Rushes Summarization by *Object* and *Event* Understanding

Feng Wang and Chong-Wah Ngo

City University of Hong Kong

Rushes Summarization

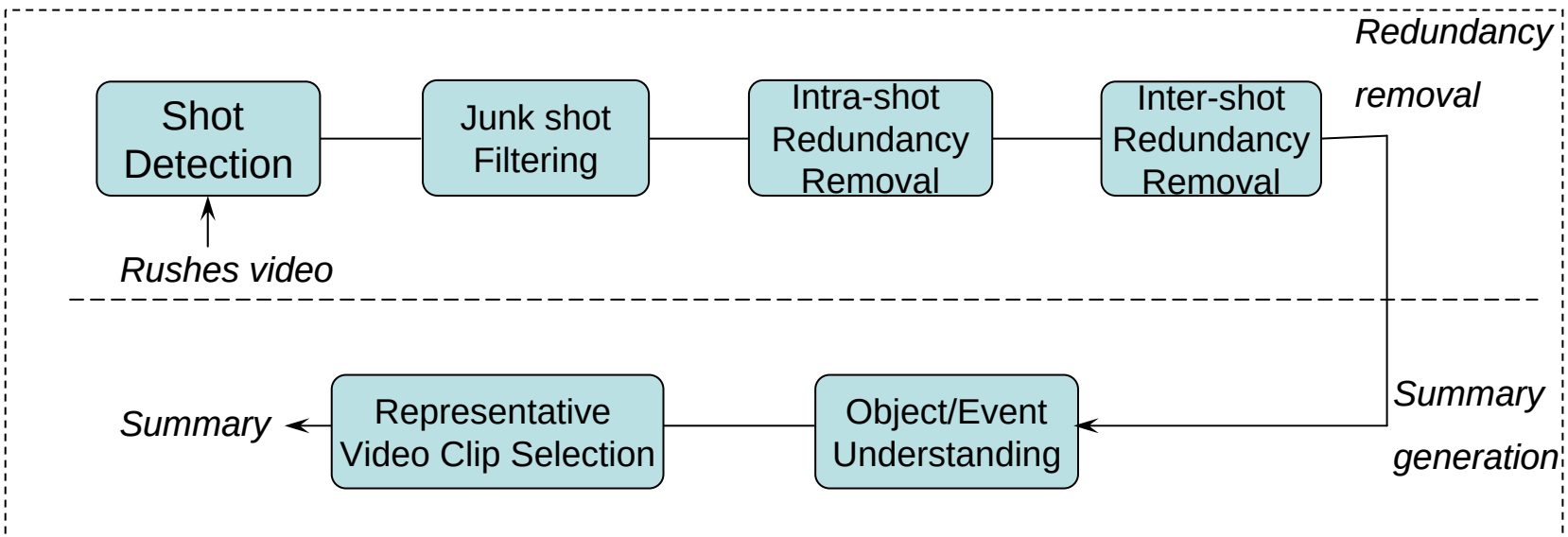
- Movie product
  - Captured by profession cameraman
  - Edited by expert
- Home video
  - Many junk shots, intermediate camera motion
  - Unedited

- *Rushes …?*
  - Like a *mixture* of movie product and home video

- We focus on:
  - *Object* / *Event* detection and understanding
  - *Audio-visual representability score* for clip selection

Overview

- Redundancy removal
  - *Junk shot:* color bars, grayscale/black frames
  - *Intra-shot:* story-irrelevant scenes
  - *Inter-shot:* retakes
- Summary generation
  - *Object / Event* detection and understanding
  - Select the most representative video clips



