SEGMENTATION OF TV SHOWS INTO SCENES

USING SPEAKER DIARIZATION AND SPEECH RECOGNITION

Context

- Exponential growth of video content
  - Mostly TV and web
- Content-based video indexing
  - Query/browse by people
    - REPERE challenge
  - Query/browse by semantic concept
    - GdR ISIS *IRIM* project
    - NIST TRECVid Semantic Indexing task
- Make content *consumption* easier
  - Automatic summarization
    - PhD thesis with IRIT (Ph. Ercolessi)

(More) Context

- Automatic summarization of TV series
  - At the episode level
    - **Shot segmentation**
    - **Scene segmentation**
    - **Plot (or substory) deinterlacing**
    - **Episode summarization**
      - « Previously, on Lost… »
      - Browse by plot
  - At the collection level
    - Cross-episode plot
    - Episode summarization wrt. whole collection

Shot, scene, sequence and plot

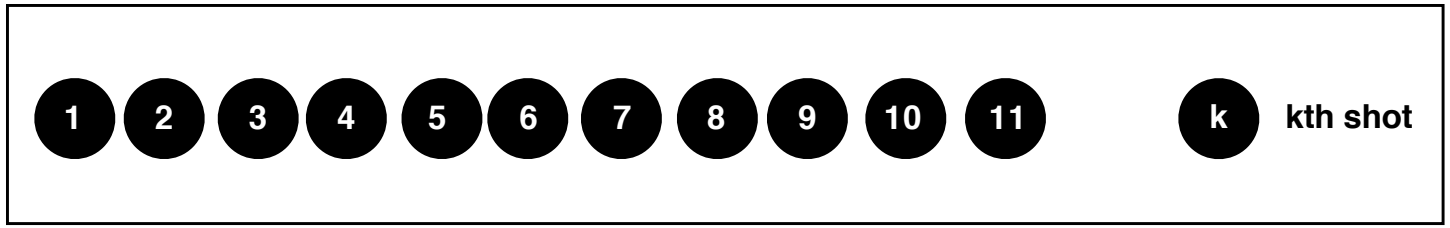
- Shot
  - a part of a film between two camera cuts
- Scene
  - *(each author working on scene segmentation uses its own definition)*
  - group of consecutive shots
  - temporal continuity / unity of time
  - semantic coherence / unity of action
- Sequence
  - group of consecutive scenes
  - semantic coherence
  - aka story or topic in TV news
    - made of multiple stories introduced by the anchor
- Plot
  - group of (not necessarily consecutive) sequences
  - modern TV show episodes usually have multiple interlaced plots
  - two plots can overlap

Outline

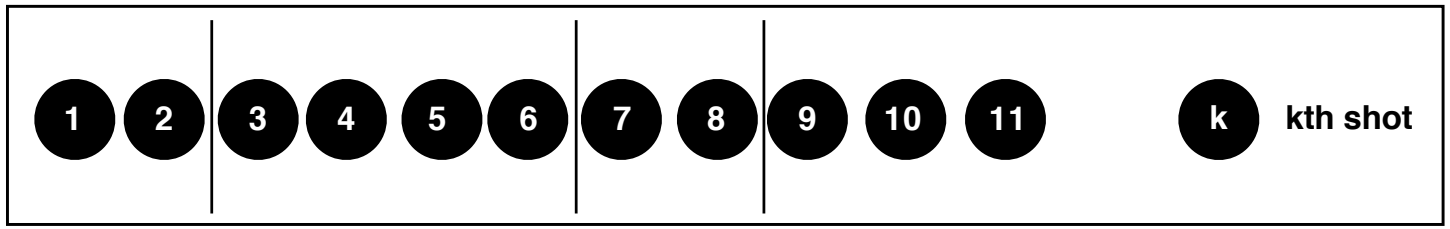
- Context
- Definitions & notations
- Principle
  - Scene transition graph with color histograms
  - Generalized STG
- Multimodal extension
  - Speaker diarization & speech recognition
  - Multimodal fusion

