow along specific scan lines
- A *cycle* corresponds to a complete heart beat, in which heart undergoes different *phases*.

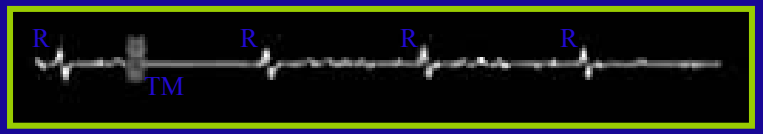
Echo Video Production Syntax: Temporal

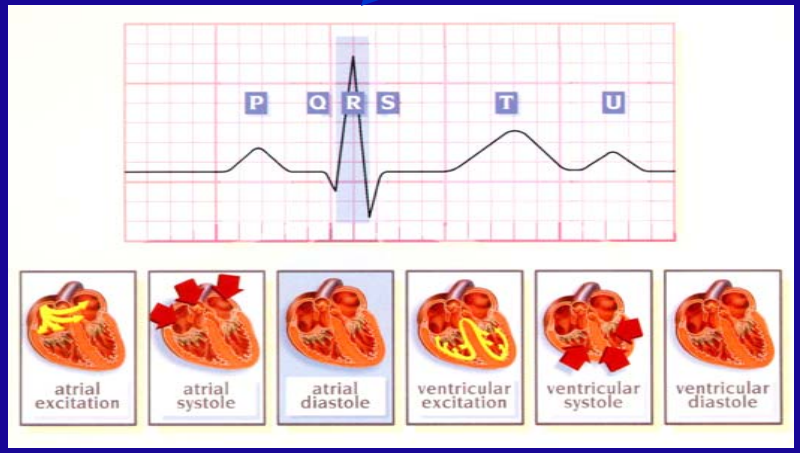


- Deterministic view transitions (*ACC*)
- Probabilistic mode transitions → separating consecutive 2D views is hard!

Echo Video Production Syntax: Spatial

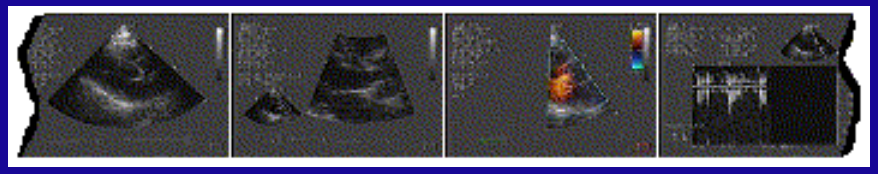






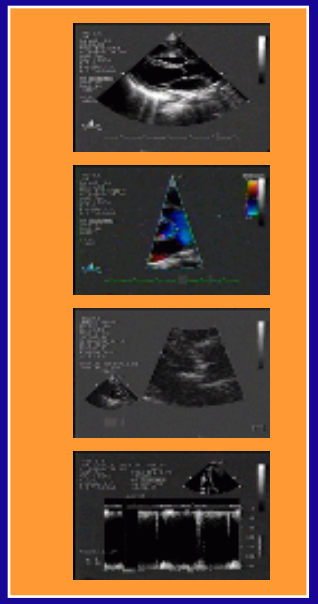
Useful information contained in window shape, objects, color, motion, and ECG images

