Multimedia Content Analysis, Organization, and Presentation

朱威達
wtchu@cs.ccu.edu.tw

國立中正大學資訊工程系 助理教授

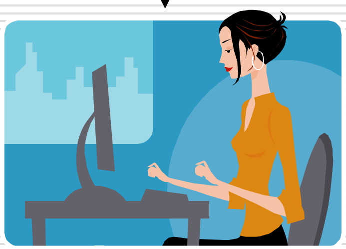
Multimedia Content Analysis

Audios
Videos
Music
Images
…

Multimedia Content

1. How to efficiently access?
2. What to access?

Content analysis & organization



Users

Find what we want in large volumes of content.

1. Classification
   - Music or video genre classification
   - Content-based clustering
2. Retrieval
   - Content-based image retrieval
   - Query by keywords

…

Get what we want in a specific data item.

1. Segmentation
   - Speech/music discrimination in audio streams
   - Structure analysis in sports videos
2. Adaptation
   - Video summarization / highlight extraction
   - Event detection

…

Information is of no use unless you can actually access it.

DEMO –
Baseball Video Analysis

Outline

- Multimedia Content Analysis
  - Video analysis
  - Audio analysis
- Multimedia Content Organization
  - Video summarization / highlight
- Multimedia Content Presentation
  - Multimodality collaborative presentation
- Summary

Outline

- **Multimedia Content Analysis**
  - **Video analysis**
  - **Audio analysis**
- Multimedia Content Organization
  - Video summarization / highlight
- Multimedia Content Presentation
  - Multimodality collaborative presentation
- Summary

Incentives to Analyze Videos

Sports video

- Find the events invoked by your favorite player.
- Find all home run events.
- Quickly view a game within ten minutes.

Movie video

- Find the most impressive scenes.
- Story segmentation
- Movie trailers.

Home video

- Find specific event.
- Automatic editing

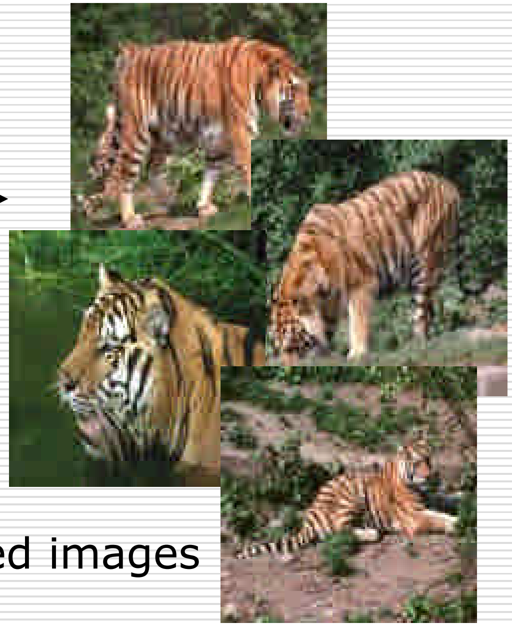
Feature-based Content Analysis

Example: Content-based Image Retrieval

Query image

Content-based image matching

Image database



Retrieved images

Metrics for matching: color, texture, edge, …

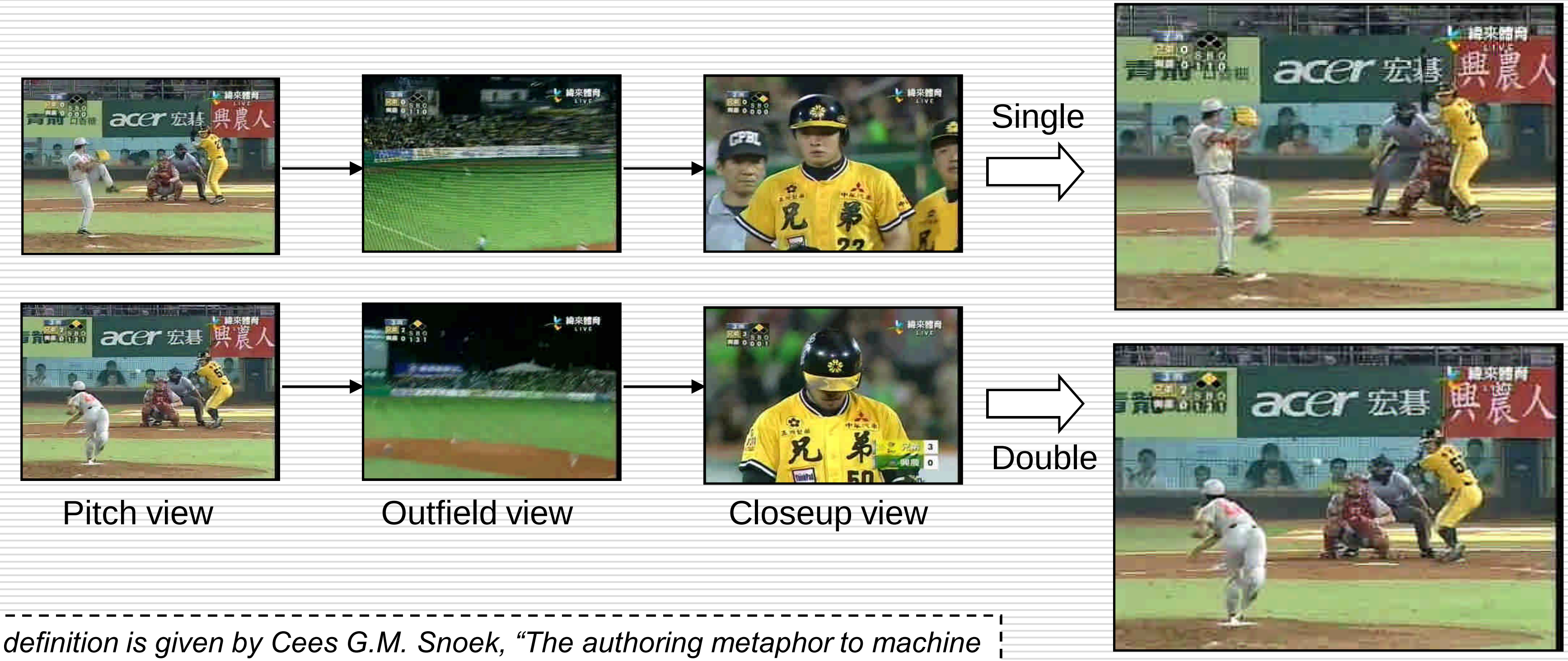
CIRES: Content-Based Image Retrieval System
http://amazon.ece.utexas.edu/~qasim/research.htm

Audio/Video Features

- Features: Data characteristics that can be directly computed from content.
- Audio Features
  - Energy, Zero-crossing rate, …
- Image/Video features
  - Color, Texture, Edge, Motion, …

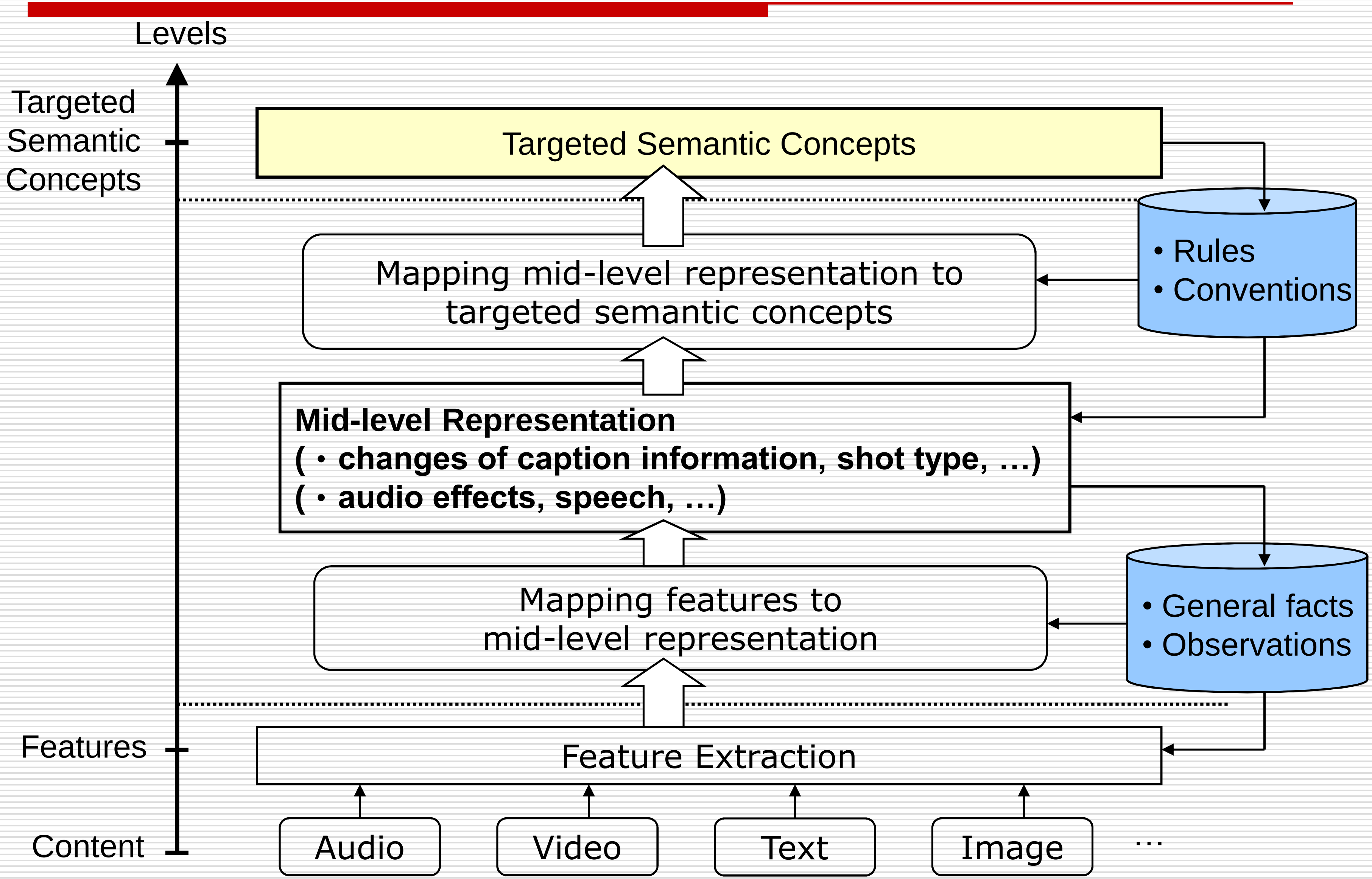
Semantic Gap Problem

- The lack of coincidence between the information that machines can extract from the multimedia data and the interpretations the user may give to the data.

