Optimal Video Adaptation and Skimming Using a Utility-Based Framework

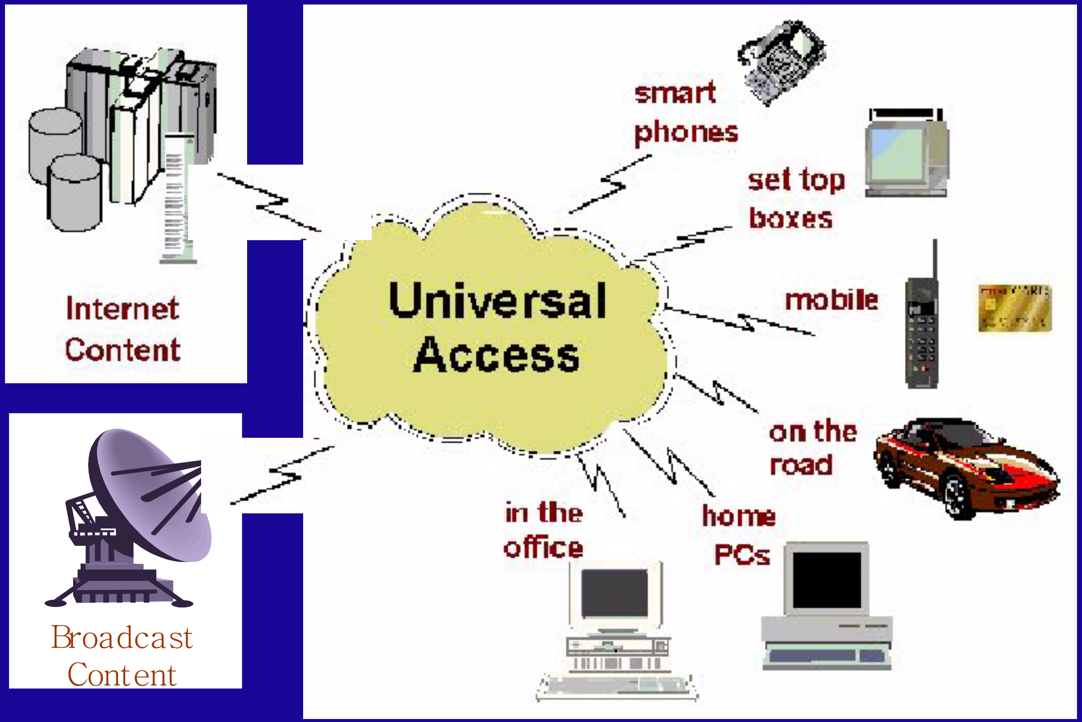
Shih-Fu Chang

Digital Video and Multimedia Lab
ADVENT University-Industry Consortium
Columbia University

Sept. 9th 2002

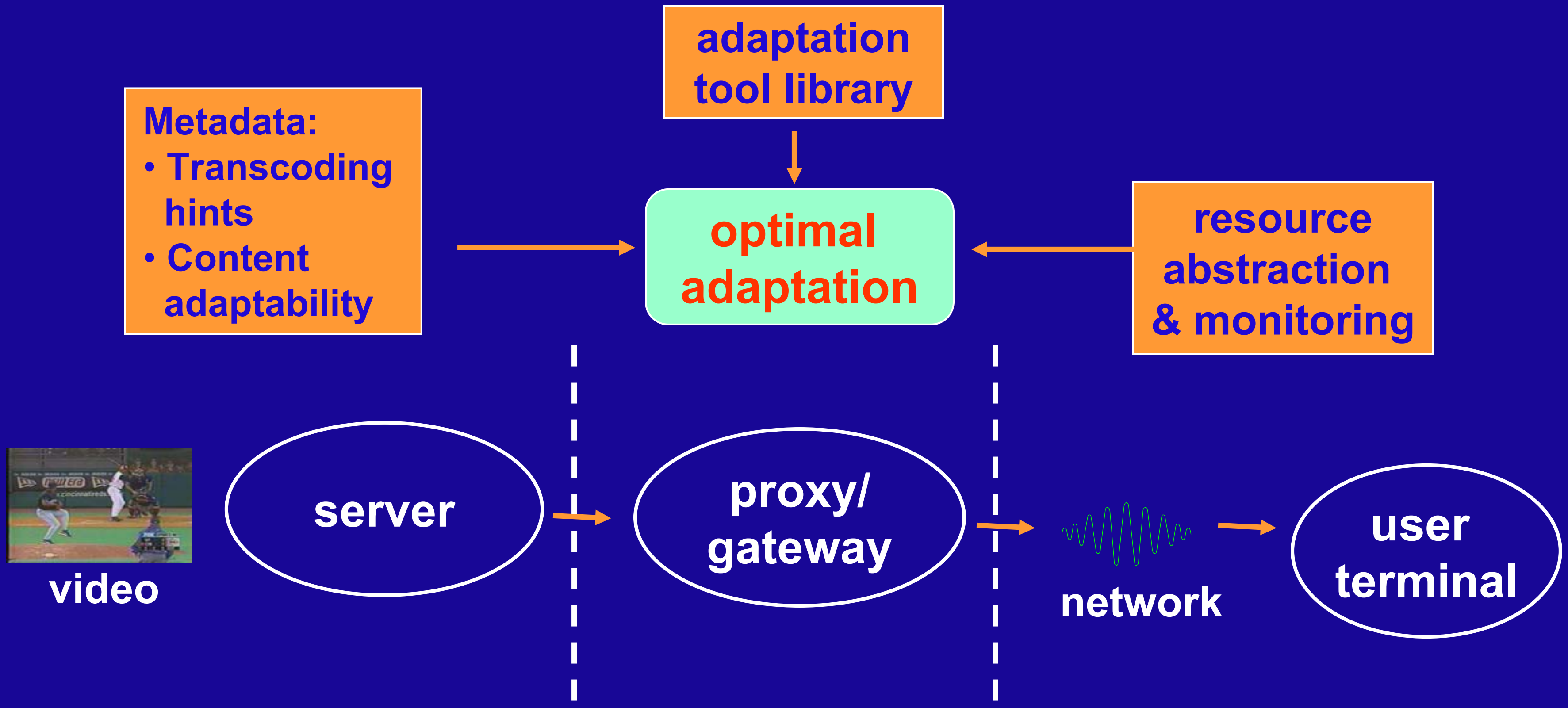
http://www.ee.columbia.edu/dvmm

The Need for Video Adaptation



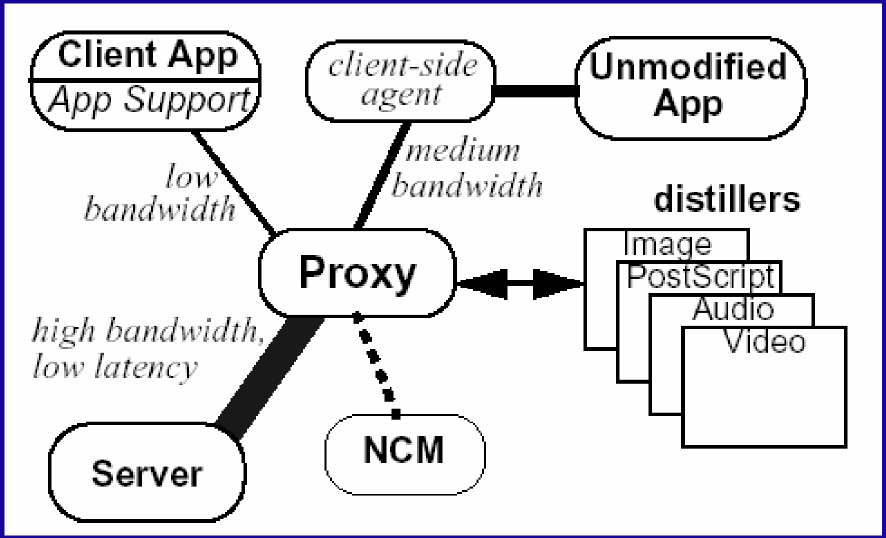
- Heterogeneous users, networks, and terminals
- Content analysis to assist video adaptation decision