Junk Shot Filtering

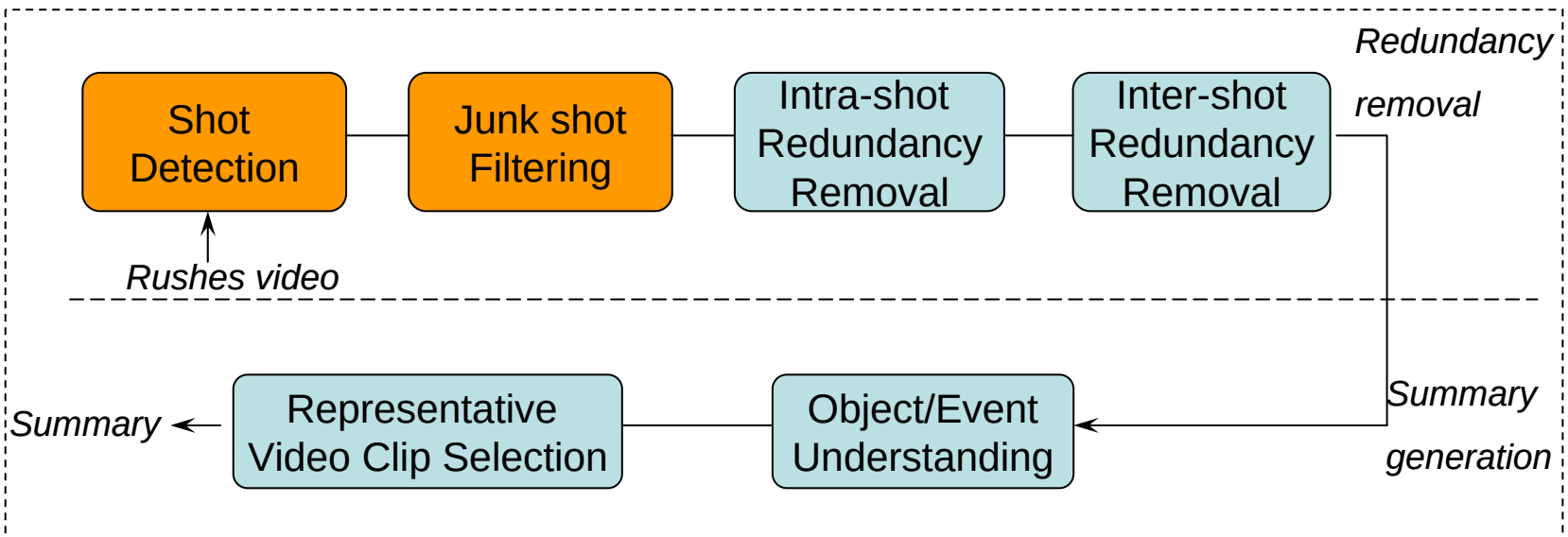
- Shot detection
  - C. W. Ngo et. al., “Video Partitioning through Temporal Slices Analysis”, *CSVT* 2001.

- Junk shots

Overview

- Redundancy removal
  - *Junk shot:* color bars, grayscale/black frames
  - *Intra-shot:* story-irrelevant scenes
  - *Inter-shot:* retakes

- Summary generation
  - *Object / Event* detection and understanding
  - Select the most representative video clips



Intra-Shot Redundancy Removal

- Separate movie storytelling from irrelevant scenes
  - ✓Cut-board scene
  - ✓Camera motion
  - ✓Audio

Cut-board Detection

- Near-Duplicate Keyframe Matching
  - W. L. Zhao, C. W. Ngo, H. K. Tan, and X. Wu, "Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning", *IEEE Trans. on Multimedia*, 2007.

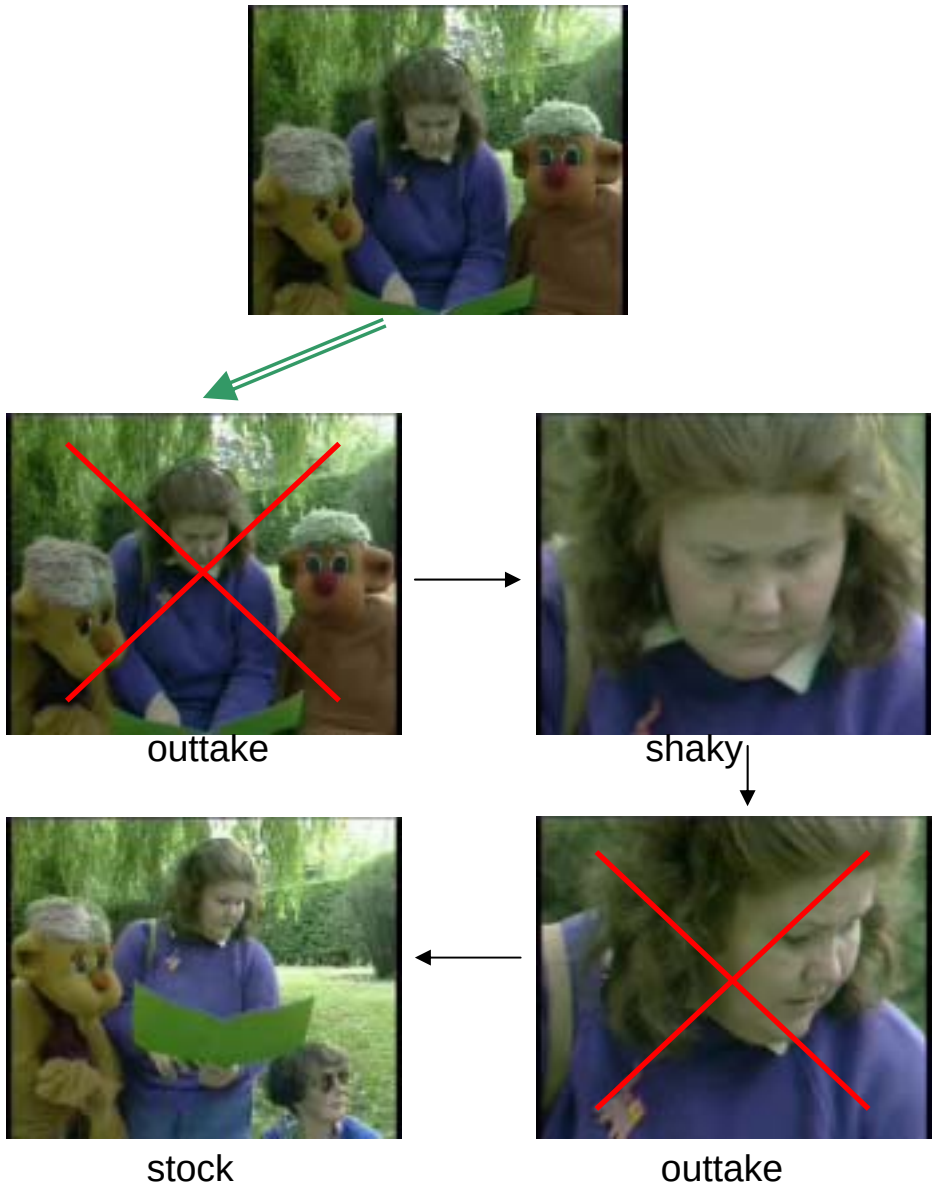
Detected cut-board scenes in test videos