Echo Video Digital Library & Remote Medicine



Echo Video

Structure Parsing



Domain Knowledge

KeyFrame, Clinic summary

View Recognizer

Diagnosis Reports

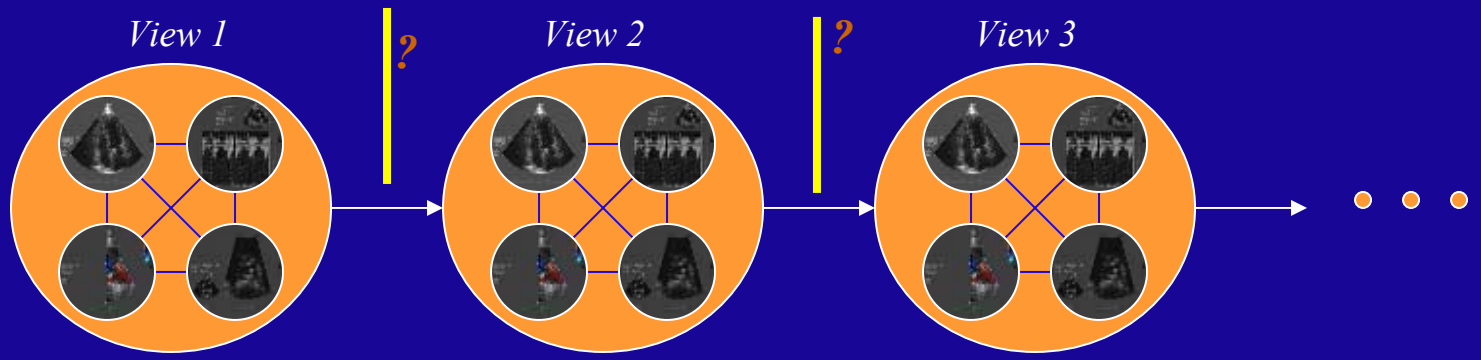
View/Object Annotation

Adaptive network delivery

Content search

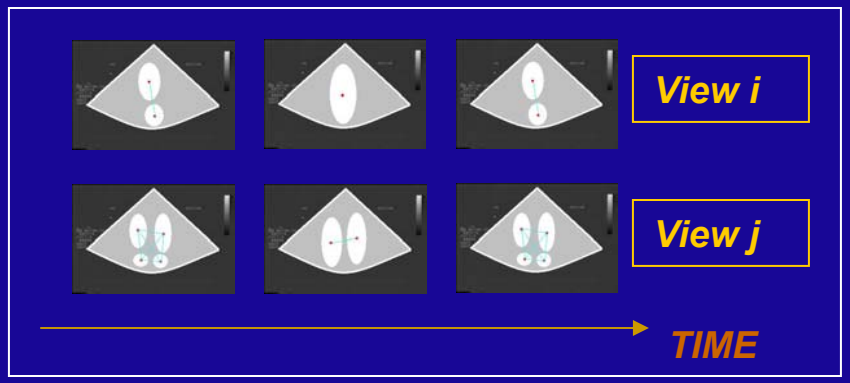
(Ebadollahi and Chang '00)

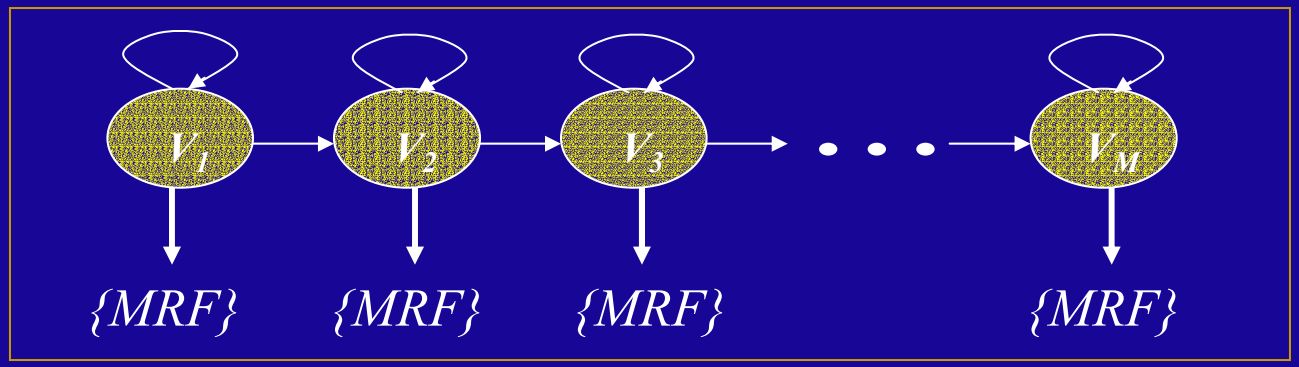
Challenge: Statistic View Recognition



Observation of Production Syntax:

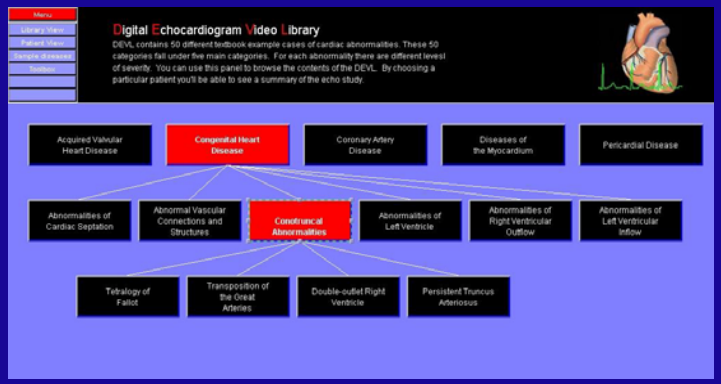
At each view angle one can see a unique spatio-temporal configuration of the cardiac objects.





DEVL Medical Echo Library Interfaces

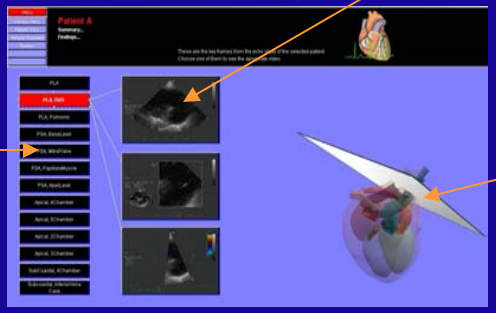
Disease Taxonomy Interface



Representative frames of modes under selected view

View Browsing Interface

Table of Contents showing list of views



3D model showing transducer angle

Automatic tools of video indexing, empowered with new visualization paradigm, greatly enhance physicians' capability of accessing and comprehending medical video.