Adaptation in 3-tier architecture



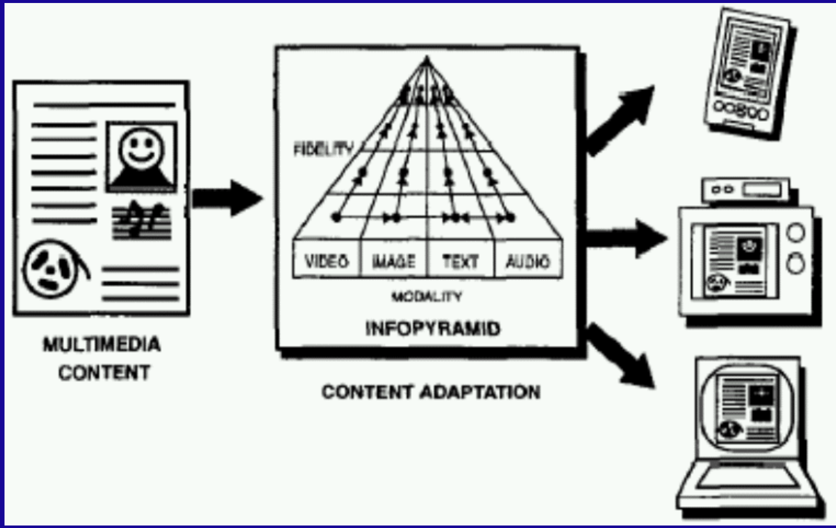
Prior works

- Active proxy dynamic distillers [Fox, Brewer, *et al* 96]
  - Datatype-specific distillation
  - intermediate proxy – low end-to-end latency
  - network/application interface separation
- Video manipulations done at the global level
  - Resolution, frame rate, color depth



Prior Works

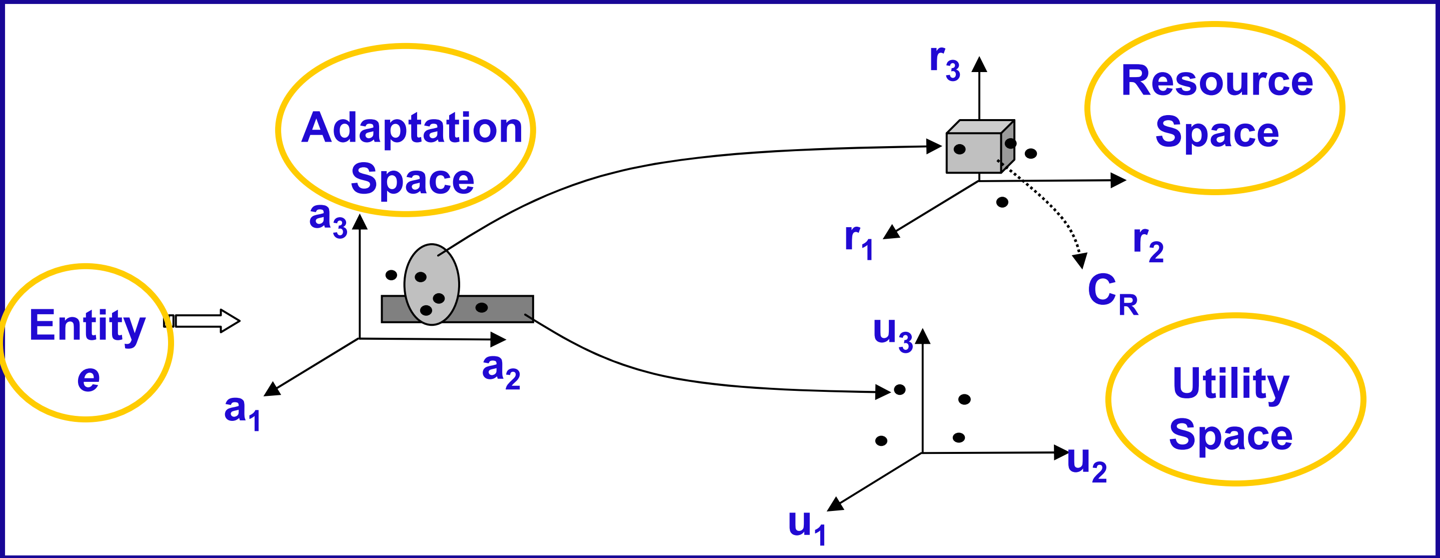
- InfoPyramid, Universal Tuner [Smith, Li, Mohand 99]
  - Content adaptation with various fidelity and modality
  - Translation and transcoding
- Transcoding focused on images
- Allowable operators are optimized at the global level
  - Resolution, bit rate, color depth, modality substitute



• Video contains rich multi-level elements and structures
• User/network conditions may change rapidly

→ Framework for Micro-level Adaptation

A Utility-Based Framework



- **Entity:**
  - **A set of video data with consistency constraints on certain attributes**
  - **Basic data unit undergoing adaptation**
  - **Examples:**
    - **Program, Shot, Scene**
    - **Frame, Object, Region**
    - **Syntactic or semantic elements, e.g., anchors, scoring segments**
    - **Synchronous multimedia entities, e.g., dialog, talking face, explosion**

Multiple Degrees of Freedom for Adaptation

- Signal level
  - change of bit rate, frame rate, resolution, color depth, SNR
- Time
  - Condensation by uniform time scaling, content-based filtering (selection/dropping)
- Modality conversion
  - Key frame shows, video posters, spatial summaries

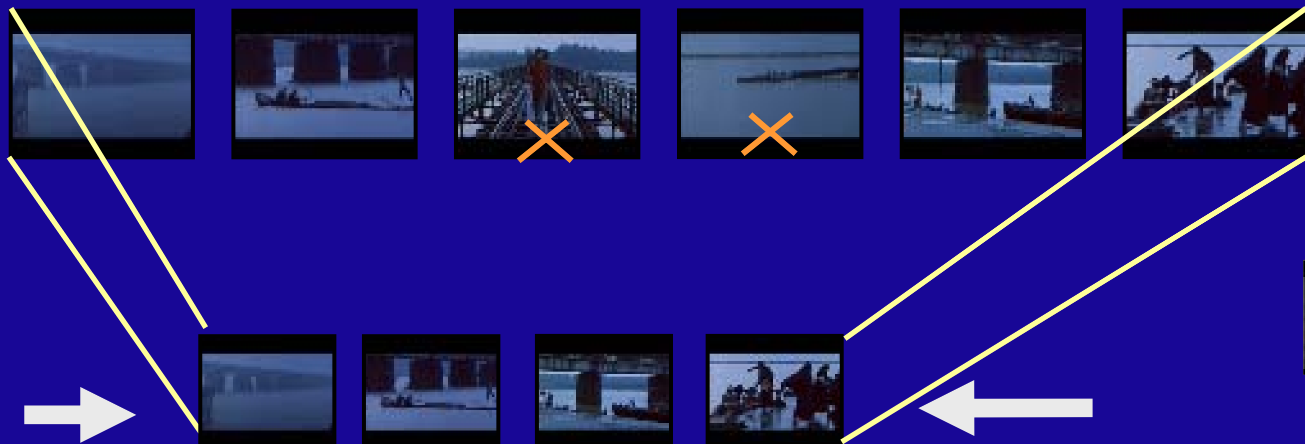
Key Issues

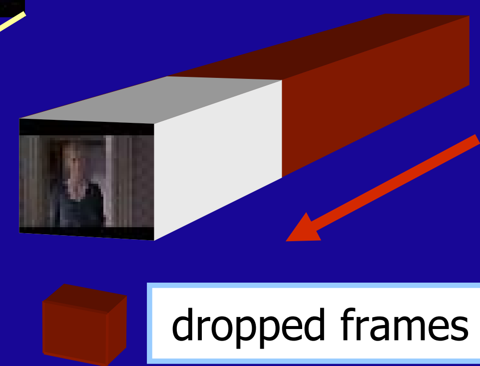
- Given an entity, what are allowable adaptations in a specific environment?
  - frames droppable from MPEG, shots removable in a scene
- Measurement and modeling of resources and utilities
- How to combine different types of resources and utilities?
  - Power, display, memory, CPU, bandwidth, and user time
  - SNR, Subjective quality and comprehension
- Description schemes of ARU spaces
  - E.g., N shots with binary selection → N(N-1) points
  - E.g., 4 frame rates, 2 resolutions → 8 adaptation points
  - How about multi-level entity, multi-dimension R/U