Example cut-board scenes from development set

- Speech Recognition
  - "Shot xx", "Take yy"

Unintentional Camera Motion Filtering

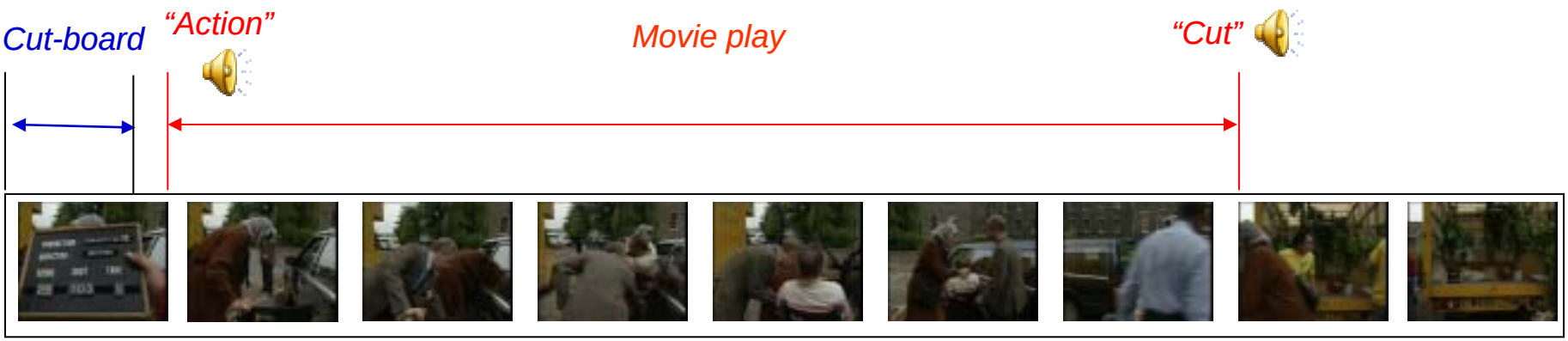
- *Stock:*
  - Intentional
  - Useful
- *Outtake:*
  - Intermediate
  - To be discarded
- *Shaky:*
  - Either useful or not useful



– C. W. Ngo, Z. Pan & X. Y. Wei, “Hierarchical Hidden Markov Model for Rushes Structuring and Indexing”, *CIVR 2006.*

Audio

- Speech recognition
  - *“Action”, “Cut”, “Standby”*
- Audio scene changes
  - Cepstralflux, multi-channel cochlear decomposition, cepstral vector, low energy fraction, volume standard deviation, non-silence ration, pitch and zero crossing rate
    - *Silence, Actor’s lines, Noise*