Scene segmentation

- Input: shot boundaries

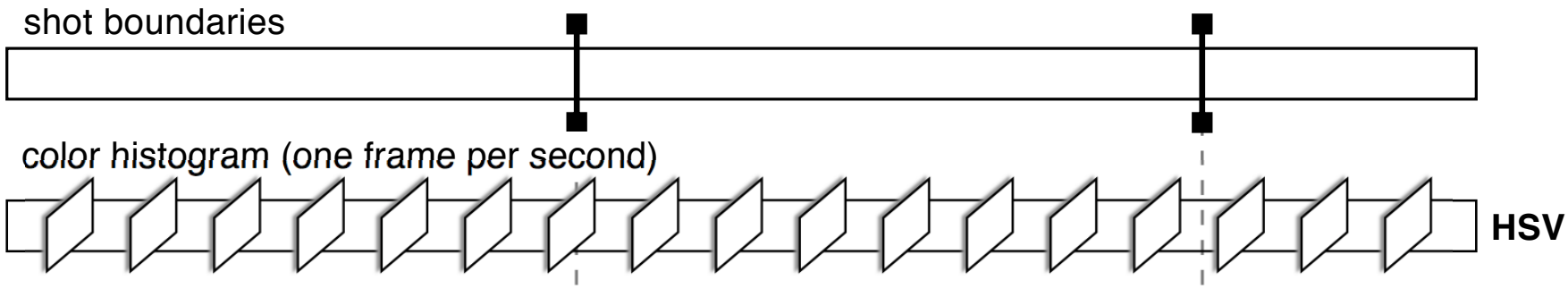


- Output: scene boundaries



- Classification problem on shot boundaries
  - precision, recall, F_1-measure

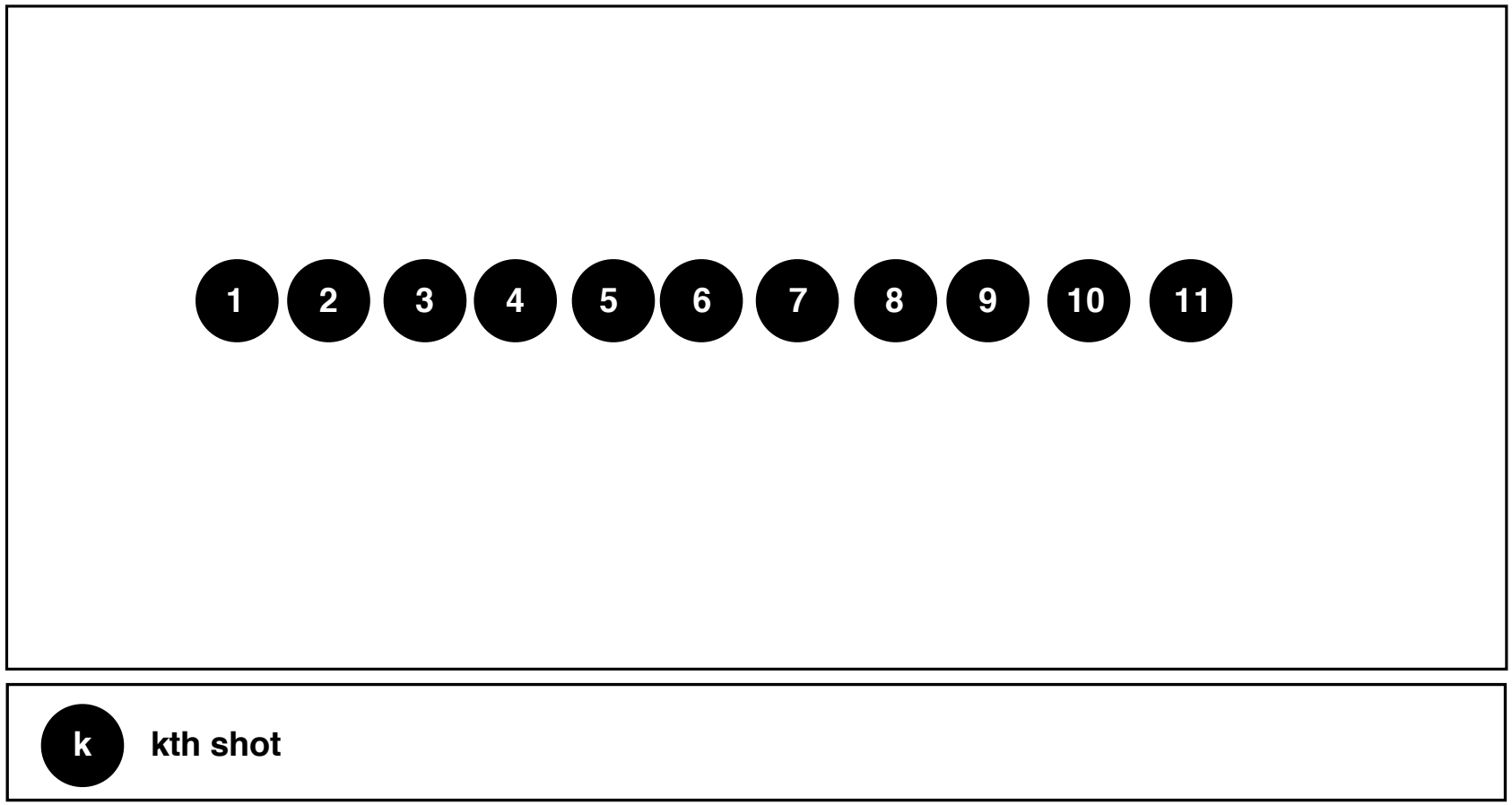
Color



- Multiple frame per shot, one color histogram per frame
- d_{ij}^{HSV}: minimum distance between all possible pairs of histograms from shots i and j