Case 1: Video Skim Generation

(with H. Sundaram 2001)

Original scene → condensed clip (<u>video skims</u>)





1. How much time is available? (C_R)
2. What are allowable operations? (A)
3. How will operations alter aesthetic affects, perceptual quality, and semantic understanding? (U)
4. A Resource Constrained Utility Maximization Problem

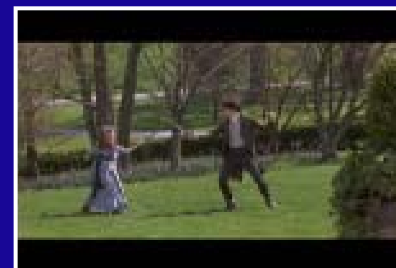
→ Generalized utility framework for video skimming

Original-1 **30% Skim**

Original-news **17% Skim**

How to Construct Computable Utility Model?

(a)



(b)

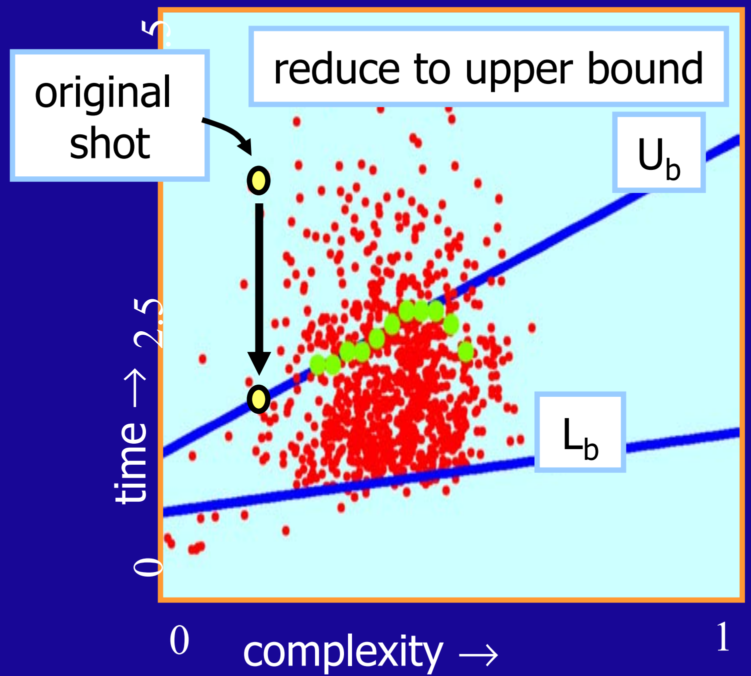
- **How much time is required for generic comprehension (who, what, where, when)?**
- **Is comprehension time related to the computable spatio-temporal complexity of the shot ?**
- **Explore Viewer Perceptual Model from Film Theory**
  - The presence of detail robs a shot of its screen time [Sharff 1982].

Measuring Comprehension Time

- Represent the shot by its key-frame
- A shot is selected at random [3600 shots]
- The subject was asked to *correctly* answer four questions in minimum time:
  - Who ?
  - What ?
  - When ?
  - Where ?

"Why" was not asked

Utility Function



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

$$U_b(c) = 2.40c + 1.11$$
$$L_b(c) = 0.61c + 0.68$$

Utility function between the bounds

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

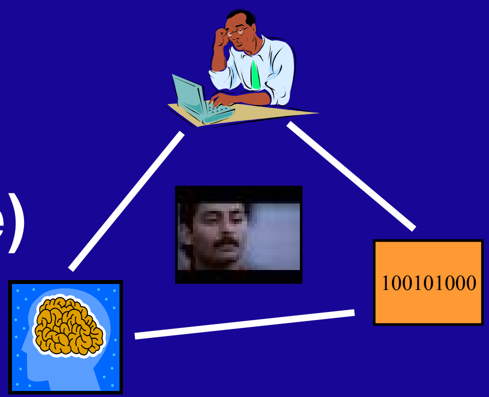
$$U(\vec{t},\vec{c},\phi) = \frac{1}{N_\phi} \sum_{i:\phi(i)=1} S(t_i, c_i)$$

t: duration, c: complexity

$\phi(i)$ **: selection indicator sequence**

Factors affecting comprehension

- The comprehension time is influenced by many factors:
  - Visual complexity
  - The viewer task (active vs. passive)
  - Prior knowledge of the viewer

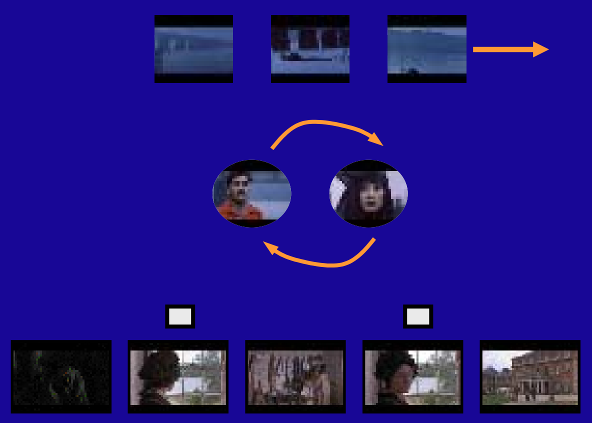


- So far, we only focus on visual complexity since it is measurable.

Rich Structure: Considering 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 (cut)
- The specific duration of the shots, to direct viewer attention
- Changing the scale of the shots

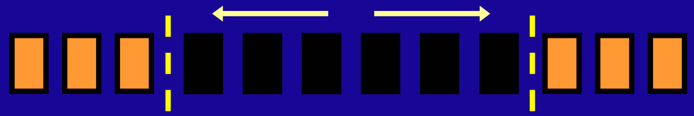


Film makers think in terms of phrases of shots and not individual shots → *choose the right entity for adaptation*

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.

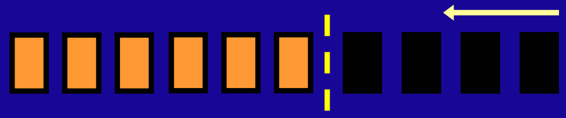


Maximal shot removal:
eliminate all the dark shots.