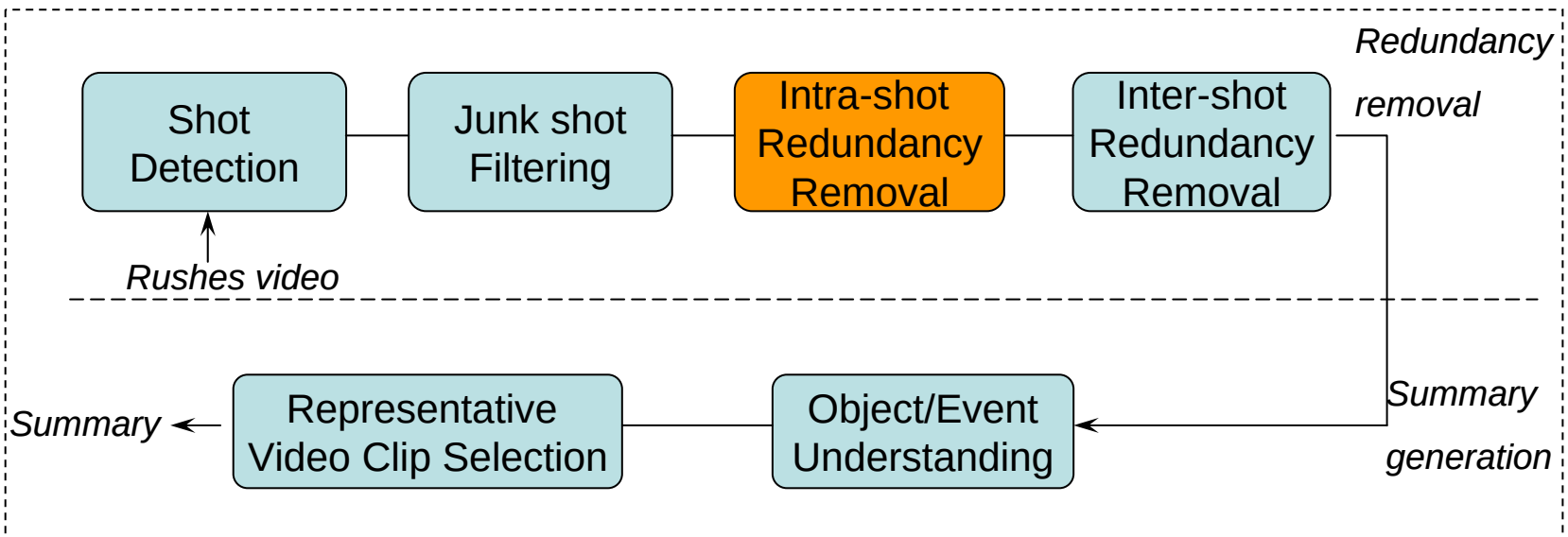


Video “MS044500”

Overview

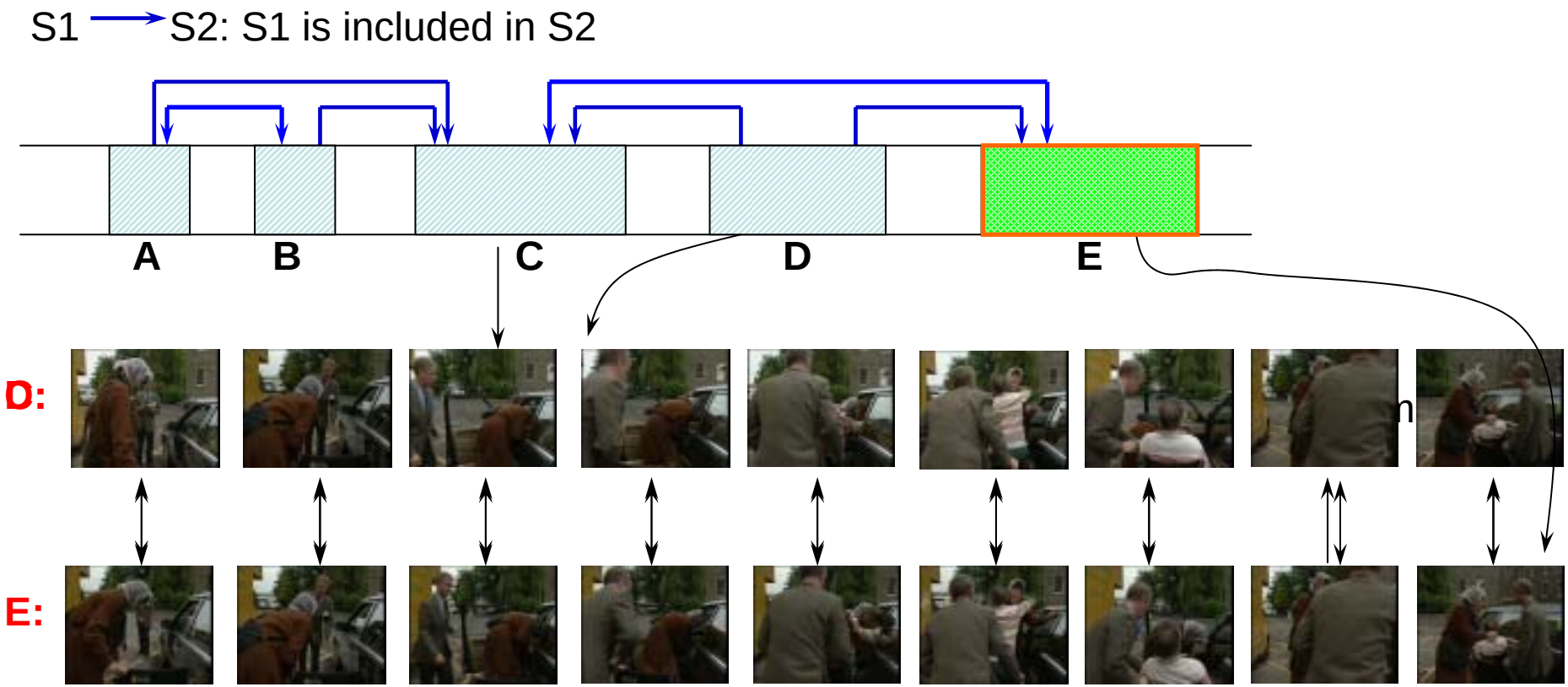
- Redundancy removal
  - *Junk shot:* color bars, grayscale/black frames
  - *Intra-shot:* story-irrelevant scenes
  - *Inter-shot:* retakes