Scene transition graph

Segmentation of Video by Clustering and Graph Analysis / **Yeung (1998)**



Notations

d_{ij} dissimilarity between shots *i* and *j*

t_{ij} temporal distance between shots *i* and *j*

D_{ij} combined distance between shots *i* and *j*

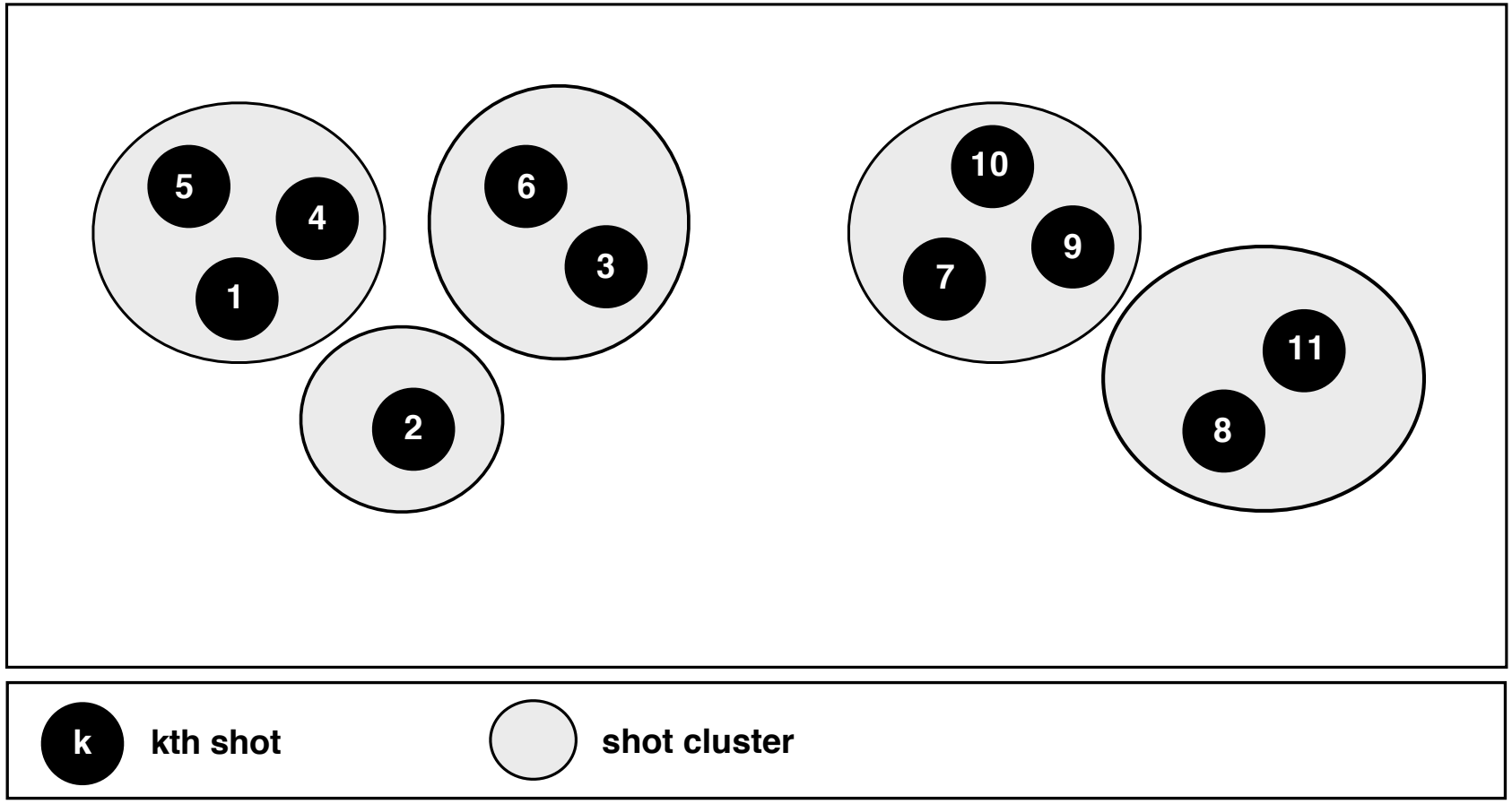
$$D_{ij} = \begin{cases} d_{ij} & \text{if } t_{ij} < \Delta_t \\ +\infty & \text{otherwise} \end{cases}$$