Structure (dialog)

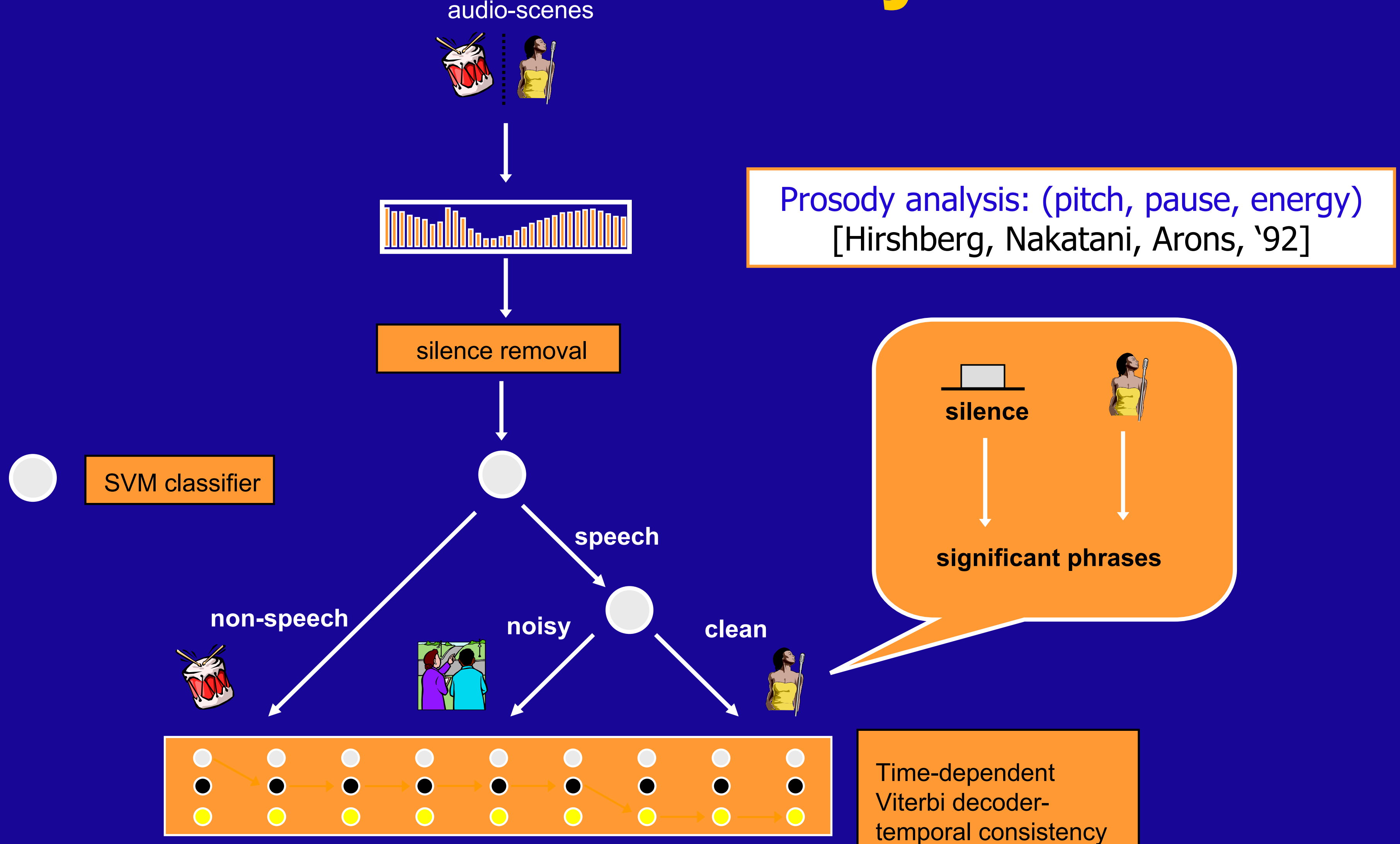
"Depicting a conversation between m people requires 3m shots." [sharff 82].

Hence, a dialog must at least have six shots

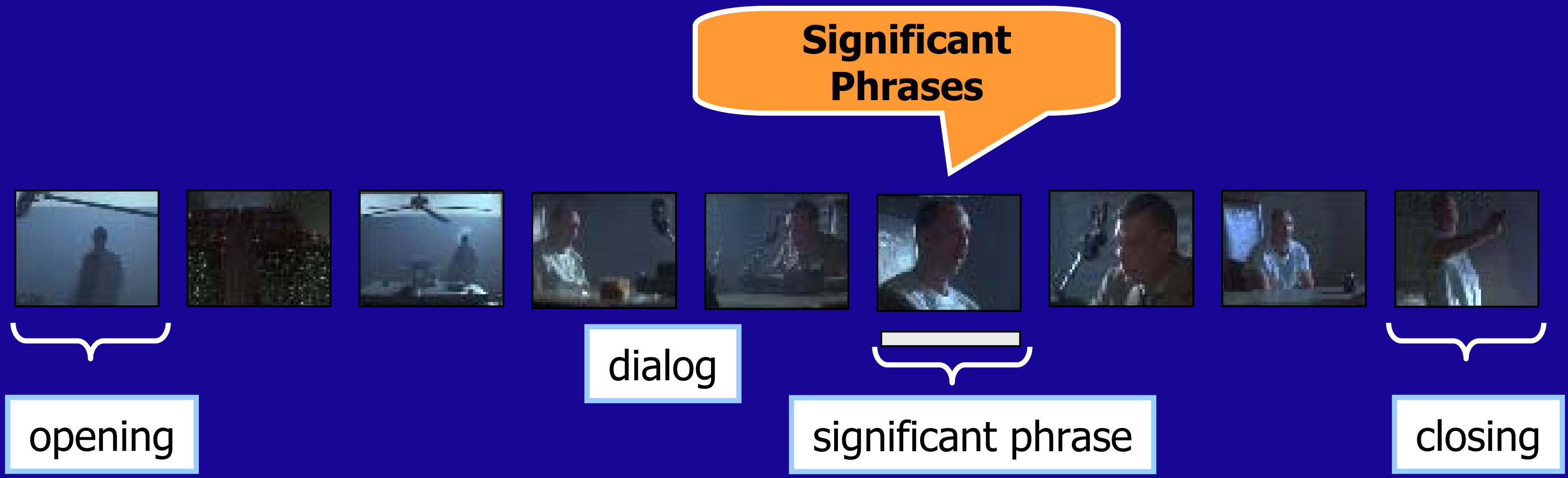


Maximal adaptation:
eliminate all the dark shots.