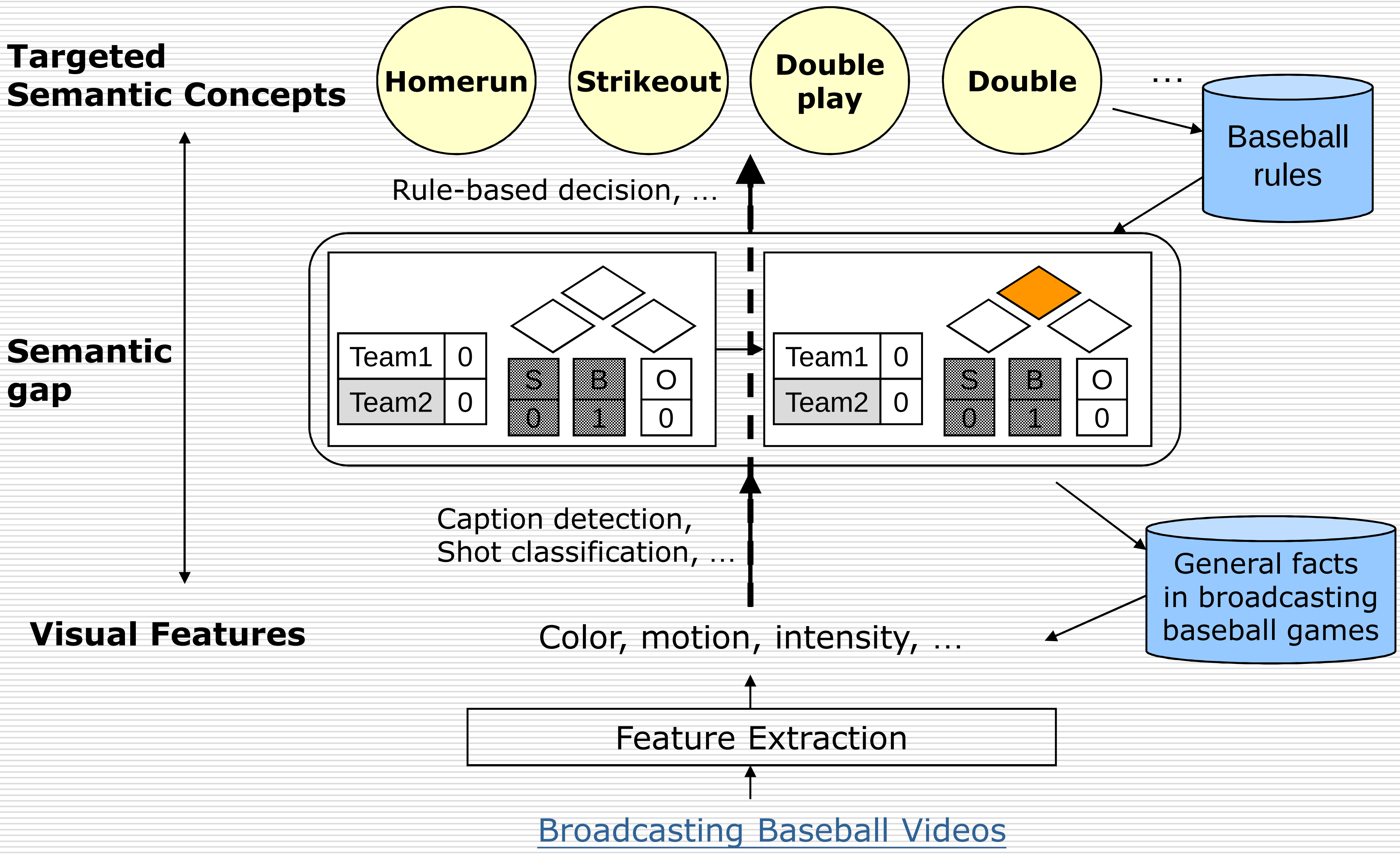


** The definition is given by Cees G.M. Snoek, "The authoring metaphor to machine understanding of multimedia," PhD thesis in U. of Amsterdam.*

Semantic Concept Detection Framework



Example: General Ideas in Baseball Concept Detection



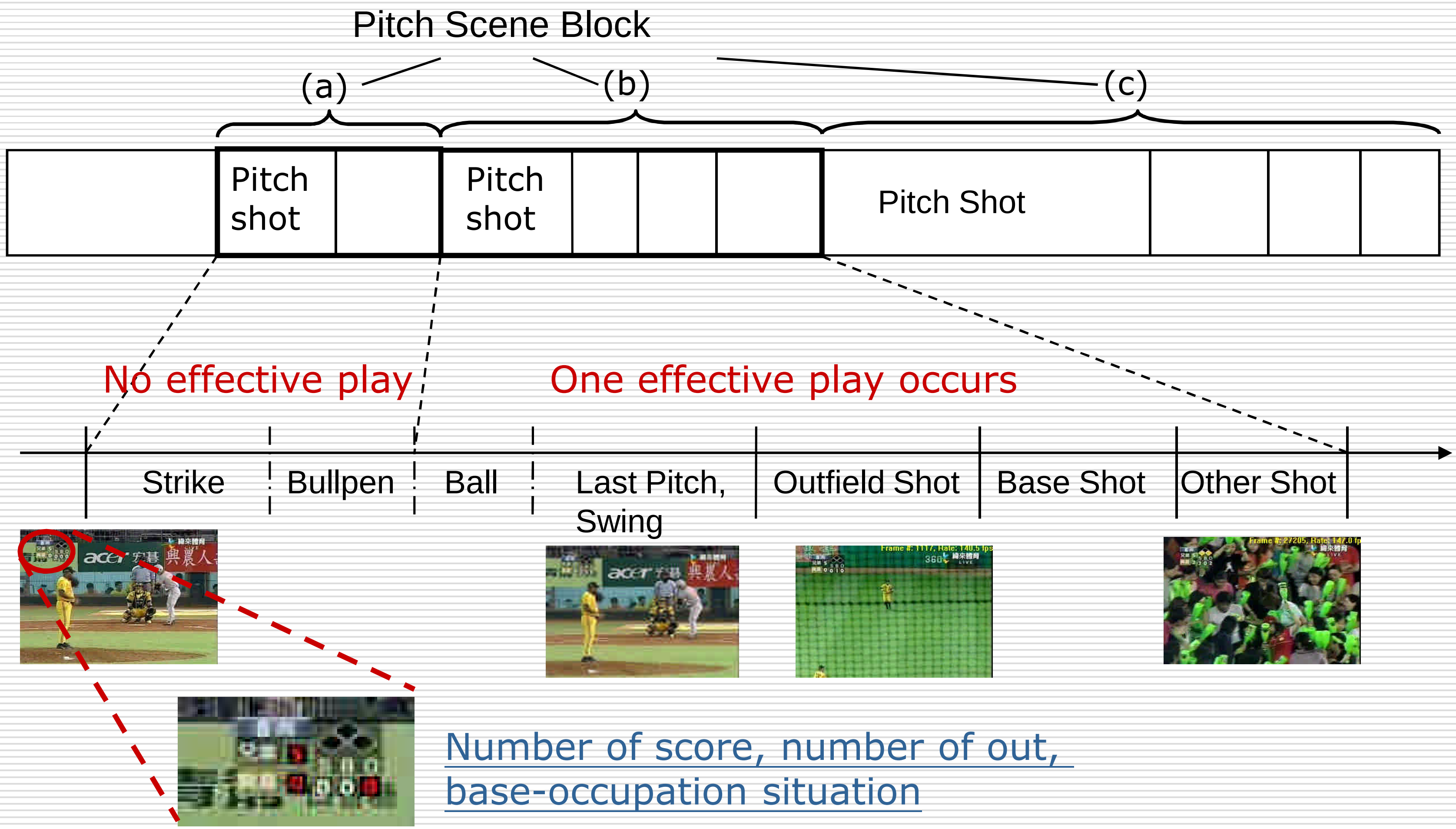
Baseball Video Analysis & Organization

Introduction

- Different plays have different meanings to the player and to the fans.
- A fan would more like to see "what really happened" rather than "rough summarization" of a baseball game.
- Official baseball rules and shot transition info. are integrated to facilitate explicit concept detection.

Baseball Game Progress

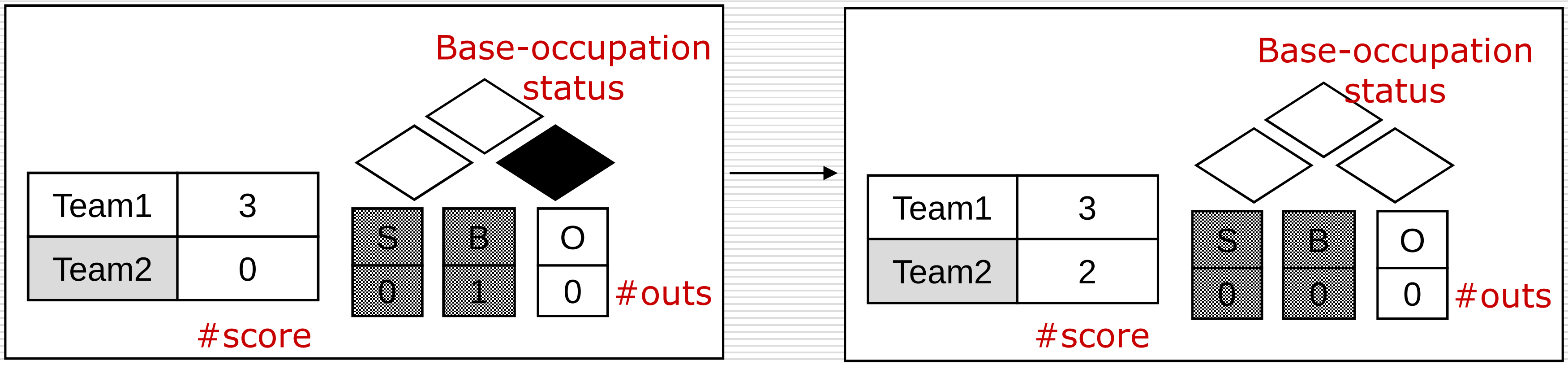
PSB: The video segment between two pitch shots
PSB is the basic unit for baseball events.