- Summary generation
  - *Object / Event* detection and understanding
  - Select the most representative video clips



Inter-shot Redundancy Removal

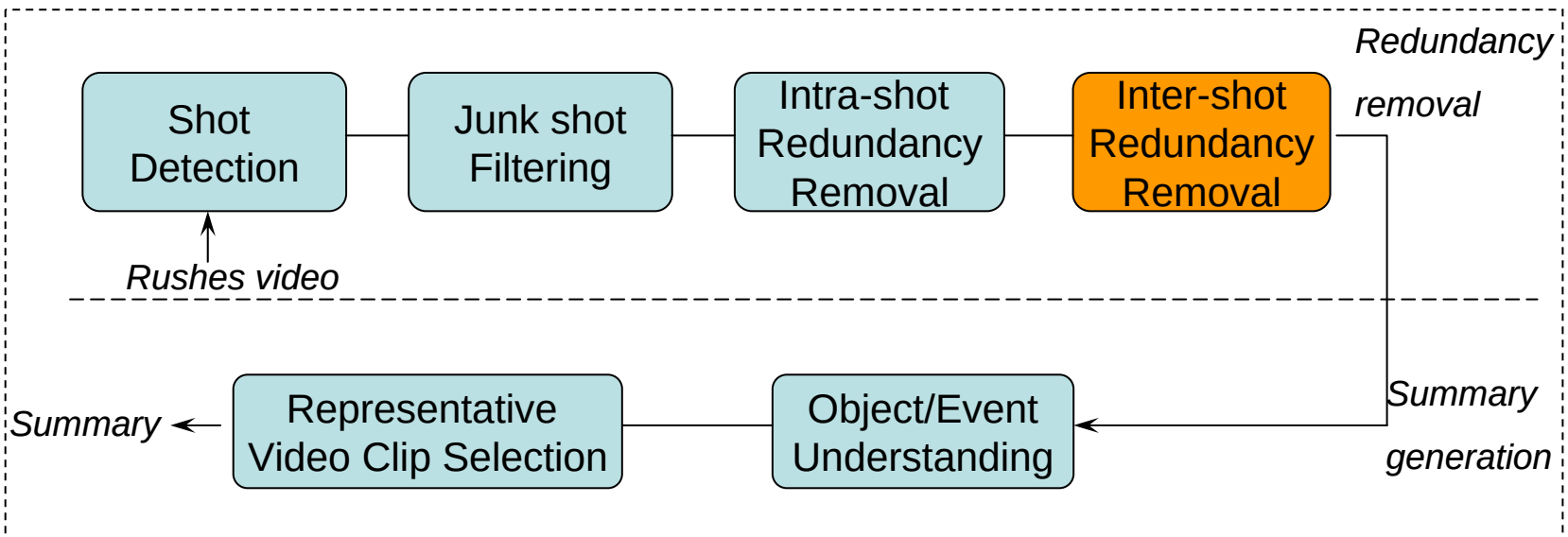
- Keyframe matching
  - Remove incomplete shots
  - Heuristically select the last one of many retakes