Audio content analysis

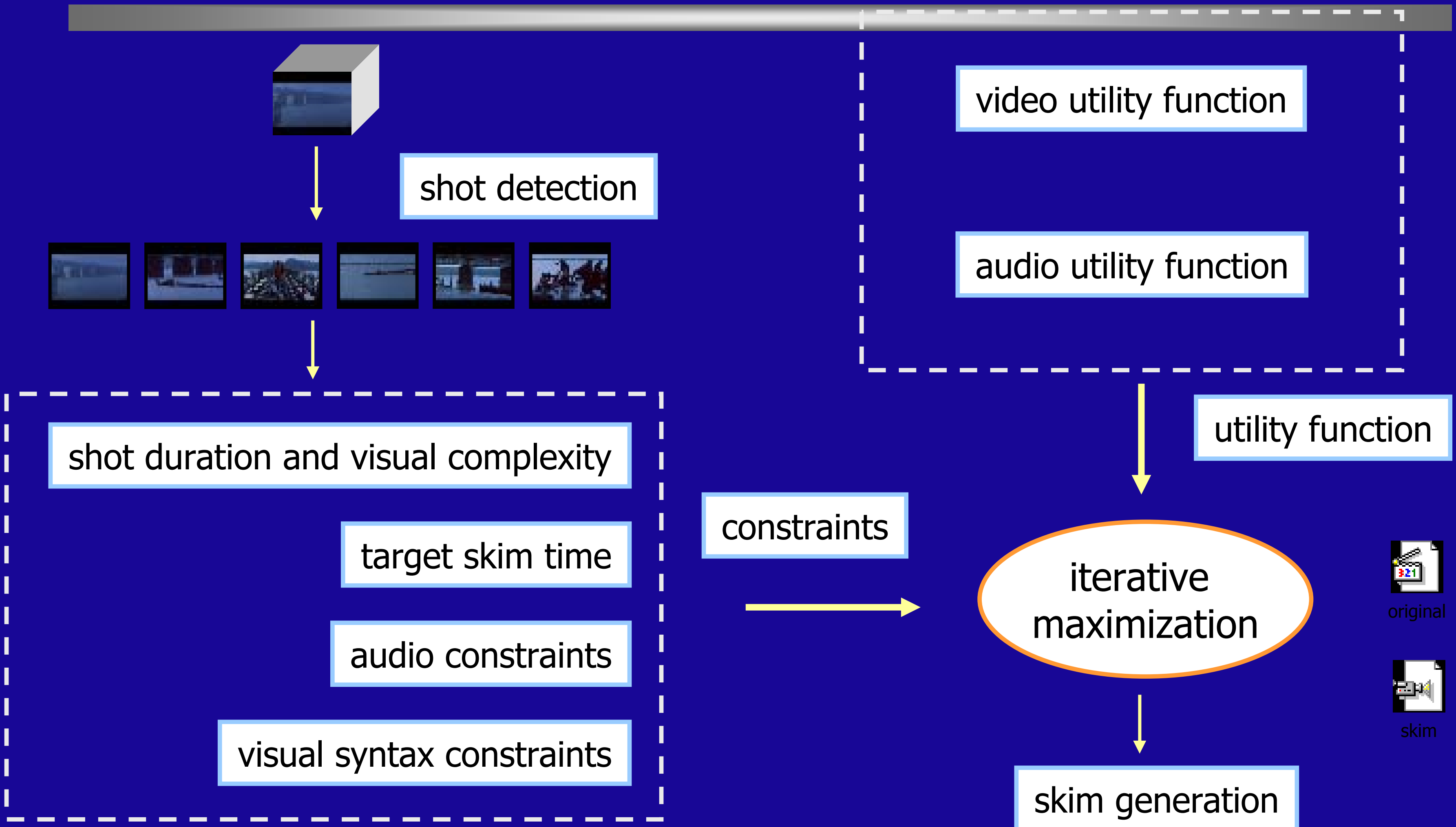


Tied Audio-Video Constraint



- Tied segments:
  - Include all significant
  - Audio and video boundaries are fully synchronized
  - Cannot be condensed or de-synchronized
  - 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

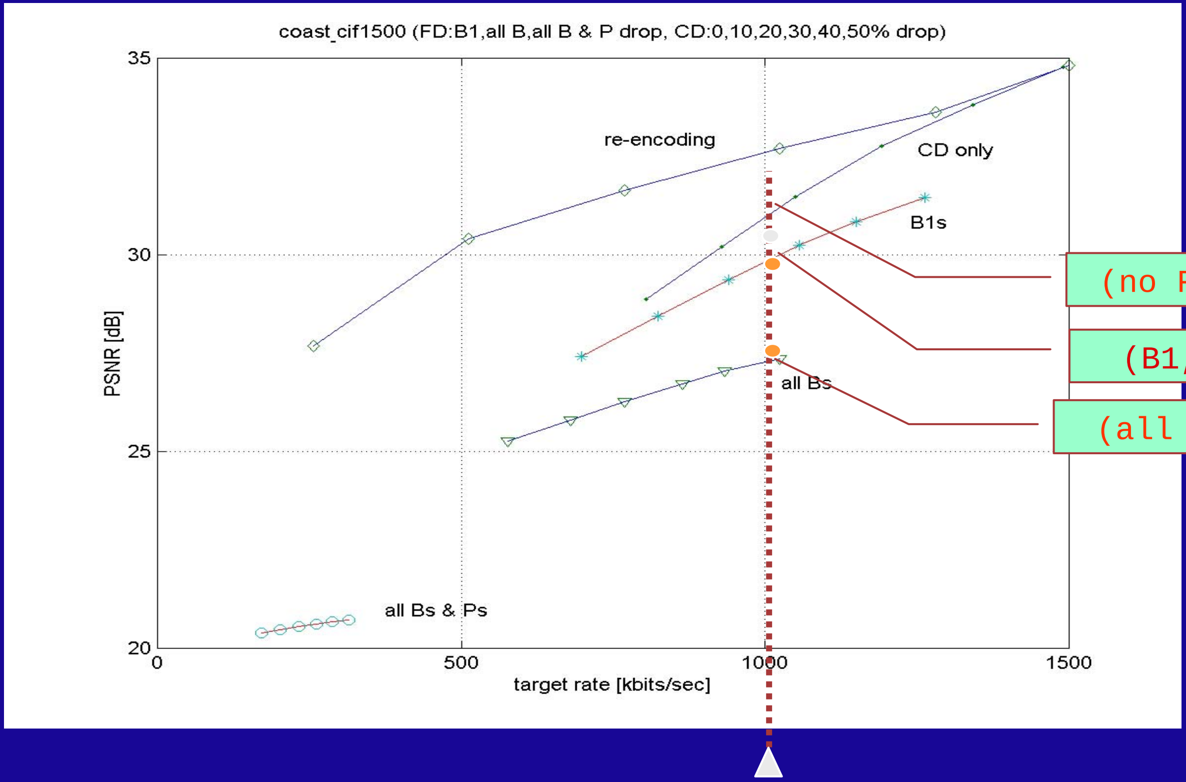
Utility Framework for Skim Generation [Chang, IWDC 02]



Case 2: Utility-Based MPEG-4 Video Transcoding

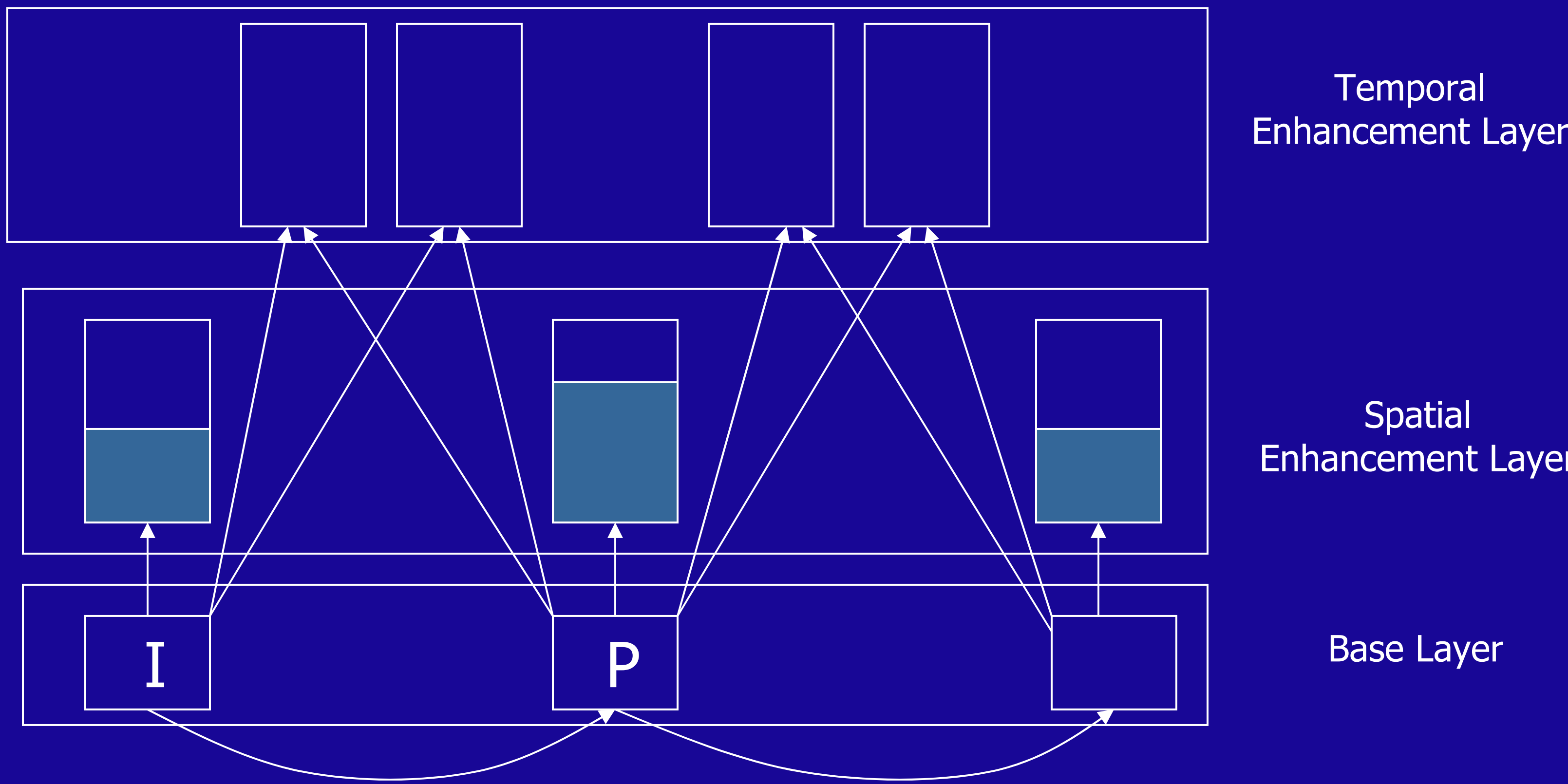
(with J. Kim and Y. Wang)