Rule-Based Decision

Caption in *i*th pitch shot

Caption in *(i+1)*-th pitch shot



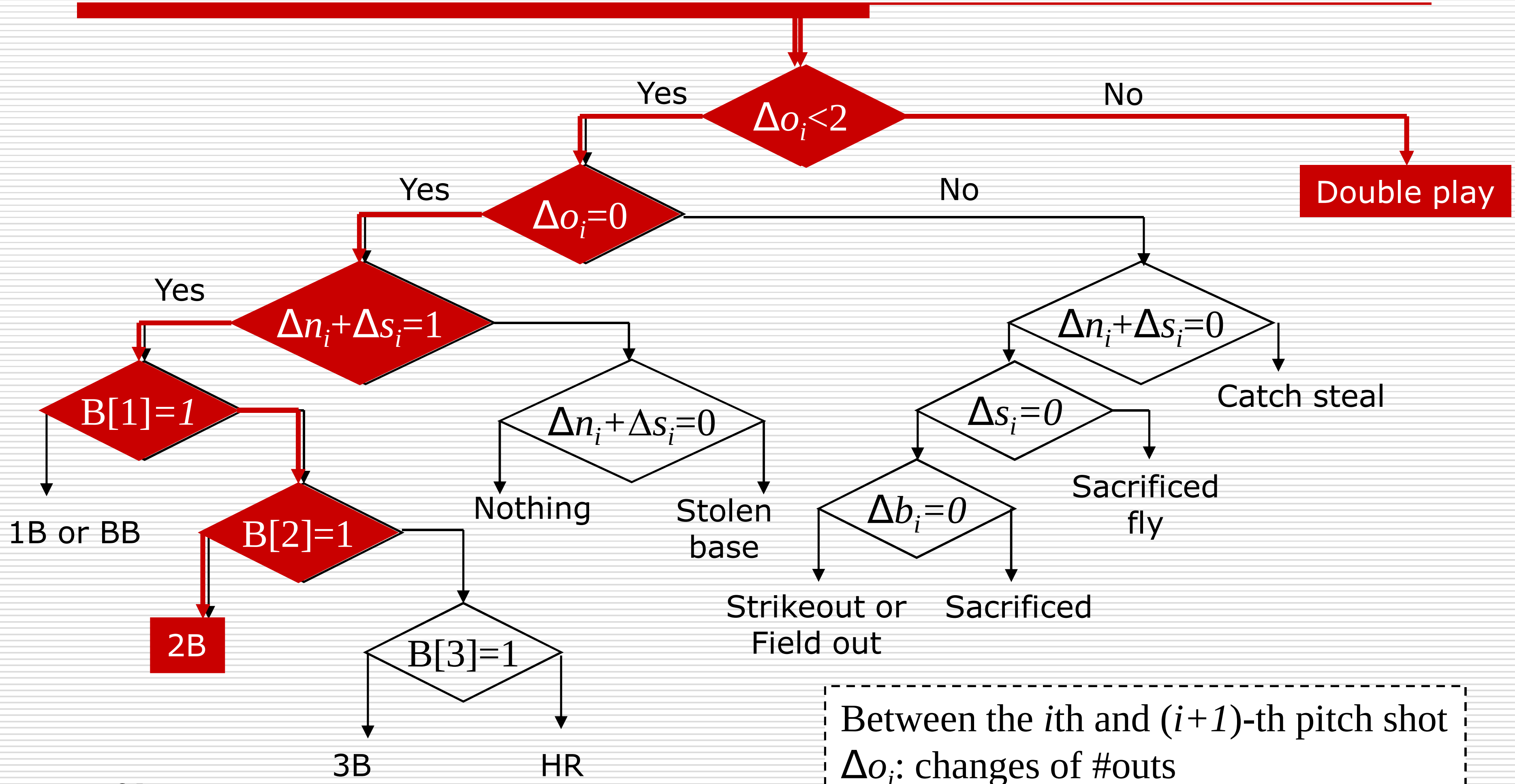
(#outs1, #score1, _base1) = (0, 0, 1)

(#outs2, #score2, _base2) = (0, 2, 0)

compare

(Δouts, Δscore, Δbase) = (0, 2, -1) → Home Run

Rule-based Decision Tree



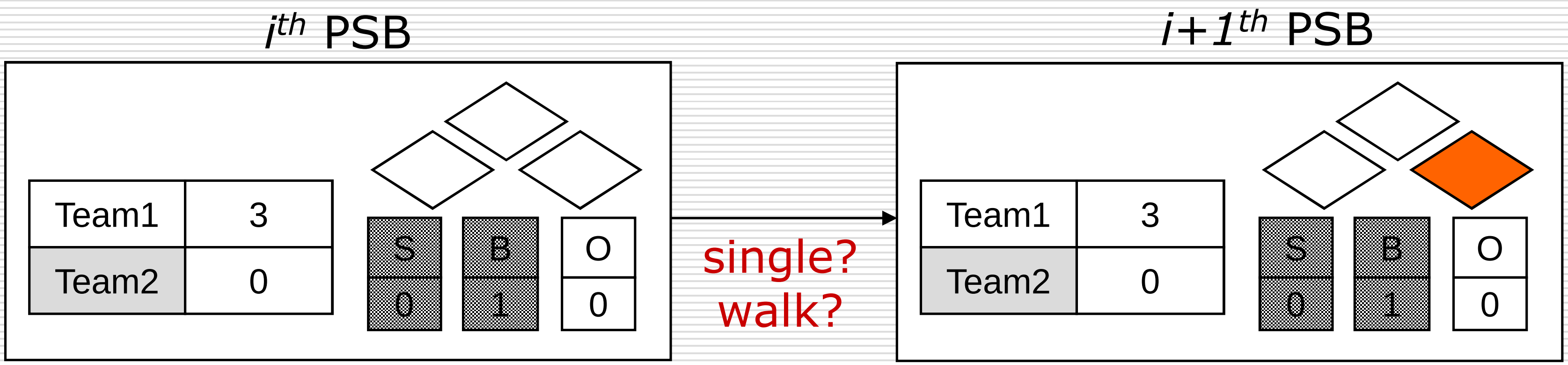
Status filtering:

$$f_{i,i+1}=\begin{cases} legal, & \text{if } (\Delta n_i+\Delta s_i+\Delta o_i)=0 \text{ or } 1, \\ illegal, & \text{otherwise.}\end{cases}$$

Between the *i*th and (*i*+1)-th pitch shot
Δo_i: changes of #outs
Δs_i: changes of #scores
Δn_i: changes of #occupied bases
B[*j*]: whether the *j*th base is occupied

Confused Concepts

- Some concepts can't be discriminated by simply using baseball rules.
  - single vs. walk
  - strikeout vs. field out

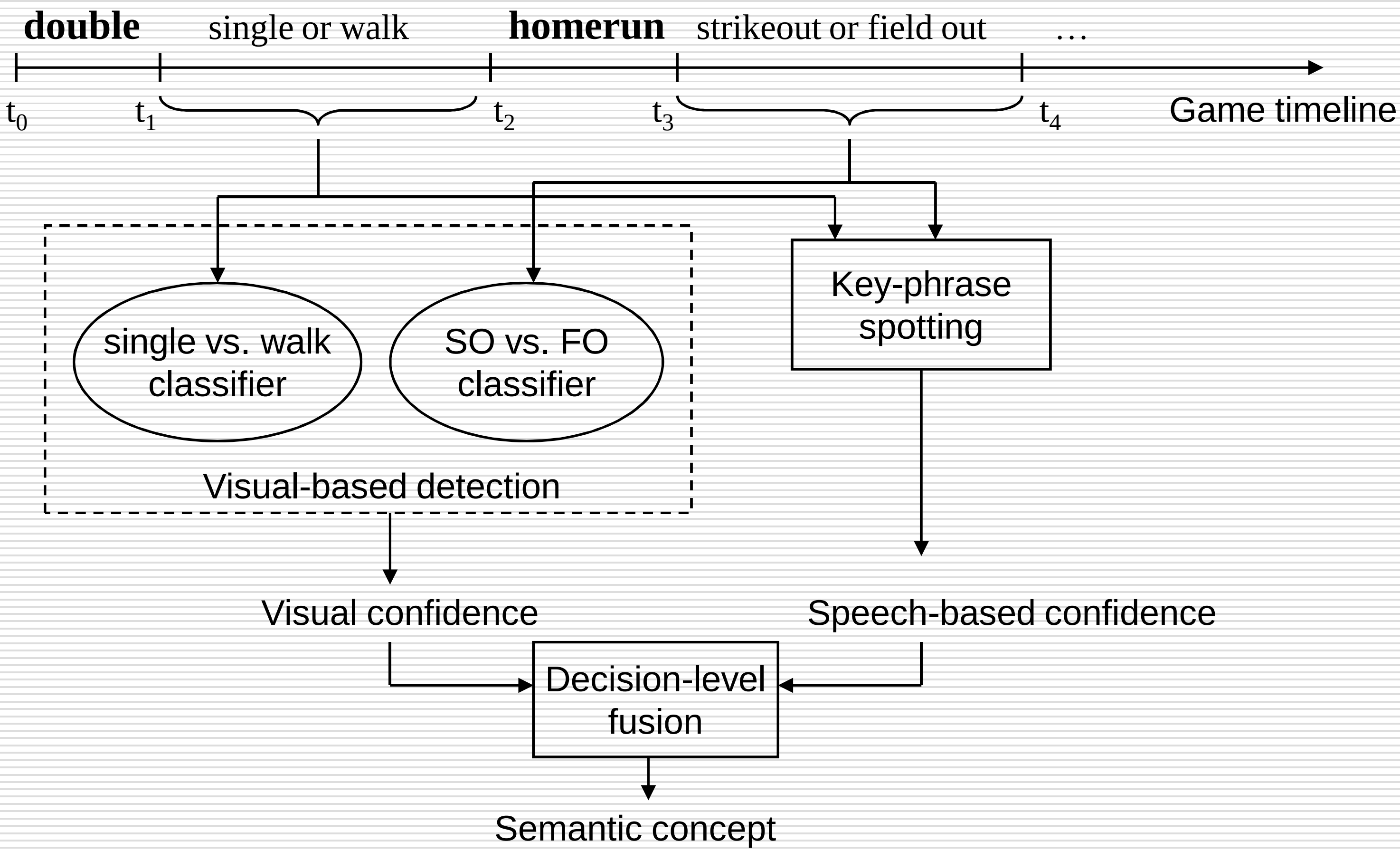


No bases is occupied. n_i = 0.
Team2's score is 0.

The first base is occupied. n_i = 1.
Team2's score is 0.

Confused Concept Discrimination

- Combine visual and speech information



Visual-Based Detection

- The shot context features are modeled by K-nearest neighbor method for training and testing.
- "Single vs. Walk" classifier and "Strikeout vs. Field out" classifiers are constructed.
- Confidence of visual-based detection

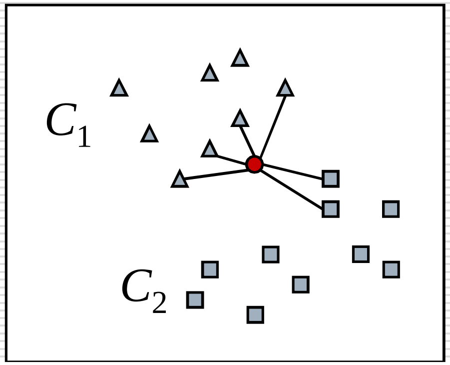
$$P(C_1 | \mathbf{x}_1) = \frac{K_1}{K}$$

$$P(C_2 | \mathbf{x}_1) = \frac{K_2}{K}$$

$$K = K_1 + K_2$$

$\mathbf{x}_1$: shot context features

K_1(K_2) is the number of patterns among $\mathbf{x}_1$'s K nearest neighbor that belong to class C_1(C_2).



Key Phrases in Anchorperson’s Speech

- Recognize key-phrases from the anchorperson’s speech.

Concepts	**Corresponding Key-phrases**
Single (C_1)	R_1={安打(hit), 一壘安打(single)}
Walk (C_2)	R_2={觸身球(hit by pitch), 保送(walk), 四壞球(four balls)}
Strikeout (C_3)	R_3={三振(strikeout), 三振出局(strikeout)}
Field out (C_4)	R_4={刺殺(‘touch out’ or ‘out before reaching bases’), 接殺 (catch out)}

...這個球越過三壘防區, 形成了一支安打,...