Overview

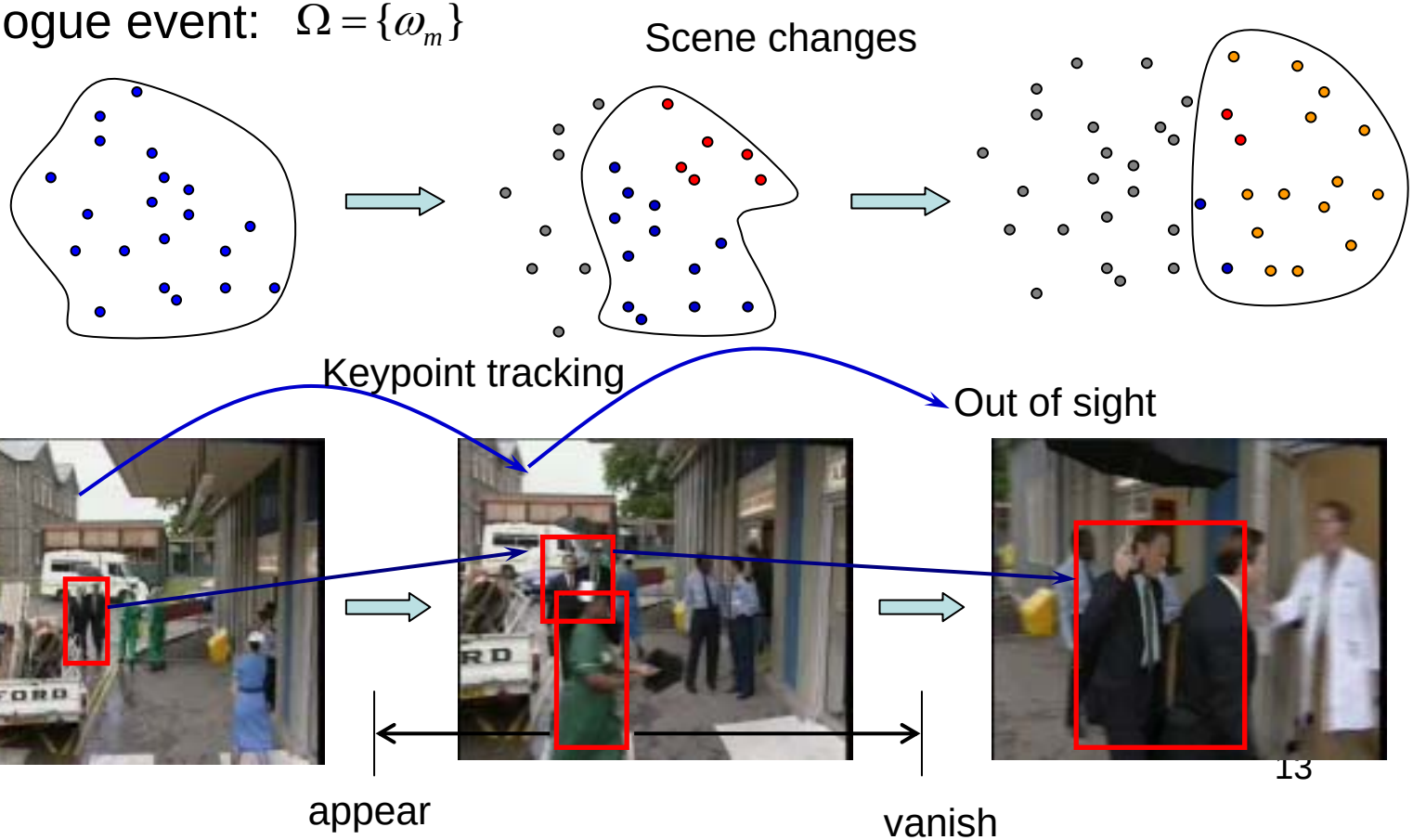
- Redundancy removal
  - *Junk shot:* color bars, grayscale/black frames
  - *Intra-shot:* story-irrelevant scenes
  - *Inter-shot:* retakes

- Summary generation
  - *Object / Event* detection and understanding
  - Select the most representative video clips



Object and Event Detection

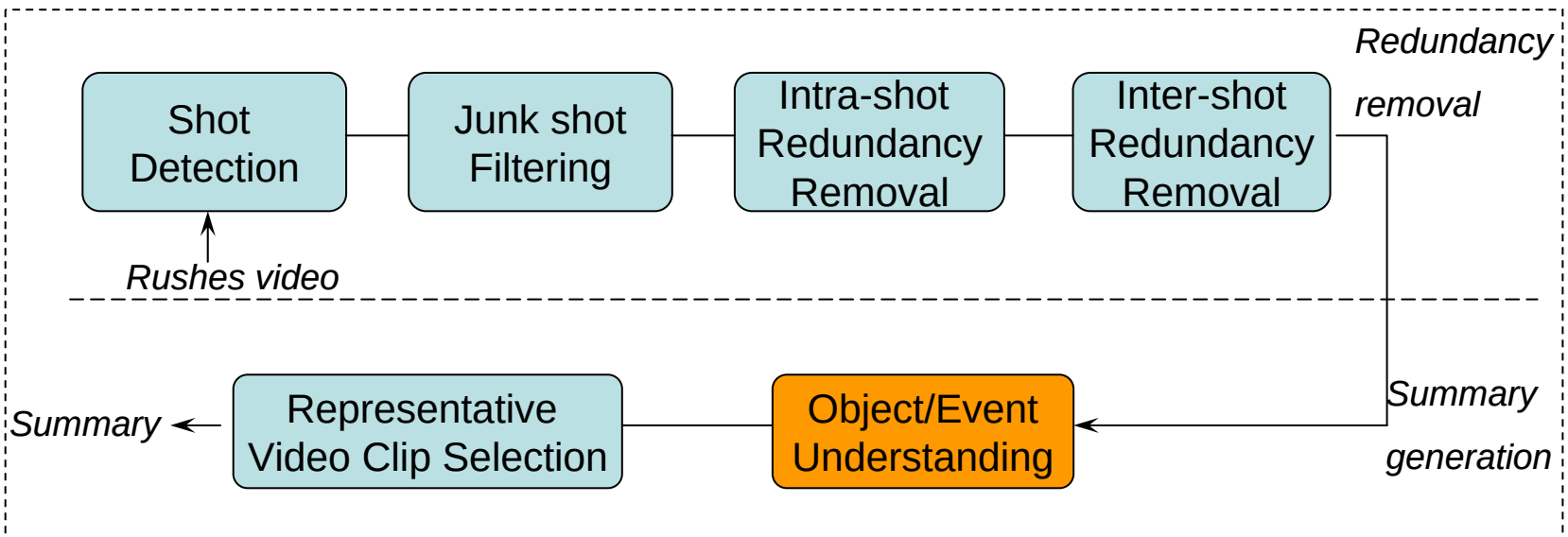
- A set of objects: $O = \{o_i\}$
- Object motion activities: $\Phi = \{\varphi_j\}$
- Scene changes: $\Delta = \{\delta_t\}$
- Camera motion: $\Gamma = \{\gamma_k\}$
- Speech/dialogue event: $\Omega = \{\omega_m\}$