Δ_t temporal distance threshold

Δ_d combined distance threshold

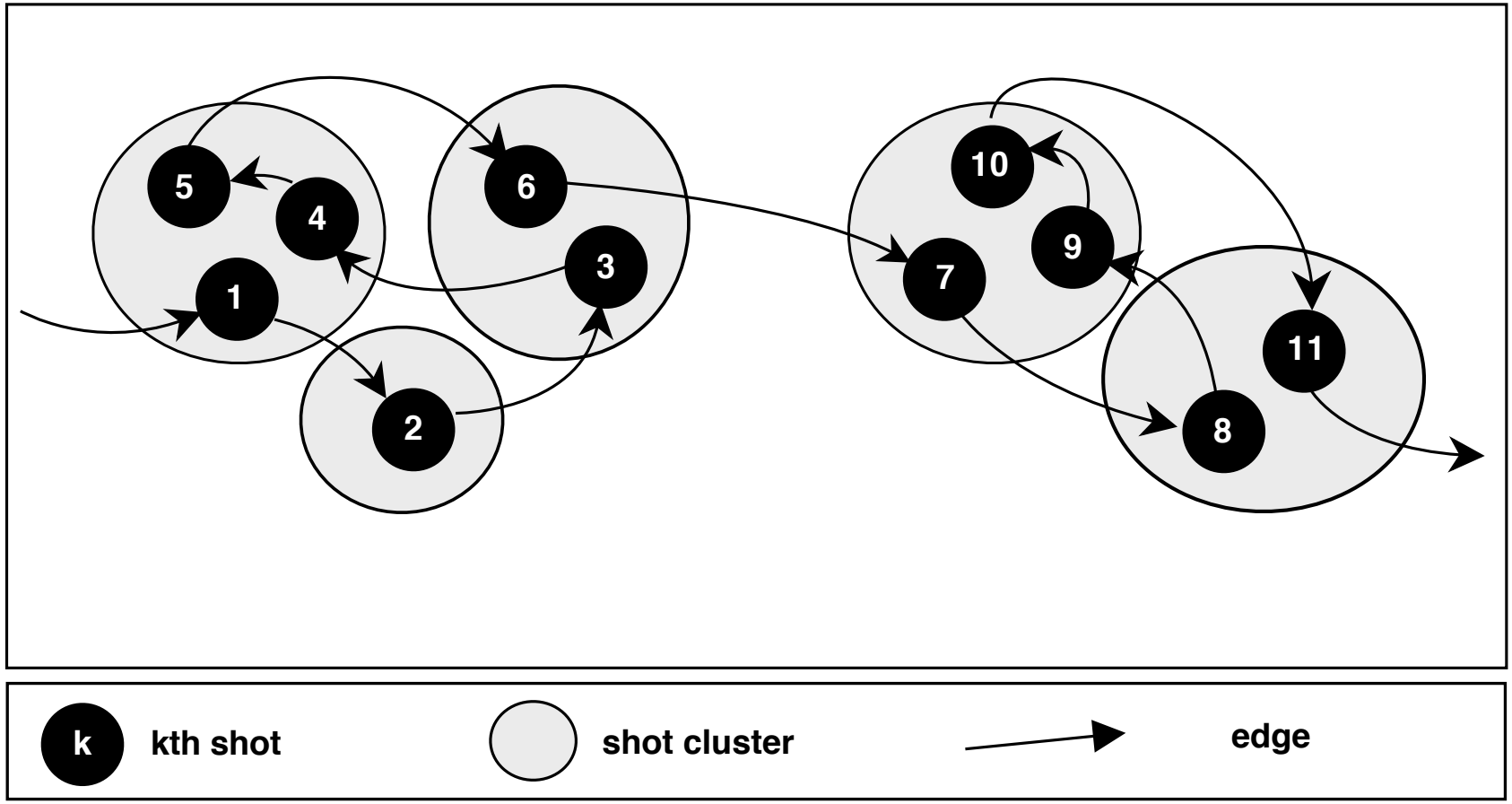
Scene transition graph

- Step 1: complete-link agglomerative clustering Δ_d



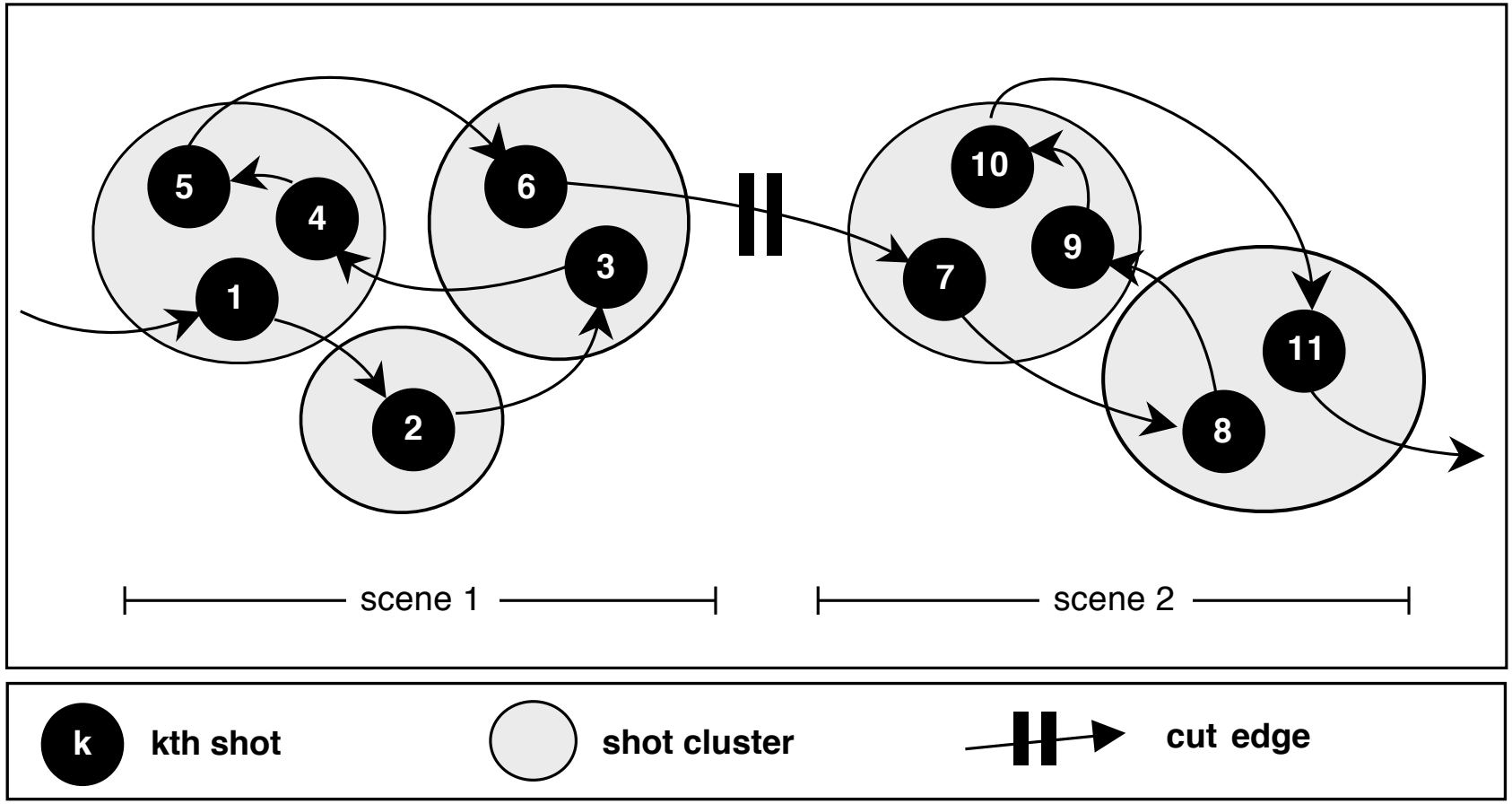