3D Heart Model courtesy of New York University School of Medicine

Conclusions

- Syntactic structures convey semantics in specific domains
  - Recall content chain principles
  - Sports, film, medical
- Syntactic structure discovery poses an interesting research problem
  - Supervised vs. unsupervised
  - Different promising statistical inferenceing models
- Applications:
  - Real-time content-adaptive video streaming
  - Syntax-preserving content re-generation, e.g., skim
  - Browsing, augmented visualization

Acknowledgements

- **Content-adaptive delivery:**
  Di Zhong, Raj Kumar, Jaegon Kim, Yong Wang
- **VOCR:**
  DongQing Zhang
- **Statistical Video Mining:**
  Lexing Xie, Peng Xu, Ajay Divakaran, Huifang Sun
- **Syntax Preserving Video Skimming:**
  Hari Sundaram
- **Medical Video:**
  Shahram Ebadollahi, Henry Wu

3D Heart Model courtesy of New York University School of Medicine

More Information

- Columbia DVMM Lab
  http://www.ee.columbia.edu/dvmm
- PI: Prof. Shih-Fu Chang
  http://www.ee.columbia.edu/~sfchang
  sfchang@ee.columbia.edu
- Publications
  http://www.ee.columbia.edu/dvmm/publications.htm

Related papers

- S.-F. Chang, *The Holy Grail of Content-Based Media Analysis*, IEEE Multimedia Magazine, Vol. 9, Issue 2, pp. 6-10, April-June 2002.
- S.-F. Chang, D. Zhong, and R. Kumar, *Real-Time Content-Based Adaptive Streaming of Sports Video*, Columbia University ADVENT Technical Report #121, July 2001. Also IEEE Workshop on Content-Based Access to Video/Image Library, Hawaii, Dec. 2001.
- R. Kumar Rajendran, M. van der Schaar, S.-F. Chang, *FGS+: Optimizing the Joint Spatio-Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding*, IEEE International Symposium on Circuits and Systems (ISCAS 2002), Phoenix, Arizona, May 2002.
- L. Xie, S.-F. Chang, A. Divakaran and H. Sun, *Structure Analysis of Soccer Video with Hidden Markov Models*, Proc. Interational Conference on Acoustic, Speech and Signal Processing, (ICASSP-2002), Orlando, FL, USA, May 13-17, 2002.
- D.-Q. Zhang, R. Kumar Rajendran, and S.-F. Chang, *General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video*, International Conference on Image Processing (ICIP-2002), Rochester, New York, USA, Sep 22-25, 2002.
- A. Jaimes and S.-F. Chang, *Learning Structured Visual Detectors From User Input at Multiple Levels*, Invited Paper, International Journal of Image and Graphics (IJIG), Special Issue on Image and Video Databases, August 2001.
- Q. Sun, S.-F. Chang, M. Kurato and M. Suto, A new semi-fragile image authentication framework combining ECC and PKI, Invited paper for Special Session on Multimedia Watermarking, ISCAS2002, Phoenix, USA, 2002.
- C.-Y. Lin and S.-F. Chang, SARI: Self-Authentication-and-Recovery Image Watermarking System, ACM Multimedia 2001, Ottawa, Canada, Sep. 30 - Oct. 5, 2001.
- H. Sundaram, S.-F. Chang, *Constrained Utility Maximization for generating Visual Skims*, IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'2001) Dec. 2001 Kauai, HI USA.
- A. Benitez and S.-F. Chang, "Semantic *knowledge Construction from Annotated Image Collections," and "Perceptual knowledge Construction from Annotated Image Collections,"* IEEE ICME 2002, August 2002, Lausanne, Switzerland.

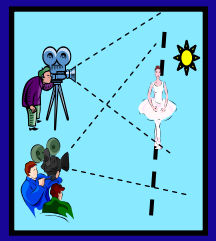
Backup Slides

Holy Grail of Content-Based Media Analysis

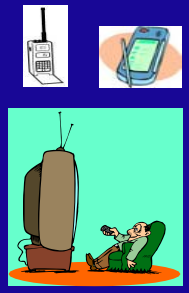
(S.-F. Chang, Multimedia, April 2002)

- Impact conditions
  - Metadata not available from production
  - Work that humans are not good at
    (e.g., feature/object computation, real-time logging)
  - Content with large volume, low individual value
  - Time sensitive
  - Well defined tasks and evaluation methods
- Promising Areas
  - Medical, sports, presentation, surveillance
  - News, meetings

Media Indexing: Challenges & Opportunities



Analysis and interaction



Production model + Media Integration + Viewer model

frameworks and methodologies

Check Application Impact Criteria

- Filtering, Browsing
- Content-Augmentation
- Streaming
- Navigation