...這個球二壘後方的小飛球遭到石志偉的接殺, 陽東益出局, 形成兩出局,...

...哇...這邊出現右外野方向的安打,..., 得分效率百分百, 就是四壞球上壘、推進、安打, 這樣好像很容易就換回一分了,...

** The key-phrase spotting module is supported by Prof. Lin-Shan Lee, National Taiwan University.*

Confidence of Speech-based Detection

- Case 1: only key phrases in R_1 are recognized

$$P(C_1|\mathbf{x}_2=R_1)=\frac{\#(C_1)}{\#(\text{only the key-phrases in } R_1 \text{ are recognized})}$$

$$P(C_2|\mathbf{x}_2=R_1)=\frac{\#(C_2)}{\#(\text{only the key-phrases in } R_1 \text{ are recognized})}$$

($\mathbf{x}_2$: recognized key phrases)

- Case 2: only key phrases in R_2 are recognized

$$P(C_1|\mathbf{x}_2=R_2)=\frac{\#(C_1)}{\#(\text{only the key-phrases in } R_2 \text{ are recognized})}$$

$$P(C_2|\mathbf{x}_2=R_2)=\frac{\#(C_2)}{\#(\text{only the key-phrases in } R_2 \text{ are recognized})}$$

- Case 3: key phrases in both R_1 and R_2 are recognized

$$P(C_1|\mathbf{x}_2=R_1,R_2)=\frac{\#(C_1)}{\#(\text{key-phrases in } R_1 \text{ and } R_2 \text{ are recognized})}$$

$$P(C_2|\mathbf{x}_2=R_1,R_2)=\frac{\#(C_2)}{\#(\text{key-phrases in } R_1 \text{ and } R_2 \text{ are recognized})}$$

They are calculated based on four different games.
There are totally about 50 “1B or walk” and 100 “SO or FO”.

Case 1:
P(single|安打)=0.969
P(walk|安打)=0.031

Case 2:
P(single|四壞)=0.091
P(walk|四壞)=0.909

Case 3:
P(single|安打,四壞)=0.455
P(walk|安打,四壞)=0.545

Case1:
P(strikeout|刺殺)=0
P(field out|刺殺)=1

Case 2:
P(strikeout|三振)=0.95
P(field out|三振)=0.05

Case 3:
P(strikeout|刺殺,三振)=0.167
P(field out|刺殺,三振)=0.833

Combine Visual and Speech Opinions

- Given the feature $Z=(\mathbf{x}_1,\mathbf{x}_2)$, decide which class it belongs to by using the sum rule.

$$\text{assign } Z \rightarrow C_j \text{ if } \sum_{i=1}^{2} P(C_j|\mathbf{x}_i) = \max_{k=1}^{2} \sum_{i=1}^{2} P(C_k|\mathbf{x}_i)$$

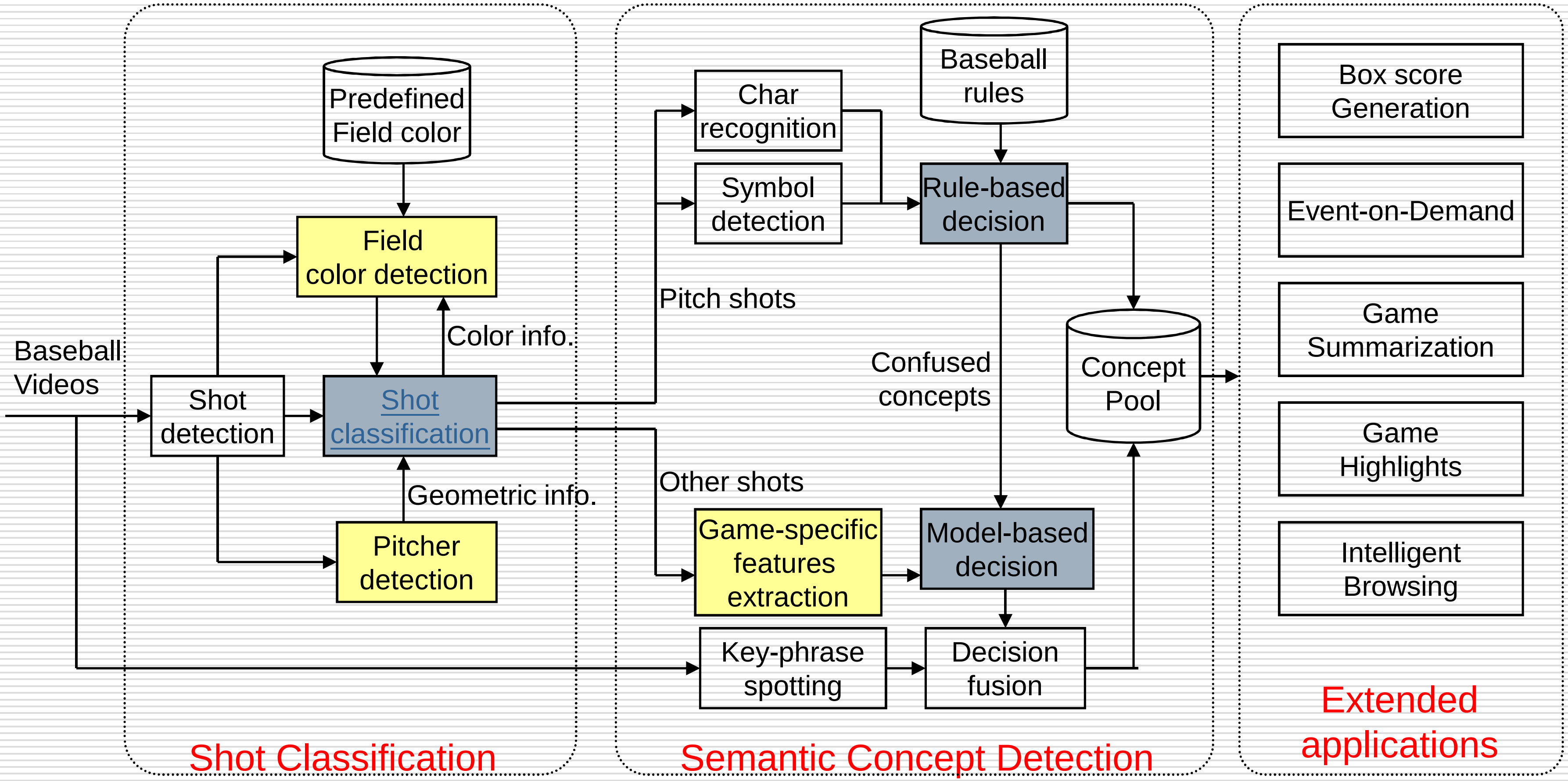
Other combining strategies:

$$\text{assign } Z \rightarrow C_j \text{ if } \prod_{i=1}^{2} P(C_j|\mathbf{x}_i) = \max_{k=1}^{2} \prod_{i=1}^{2} P(C_k|\mathbf{x}_i)$$

$$\text{assign } Z \rightarrow C_j \text{ if } \max_{i=1}^{2} P(C_j|\mathbf{x}_i) = \max_{k=1}^{2} \max_{i=1}^{2} P(C_k|\mathbf{x}_i)$$

$$\text{assign } Z \rightarrow C_j \text{ if } \min_{i=1}^{2} P(C_j|\mathbf{x}_i) = \max_{k=1}^{2} \min_{i=1}^{2} P(C_k|\mathbf{x}_i)$$

System Framework



Thirteen concepts are detected:
single (1B), double (2B), triple (3B), home run (HR), stolen base (SB), caught stealing (CS), fly out (AO), strikeout (SO), base on ball (Walk, BB), sacrifice bunt (SAC), sacrifice fly (SF), double play (DP), and triple play (TP).

Outline

- Multimedia Content Analysis
  - Video analysis
  - Audio analysis
- **Multimedia Content Organization**
  - **Video summarization / highlight**
- Multimedia Content Presentation
  - Multimodality collaborative presentation
- Summary

Extended Applications

- Automatic Generation of Box Score
- Automatic Game Summarization
  - Maintain “informativeness” in a short duration.
  - Select plays based on their contributions.
- Automatic Highlight Generation
  - Maintain “enjoyability” in a short duration.
  - Select plays based on contributions, occurrence time, and audio energy dynamics.

Game Summarization
Game Highlight

2005.04.08 興農 vs. 統一
比賽時間：3小時14分

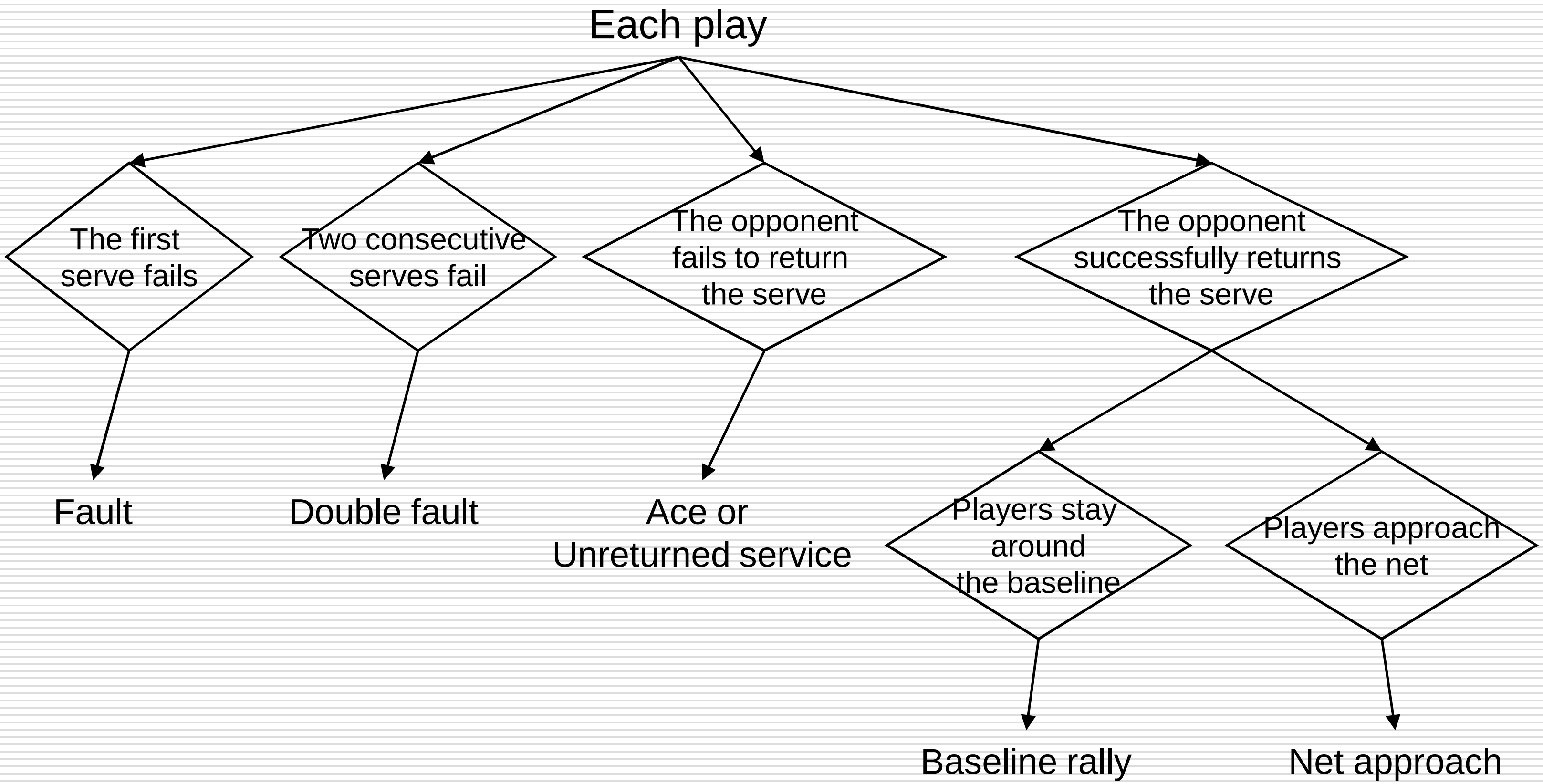
Man-made summary	Automatic summary	Automatic highlight	Automatic highlight
			
	16分鐘	3分25秒	6分鐘

31 plays are selected.
25 plays are in the
man-made sum.