Scene transition graph

- Step 2: scene transition graph



Scene transition graph

- Step 3: cut-edge detection



HSV/STG Results

- Corpus
  - First eight episodes of *Ally McBeal* TV shows
  - 5 hours of videos, 5564 shots and 306 scenes
- Evaluation
  - Leave-one-episode-out cross validation
  - Two thresholds Δ_t and Δ_d

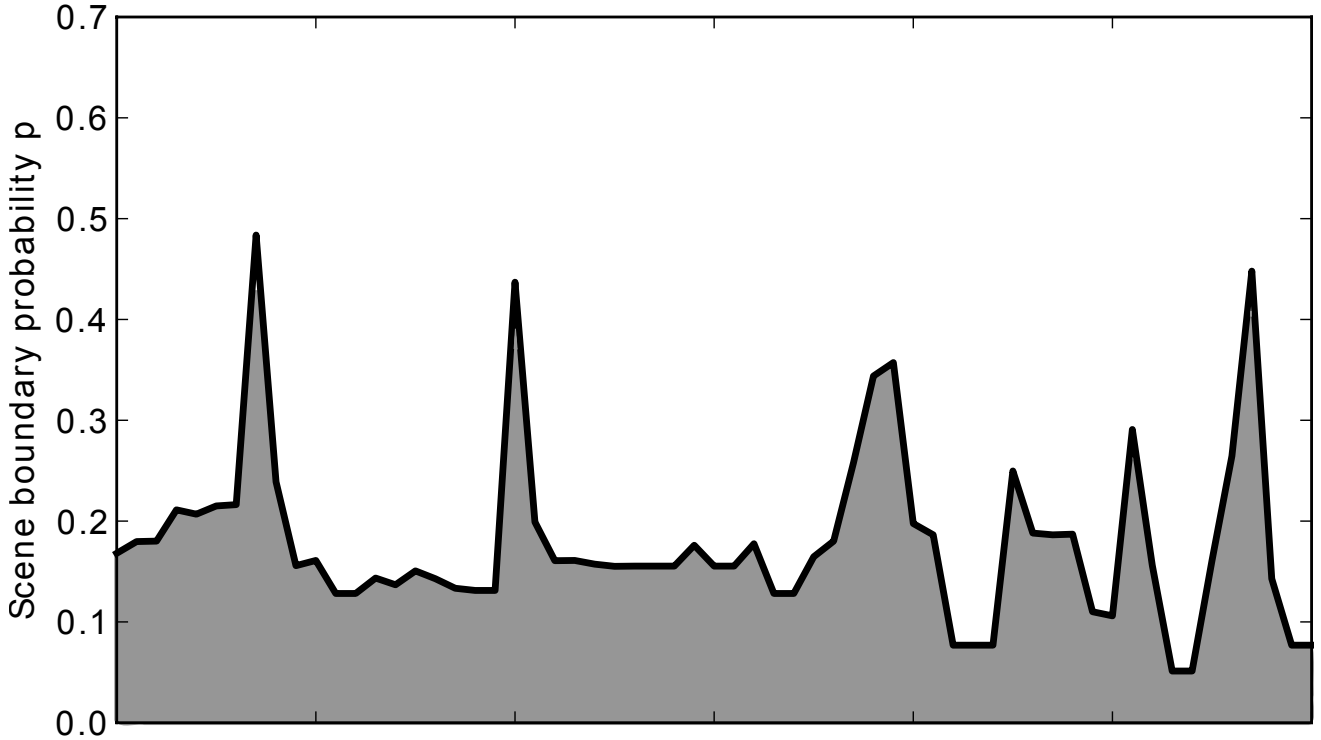
	Precision	Recall	F-Measure	# Scenes
HSV (STG)	0.256	0.533	0.449	461