- Original bit rate → reduced rate (resource change)
- Adaptation Space: FD: frame dropping, CD: coefficient dropping, and combinations



- **Utility Ranking Description**: {(R_i, Rank(A_i), U_i, Consistency-Flag), i = 1,2, …}

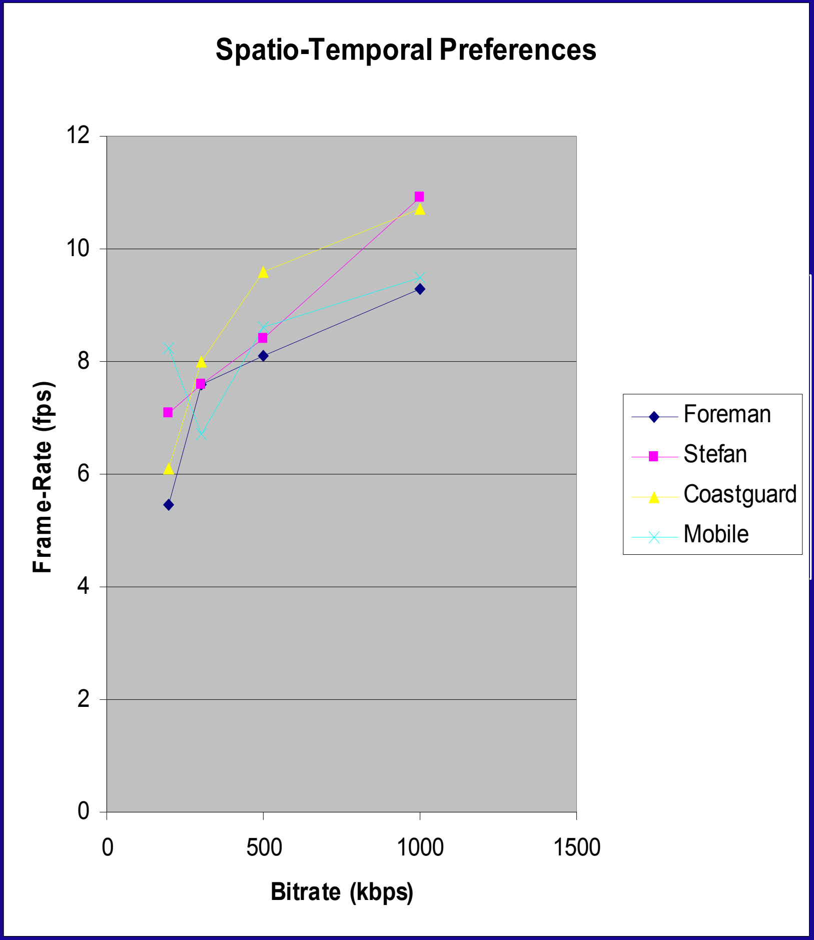
MPEG-4 Fine-Grained Scalability



- **Adaptation Space:**
  Temporal frame rate and SNR bit planes

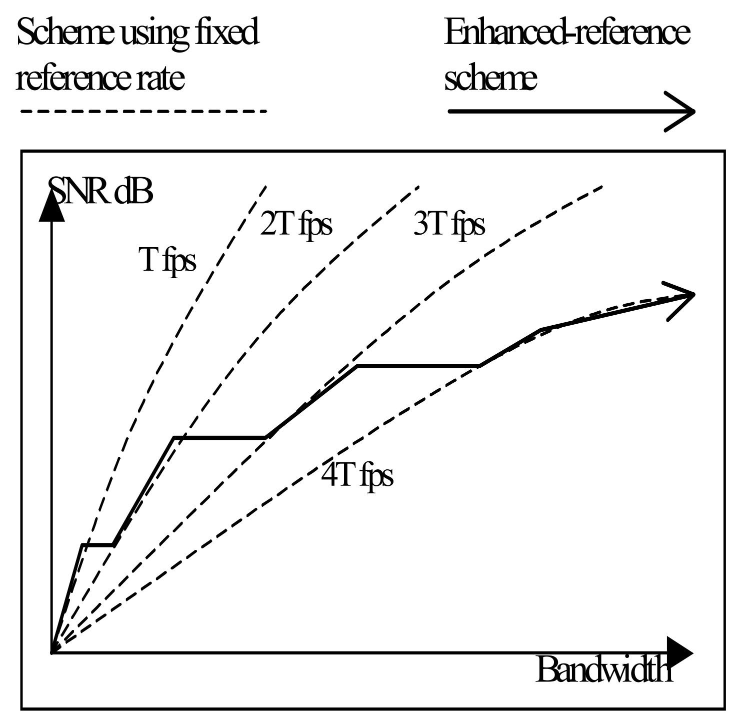
Utility Function of Subjective Spatio-Temporal Preference
(w. R. Kumar and M. van der Schaar)

- SNR is inadequate for measuring utility
- We conduct study where users chose preferred frame-rate at different bit-rates
- As the bit-rate goes up, people prefer better frame-rates
- Preference varies with video category
  - High-motion videos (Stefan, Coastguard) require a higher frame rate



Utility Model Guided FGS

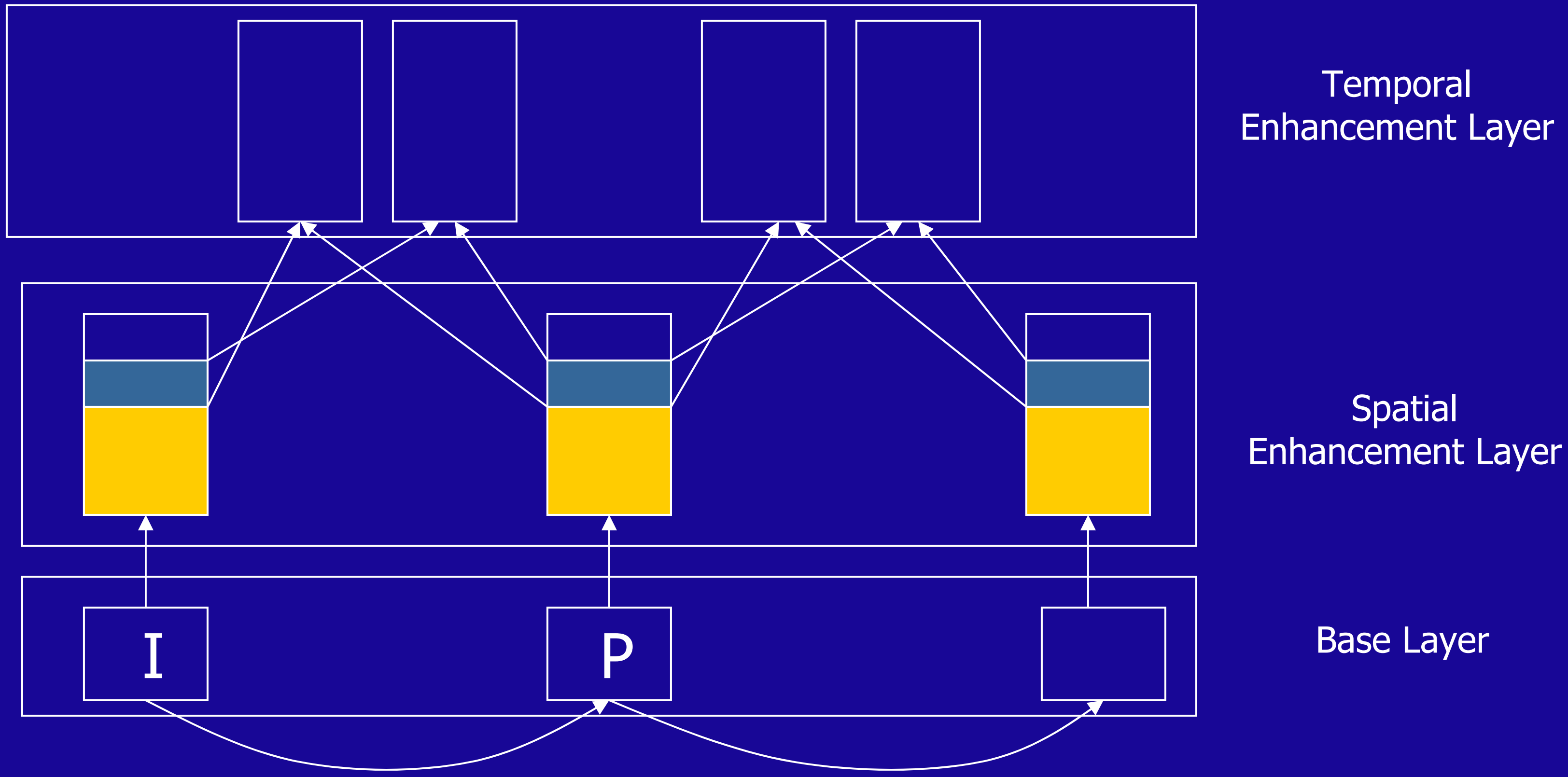
- When more bit rate available, increase SNR quality and fix temporal rate to a predetermined bitrate
- Then improve temporal quality
- Further improve SNR quality at the new frame-rate
- And so on….