Precision=0.806
Recall=0.833

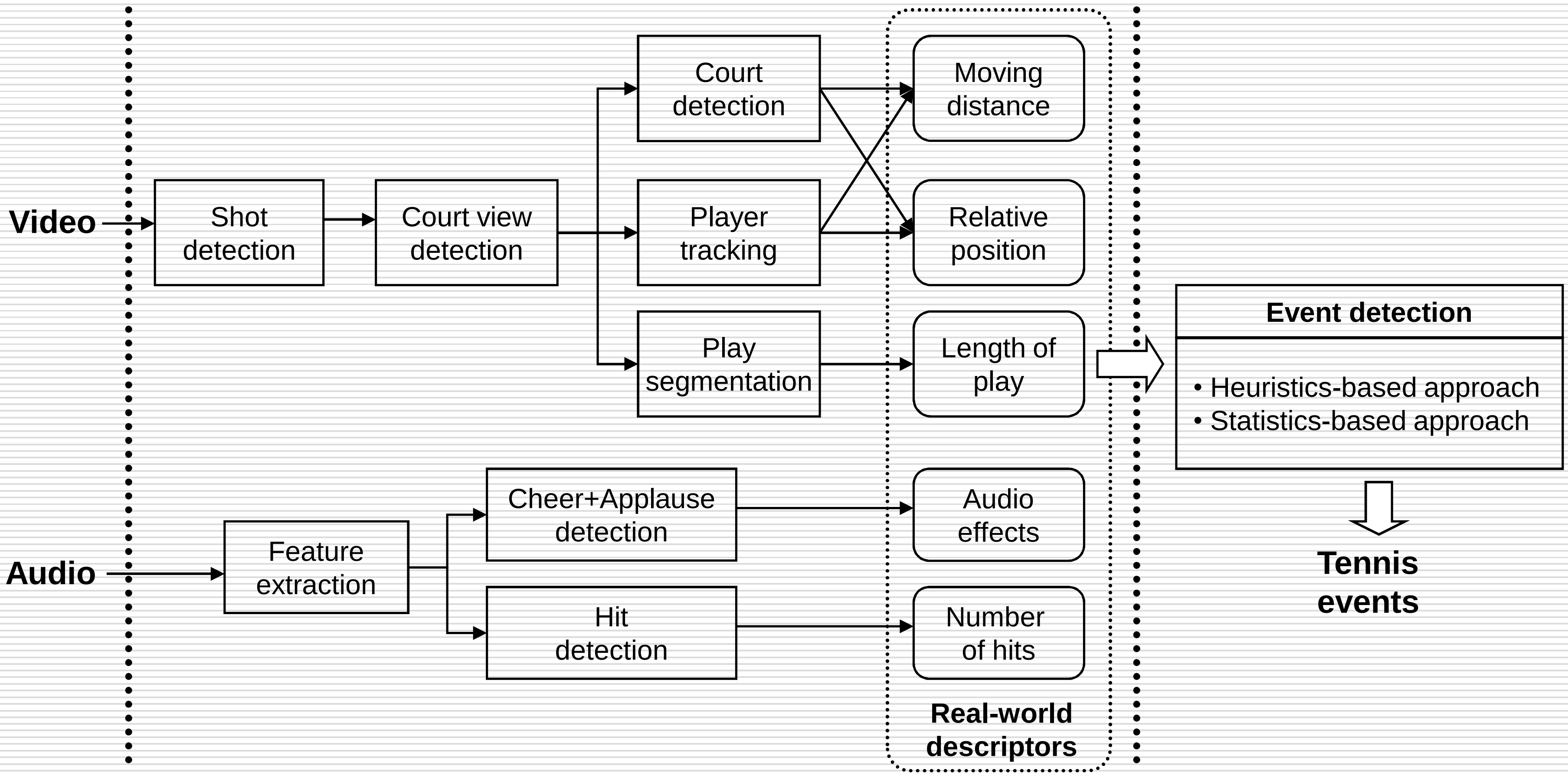
Tennis Video Analysis & Organization

Tennis Events



play　out-of-play　play　out-of-play　play

System Framework



DEMO –
Tennis Video Analysis

Outline

- Multimedia Content Analysis
  - Video analysis
  - Audio analysis
- Multimedia Content Organization
  - Video summarization
- **Multimedia Content Presentation**
  - **Multimodality collaborative presentation**
- Summary

Presentation

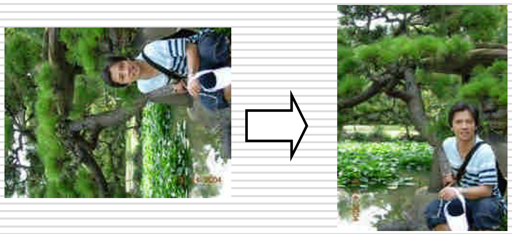
- Better presentation facilitates users browse massive multimedia data efficiently.
- Better presentation eases users in capturing the concept conveyed by multimedia data.

Tiling Slideshow

Motivation

- Large amounts of **consumer photos** derive the following problems:
  - Filtering or correcting are annoying.
  - Browsing photos takes much time.
  - Sequential presentation makes users boring.

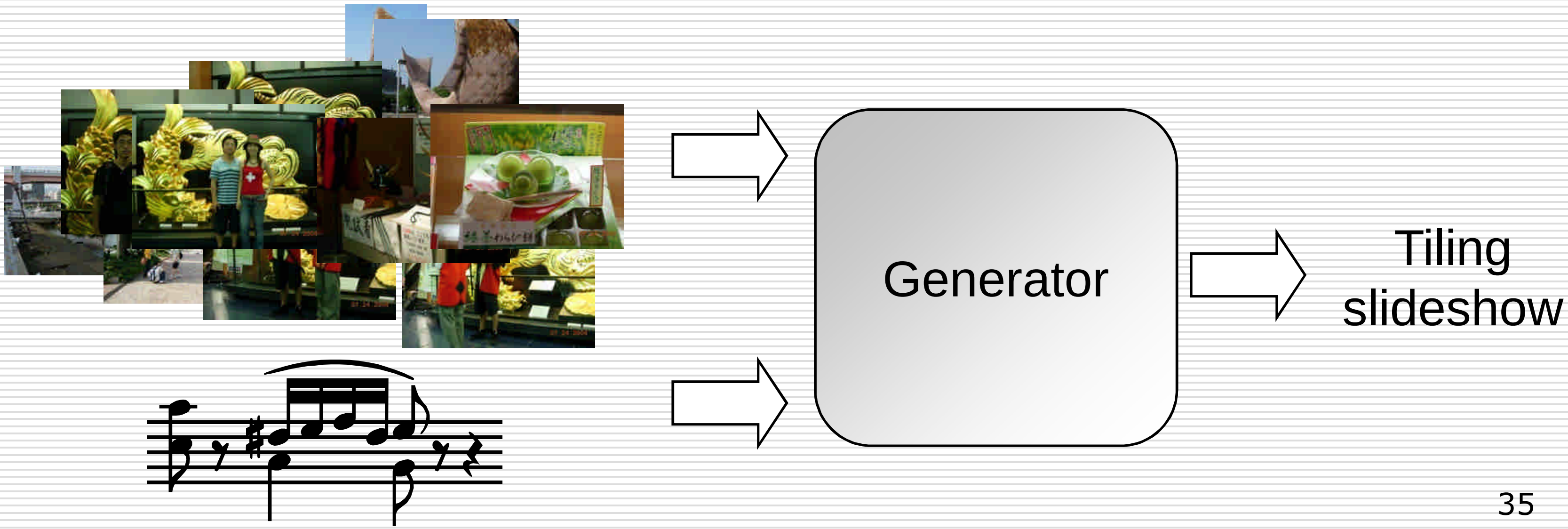
blurred photo



orientation correction

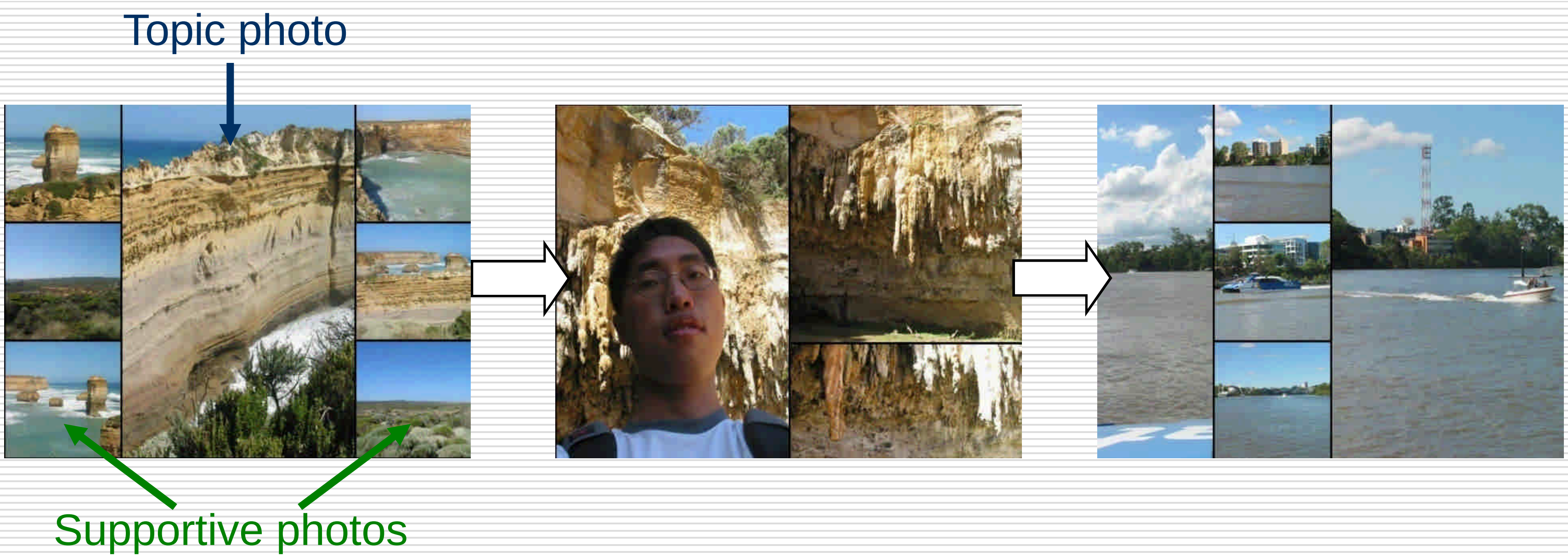
Goal

- Generate a kind of **new media** that provides user elaborate **photo browsing experience**.
  - Photo filtering & organization
  - Vivid audiovisual presentation
  - Value-added results