Overview

- Redundancy removal
  - *Junk shot:* color bars, grayscale/black frames
  - *Intra-shot:* story-irrelevant scenes
  - *Inter-shot:* retakes

- Summary generation
  - *Object / Event* detection and understanding
  - Select the most representative video clips



Video Clip Representability

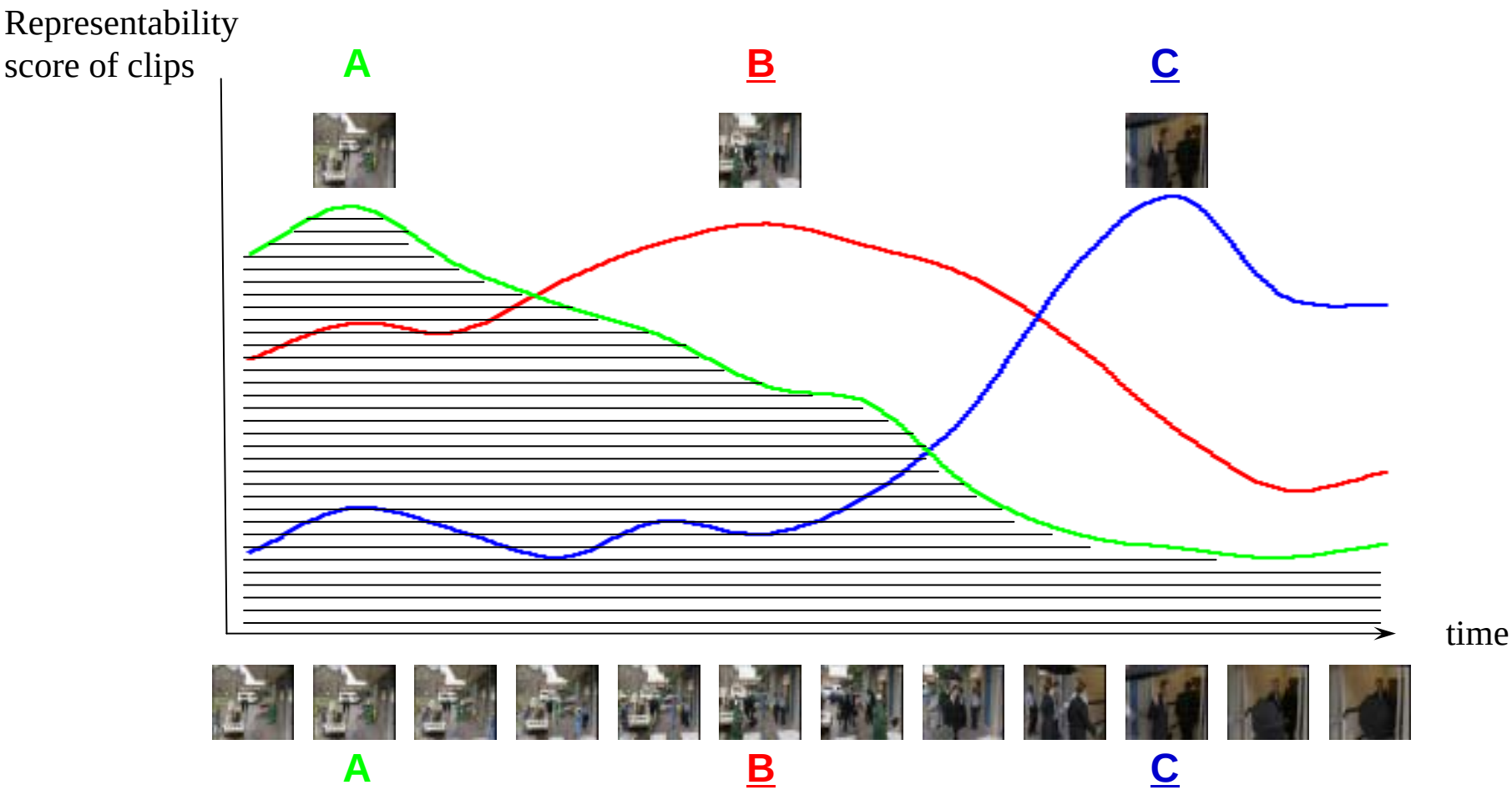
- The representability of a video clip v for the objects and events at time (t_1, t_2)