(B) SNR quality of enhanced-scheme

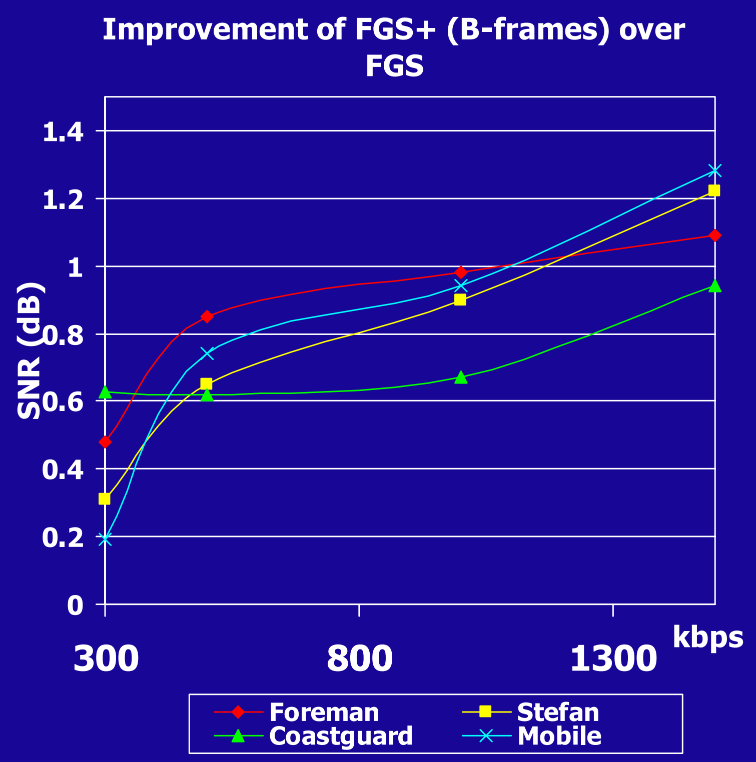
FGS+ Scheme

- **Solve the issue of determining optimal rate for motion prediction reference**



Performance

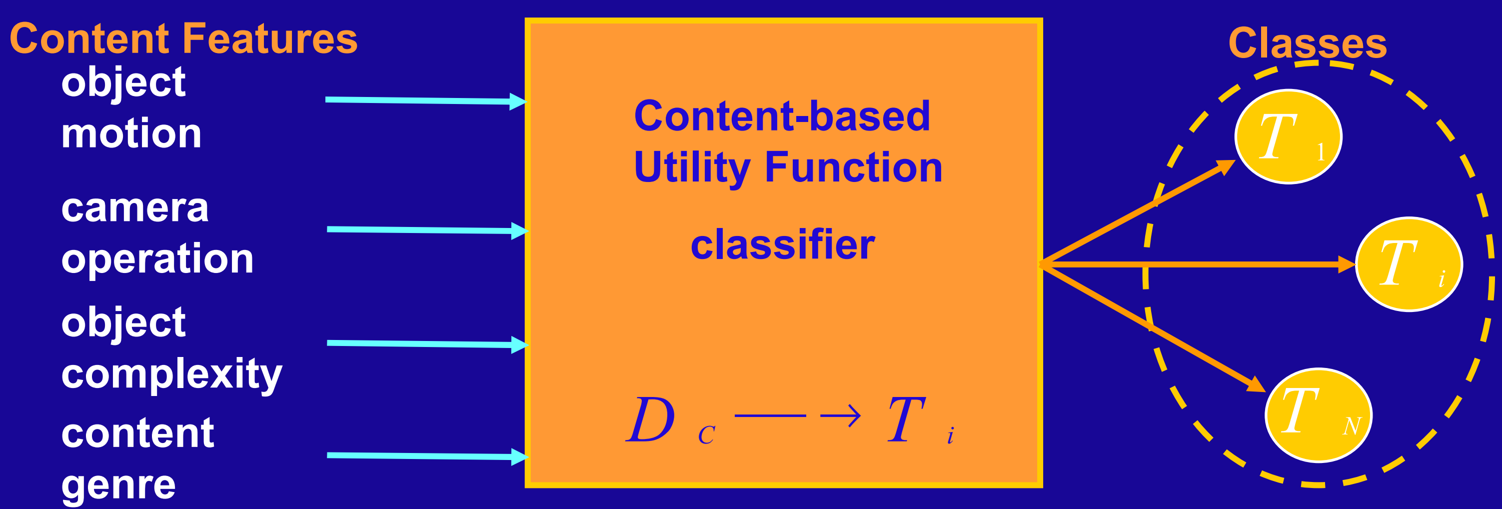
- Improvement over FGS varies from 0.19 dB to 1.28 dB
- At low bit-rates simple videos benefit (Coastguard)
- At high bit-rates complex videos benefit (Mobile)



Content-based utility classification and prediction

Challenge:
predict utility function for an arbitrary video?