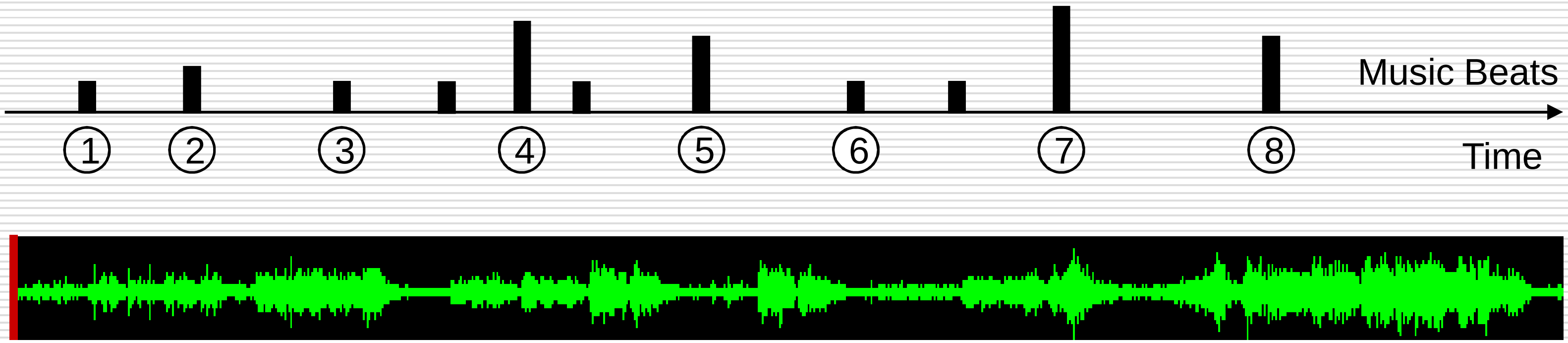


Photographic Story

- Paragraph: describe by text
  - Contains a topic sentence and several supportive sentences.
- Photographic paragraph: describe by photos
  - Contains a topic photo and several supportive photos



The Proposed Slideshow



System Overview

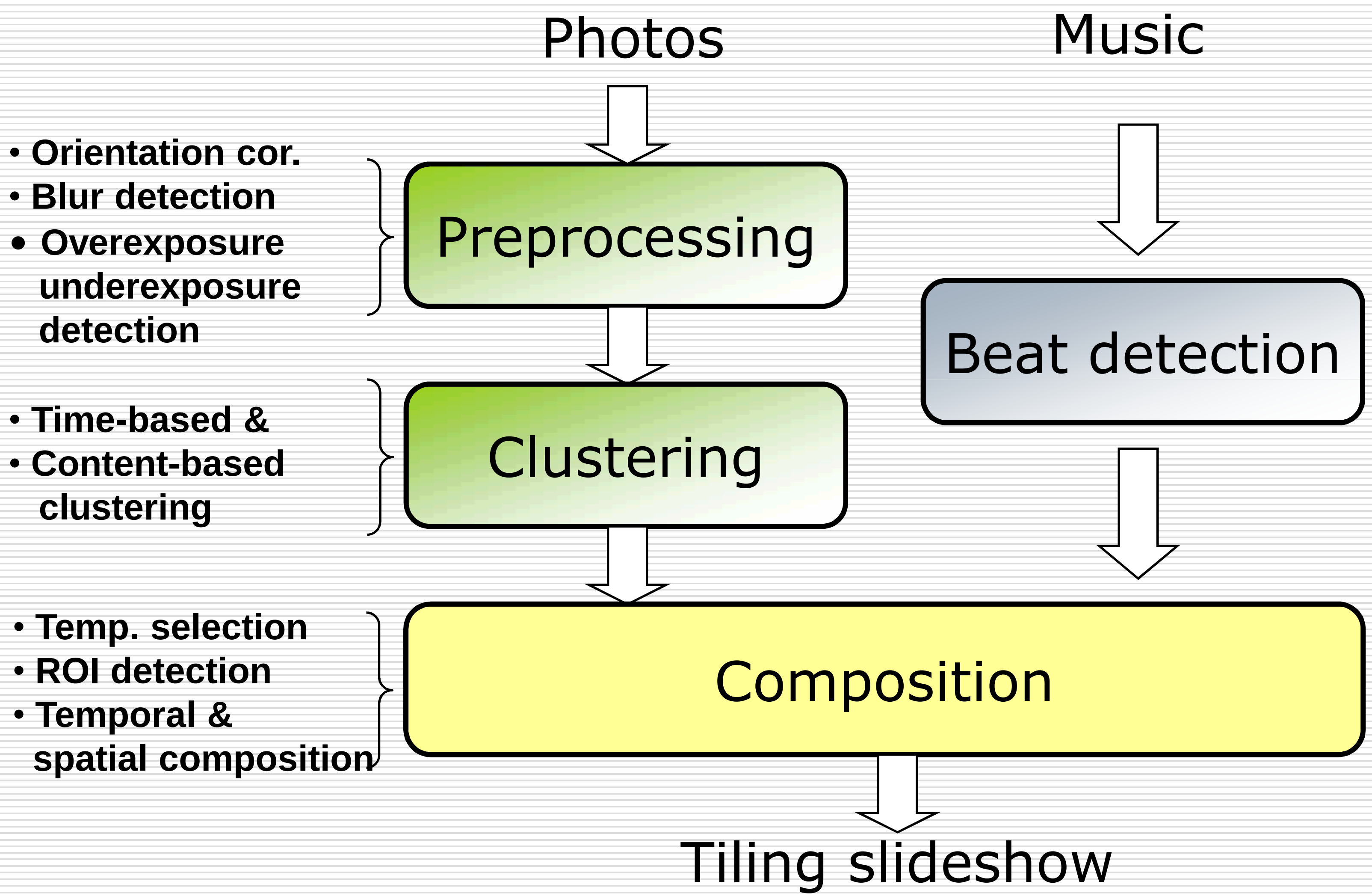


Photo Processing

- Orientation correction
  - EXIF (Exchangeable Image File Format) metadata
- Photo Filtering
  - Blur detection
    - Check edge information in diff. resolutions
  - Overexposure/Underexposure detection
    - Check intensity information of each photo

Blurred photo

Underexposure photo

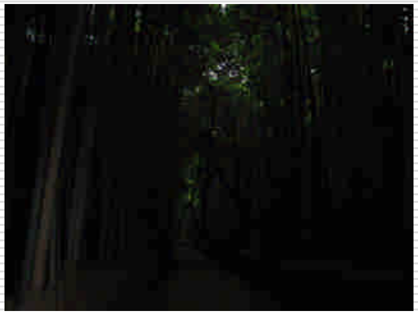


Photo Clustering

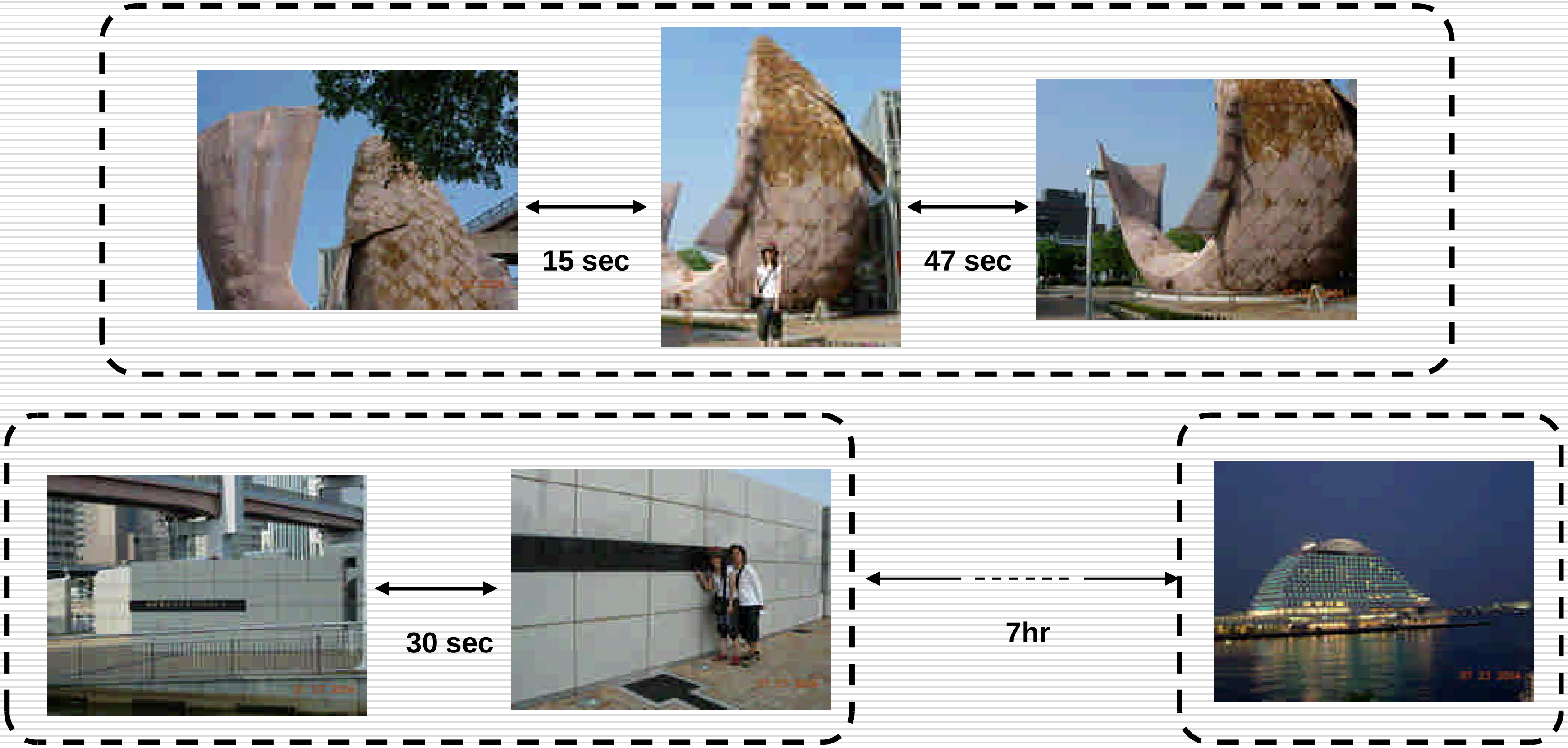
- Displaying photos that are in the same scenic spot or the same event would strengthen audiovisual perception.
- Clustering
  - Time characteristics – event
  - Content characteristics – visually homogenous