$$
\begin{aligned}
R_v(O) &= \sum_{o \in O} \int_{t_1}^{t_2} (1 - \frac{|t - (t_{so} + t_{eo})/2|}{t_{eo} - t_{so}}) dt \\
R_v(\Phi) &= \frac{\sum_{\phi \in \Phi} \int_{t_1}^{t_2} f(t) dt}{\sum_{\phi \in \Phi} \int_{t_{s\phi}}^{t_{e\phi}} f(t) dt} \\
R_v(\Gamma) &= \sum_{\gamma \in \Gamma} \int_{t_1}^{t_2} (1 - \frac{|t - (t_{s\gamma} + t_{e\gamma})/2|}{t_{e\gamma} - t_{s\gamma}}) dt \\
R_v(\Delta) &= \frac{\int_{t_1}^{t_2} \delta(t) dt}{\int_{t_{s\delta}}^{t_{e\delta}} \delta(t) dt} \\
R_v(\Omega) &= \frac{\|W(v) \bigcap W(\Omega)\|}{\|W(\Omega)\|}
\end{aligned}
$$

- The representability of video clip v_i for v_j

$$
\begin{aligned}
Rep(v_i, v_j) &= \frac{1}{\sqrt[4]{d(v_i, v_j)}} \cdot (w_O R_{v_i}(O_{v_j}) + w_\Phi R_{v_i}(\Phi_{v_j}) \\
&+ w_\Gamma R_{v_i}(\Gamma_{v_j}) + w_\Delta R_{v_i}(\Delta_{v_j}) + w_\Omega R_{v_i}(\Omega_{v_j}))
\end{aligned}
$$

Representative Clip Selection



Result

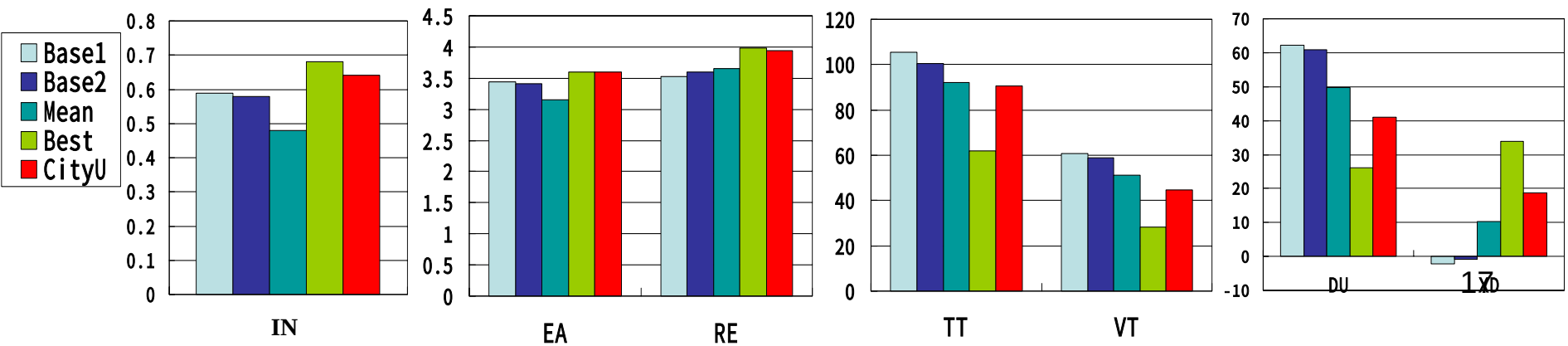
Video summary for the rushes video ‘MRS044500’:

DU	36.50
XD	14.80
TT	66.67
VT	38.00
IN	0.42
EA	3.00
RE	2.67

Object: 85

Event: 134

TRECVID Evaluation:



Conclusion

- *Object and event understanding*
  - provide a promising approach for *semantic-based video summarization*
  - excellent at describing storyline of movies
- *Domain knowledge*
  - redundancy removal: outtake, shaky, cut-board, speech, …..
- *Generalization to other domains*
  - object/event understanding
  - representability score
- *Future work*
  - more on event analysis

Thanks!

Q&A?