content-utility correlation & content-based classification

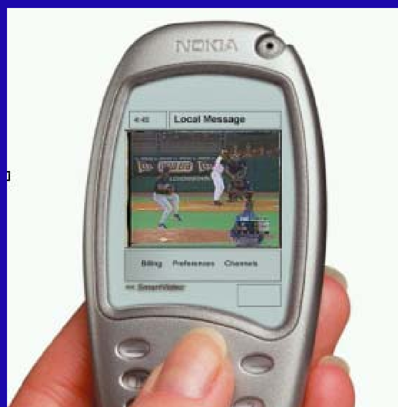


Bocheck and Chang 2000

Case: Live Sports Filtering and Adaptive Streaming

With D. Zhong and R Kumar, 2000

Real-time alert or streaming

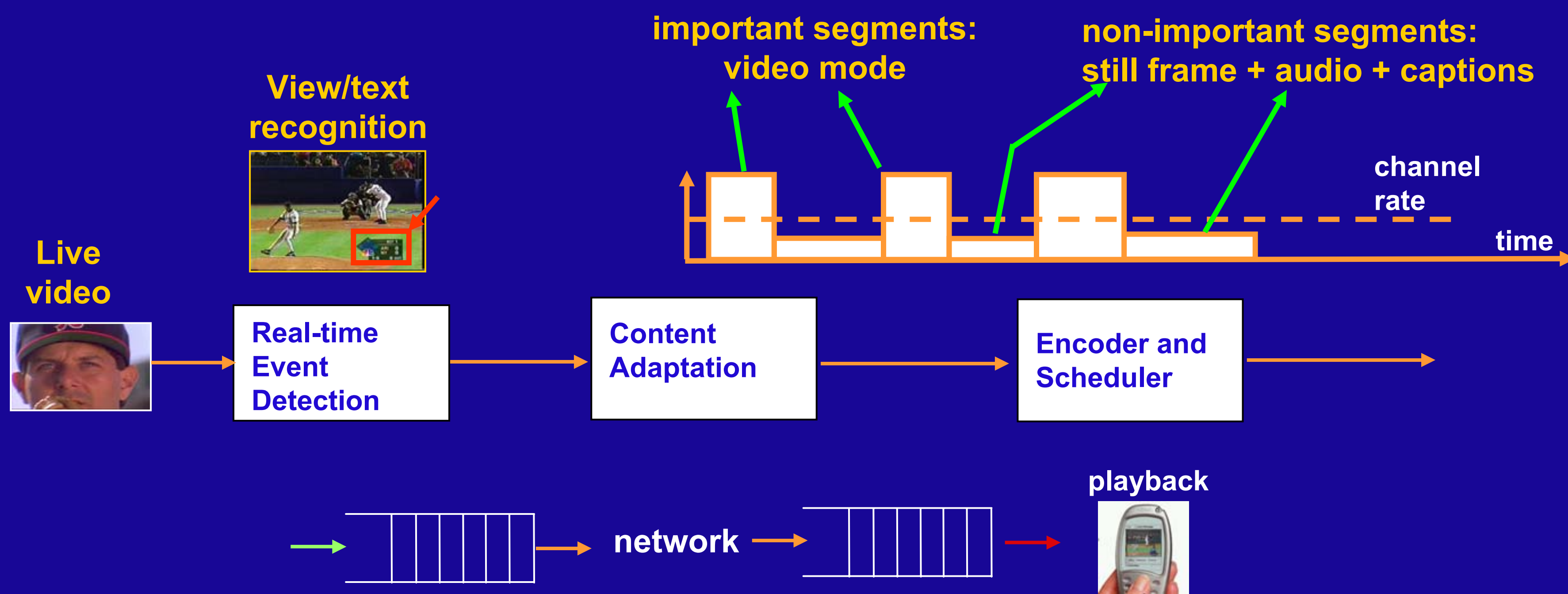


Interactive Video Filtering

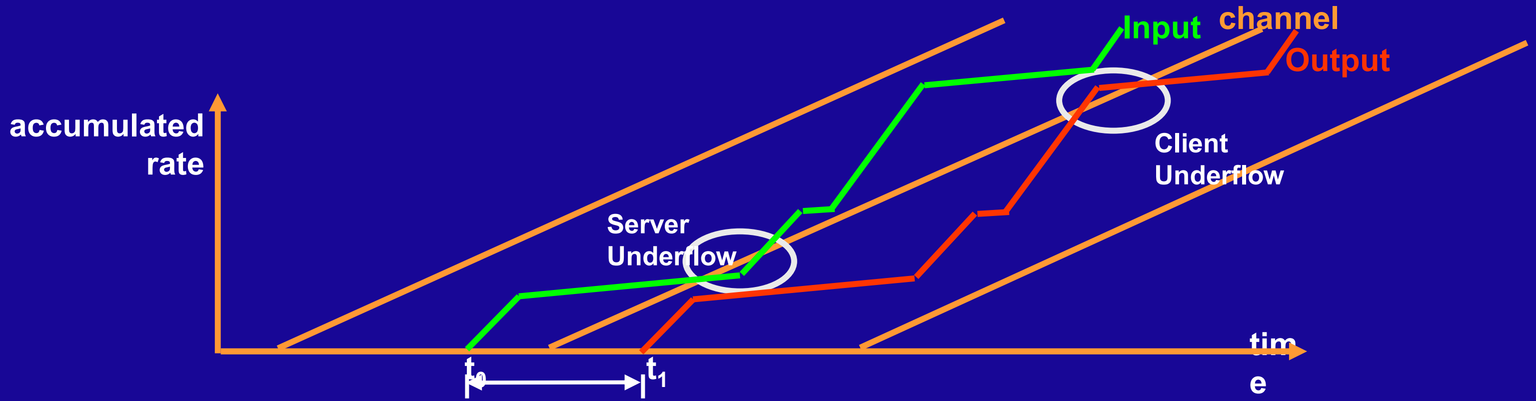
- Time sensitive interest
  - Need of real-time processes
- Time compressibility
  - Room for adaptation
- Temporal structure and production rules
  - content analysis feasibility

Utility Adaptive Video Streaming

- Model utility based on content “importance” vs. “non-importance”
- Utility-based adaptive rate allocation

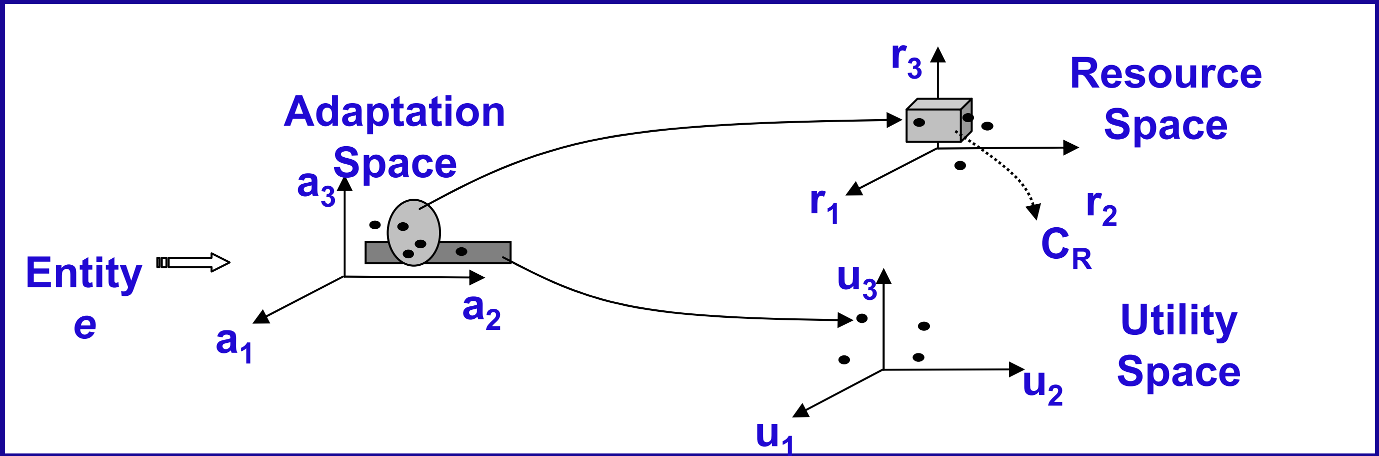


System Issues for Adaptive Streaming



- Increase buffer size feasible
  - 8MB buffer can store 2000 sec (32Kbps) - 250 sec (256Kbps)
- But playback latency is the main constraint
  - E.g., deliver real-time event within 60 sec
- Client error more serious than server error
  - Degrade to Channel Rate, Freeze and resume, or adaptive playback speed.

Conclusions



- A generalized conceptual framework for
  - Modeling relationships among content entities, adaptation processes, utility, and resources
  - Formulating optimization tasks, e.g.,
    - Time condensed skims
    - Modeling spatio-temporal utility preference
    - Content-adaptive streaming
- Several remaining issues
  - utility model, high-dimensional representation, search