(O)

(X)

Time-based Clustering

- Check the time gap between adjacent photos

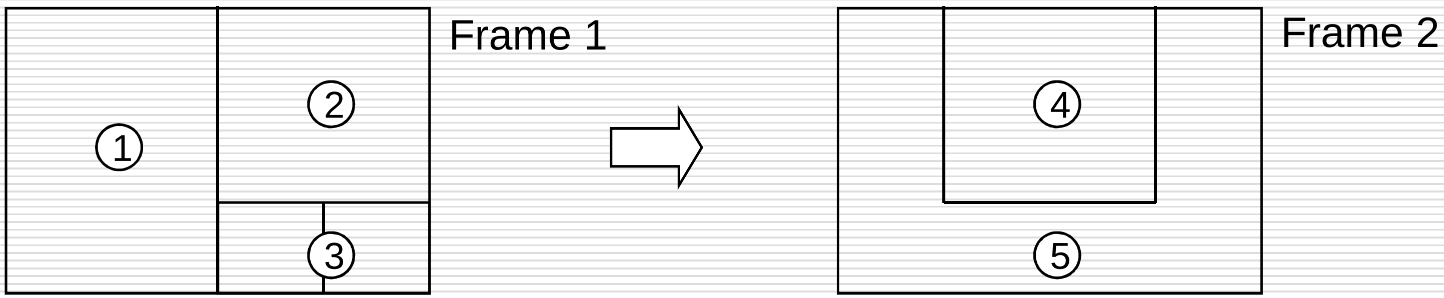


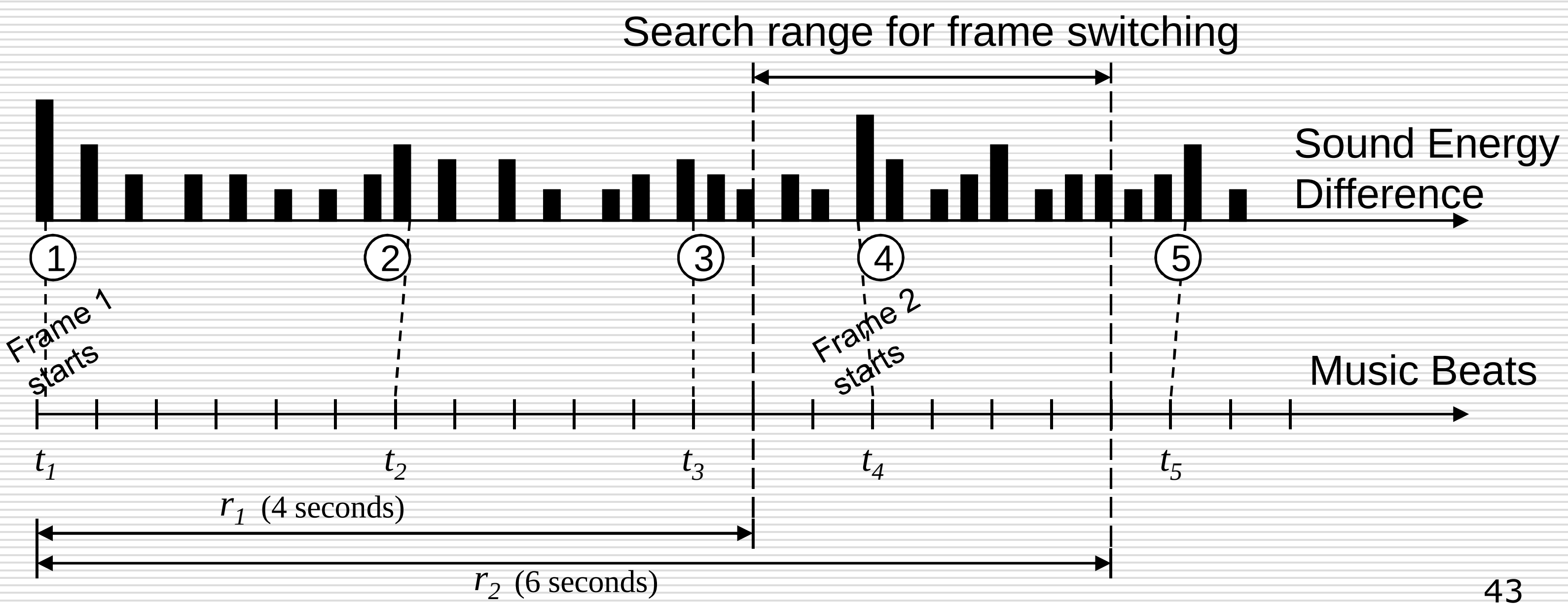
Content-based Clustering

Music Analysis

- Beat detection
- Music segmentation

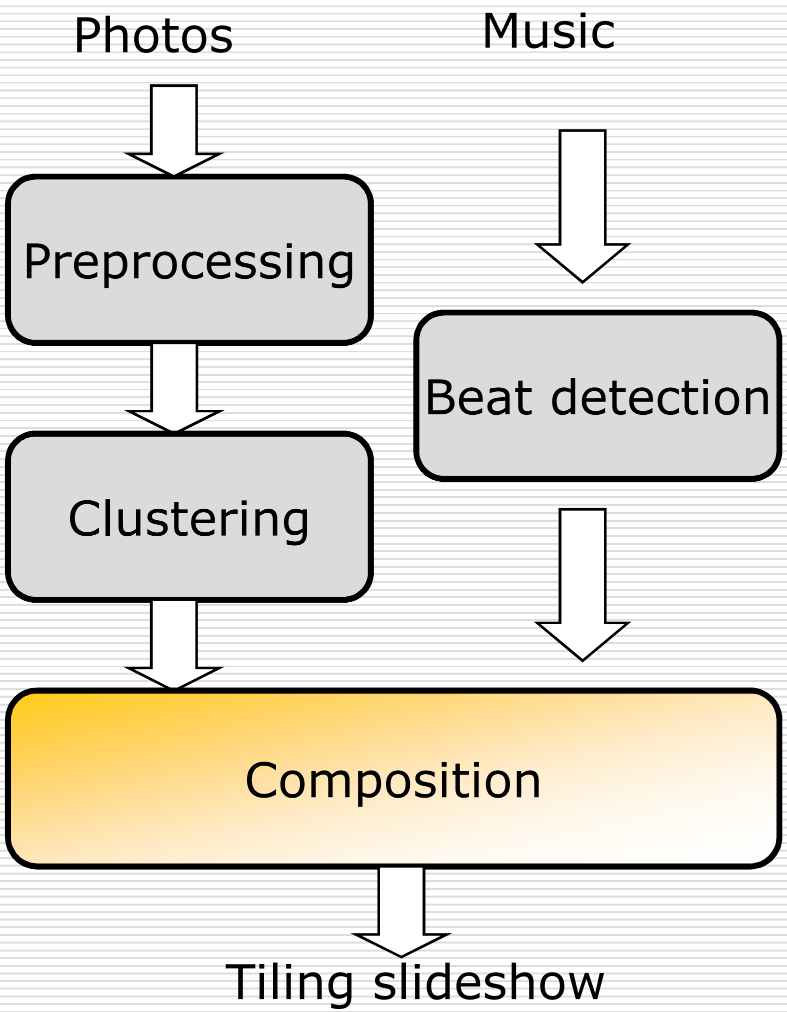
For frame switching and photo displaying





Short Summary

- Photo
  - Filter out defective photos
  - Organize photos in terms of time and content characteristics
- Music
  - Segment into smaller pieces



Tiling Slideshow Composition

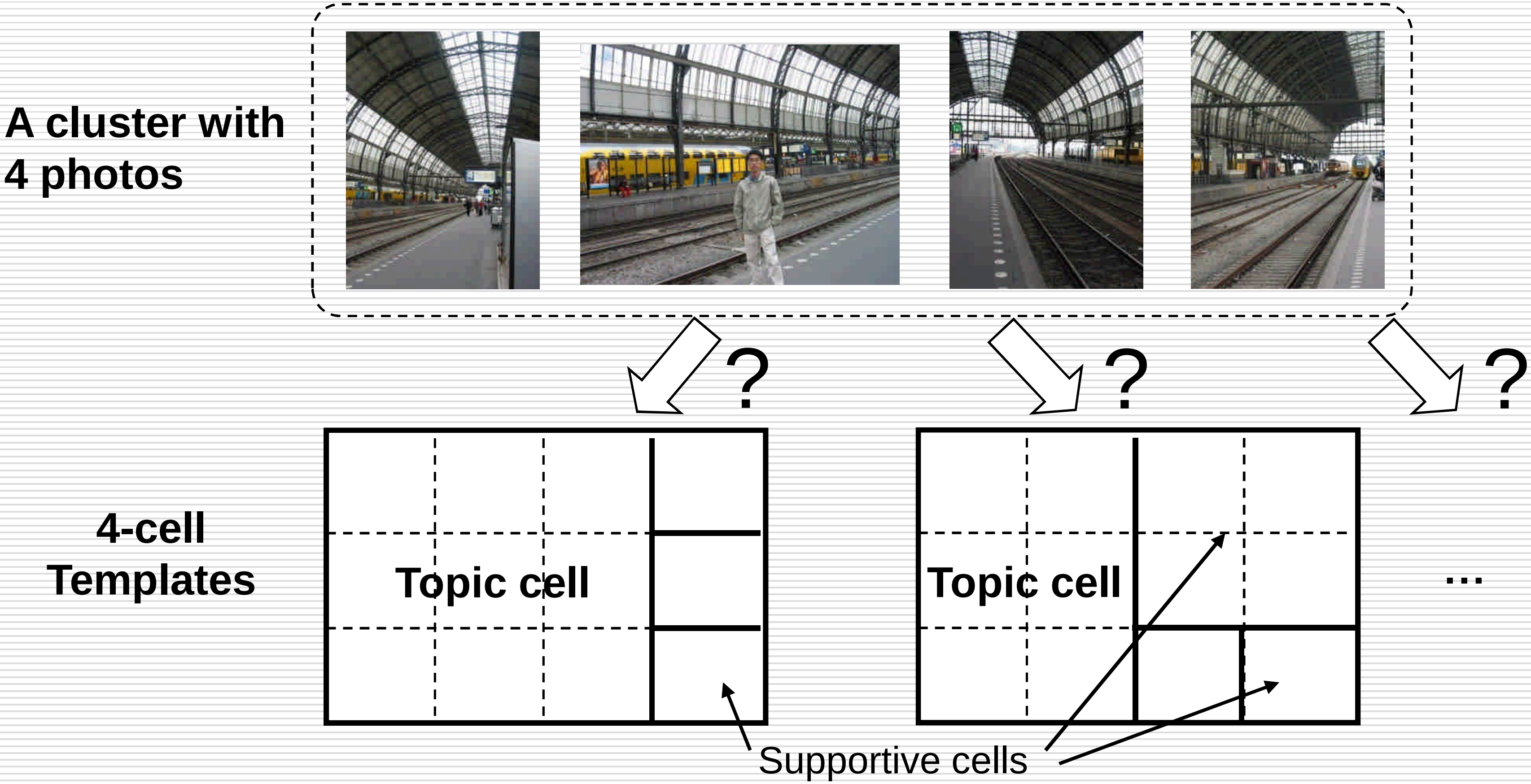
- **Challenge 1**
  - Given a time-limited music clip, only a subset of photo clusters can be displayed.
- **Challenge 2**
  - For a cluster of photos to be displayed, more important photos should occupy larger space.
- **Challenge 3**
  - Photos should be smartly manipulated to fit in with the limited displaying space.

Cluster Selection (for Challenge 1)

- Cluster-based importance
  - Defined based on “photo per minute (PPM)” and “photo conformance (PC)”
  - Higher shooting frequency (PPM), more important
  - Larger conformance (PC), more important

Templates (for Challenge 2)

- Design various templates that contain a topic cell and several supportive cells – to form photographic paragraphs.



Template Determination (for Challenge 2)

- Templates importance

$$Ic_i = Area(Tc_i)/Area(T) \qquad \left(Ic_1 \geq Ic_2 \geq \ldots \geq Ic_k\right)$$

$$TV = \left(Ic_1, Ic_2, \ldots, Ic_k\right)$$ – Template importance vector

- Photo-based importance
  - Defined based on "face region (FR)" and "attention value (AV)"

$$PI_i = W_{face} \times FR(P_i) + W_{attention} \times AV(P_i)$$

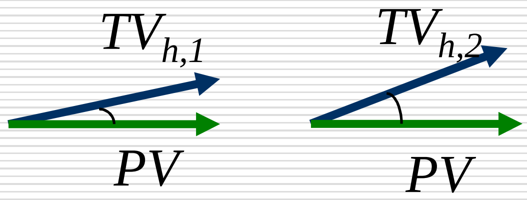
$$PV = \left(PI_1, PI_2, \ldots, PI_k\right)$$ – Photo importance vector

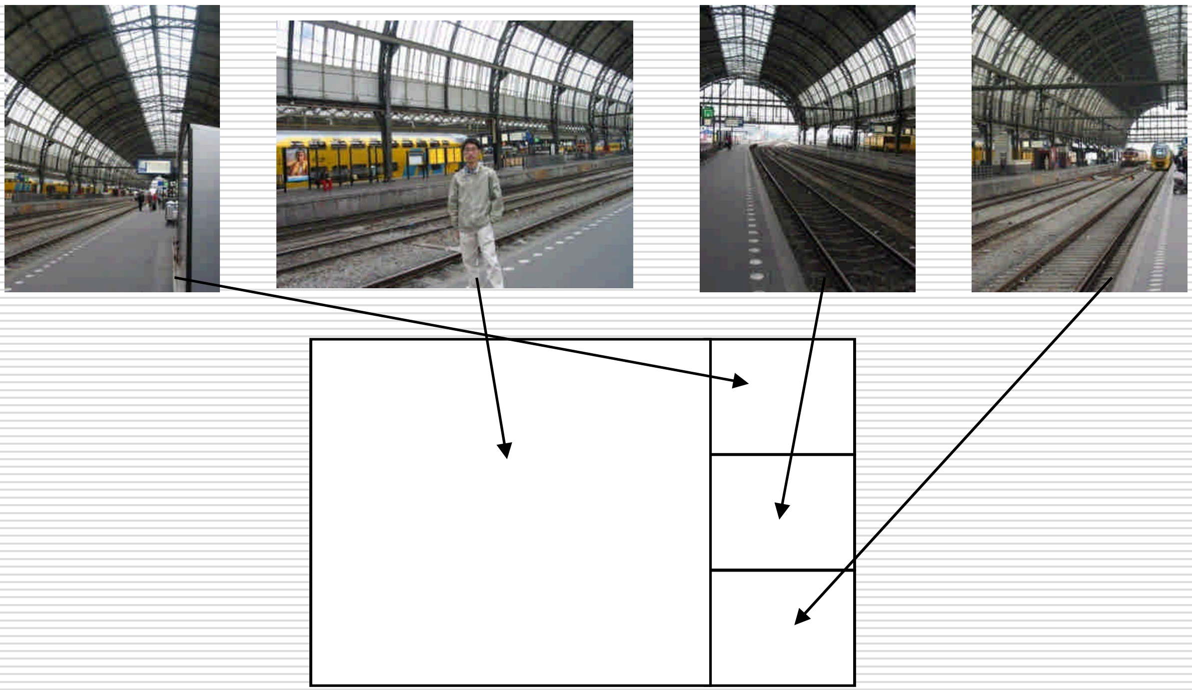
$$\left(PI_1 \geq PI_2 \geq \ldots \geq PI_k\right)$$

Template Determination (for Challenge 2)

- Find the most matching between template importance and photo importance
  - Find the minimum included angle between them

$$T_{h,i} = \arg\min_{i=1,2,\ldots,s} \operatorname{acos}\left(\frac{PV \cdot TV_{h,i}}{\| PV \| \| TV_{h,i} \|}\right)$$

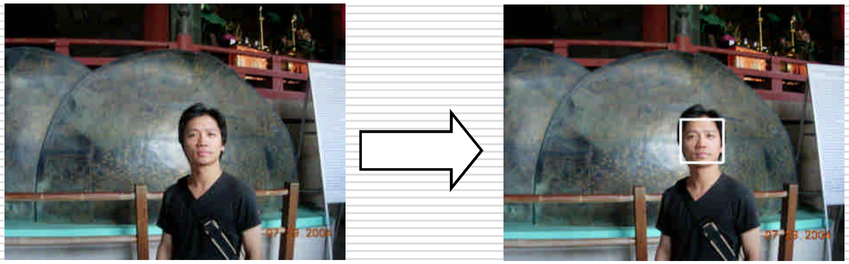




Composition (for Challenge 3)

- Find the region that conveys most “content value” and conforms to the aspect ratio of the targeted cell – constrainted optimization problem.

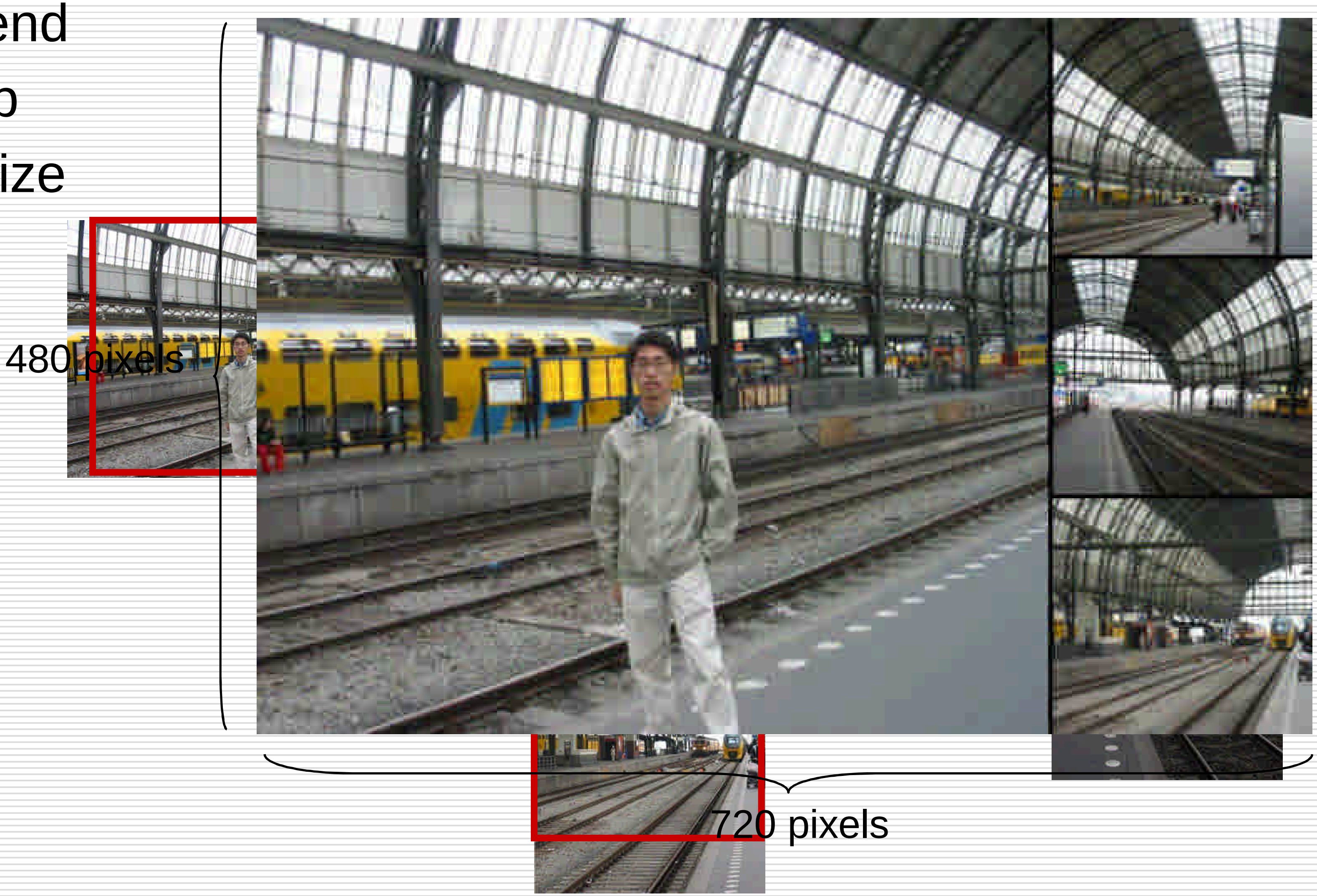
Top-down case:
(photo with face)



Bottom-up case:
(photo without face)

Composition (for Challenge 3)

1. Find ROI
2. Extend
3. Crop
4. Resize



Demo

Summary

- Semantic analysis
  - We propose a framework to bridge the semantic gap.

Domain	Mid-level representation	Modality
Semantic analysis in baseball videos	Caption information, shot context, key phrases	Video, speech
Semantic analysis in tennis videos	Court information, player, audio events	Video, audio

- Collaborative presentation
  - We propose a new type of audiovisual presentation for **consumer photos**.
  - Perform both visual and music analysis for **organized presentation**.

Open Problems

- Semantic gap problem is still unsolved.
- Web 2.0 for multimedia research.
- Social network in multimedia research.
- Knowledge discovery in multimedia content.

Thank You

朱威達 (Wei-Ta Chu)
wtchu@cs.ccu.edu.tw
http://www.cs.ccu.edu.tw/~wtchu