Scene transition graph

- Limitation:
  - Every pair (Δ_t, Δ_d) leads to a different set of detected scene boundaries.
  - The optimal values are very dependent on the video
- Proposition (Sidiropoulos, 2011):
  - Generalized STG

Generalized STG

- Large set of STGs by selecting random Δ_t and Δ_d
- Scene boundary probability

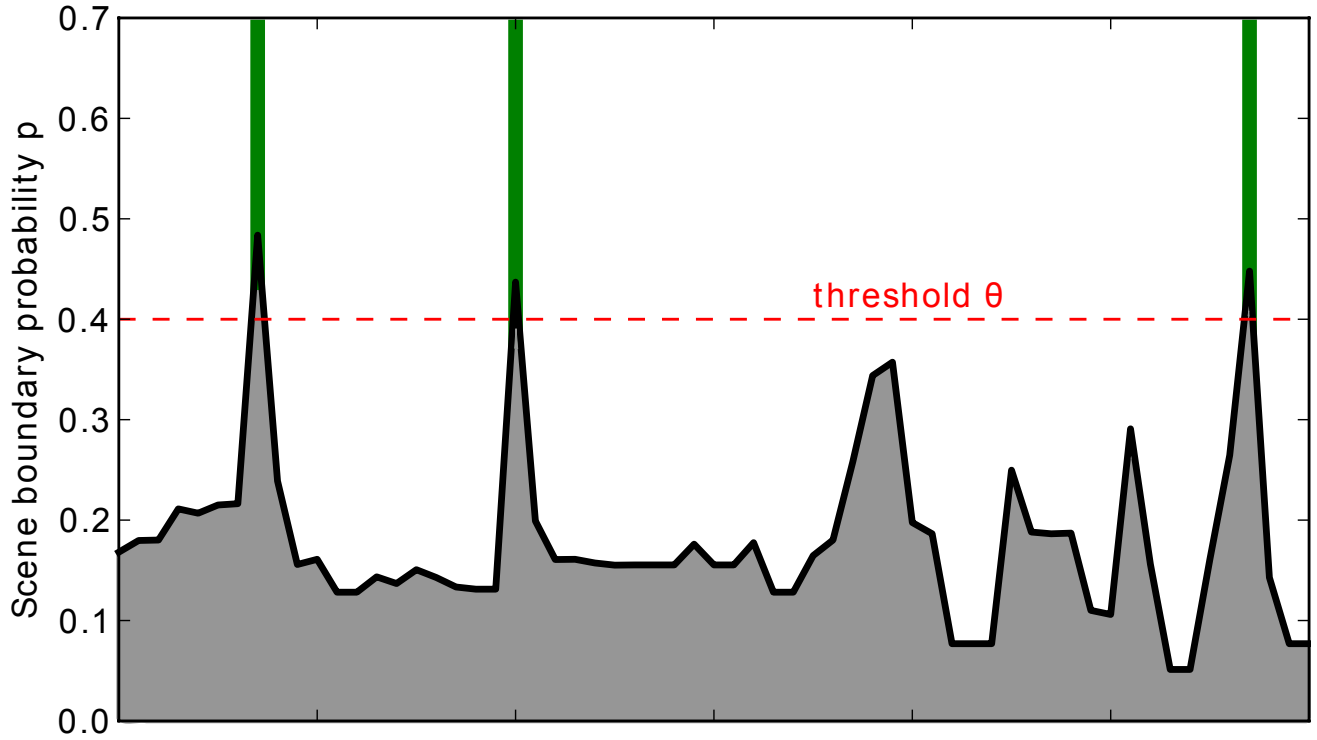


Generalized STG

– Large set of STGs by selecting random Δ_t and Δ_d

– Scene boundary probability

– Unique threshold θ

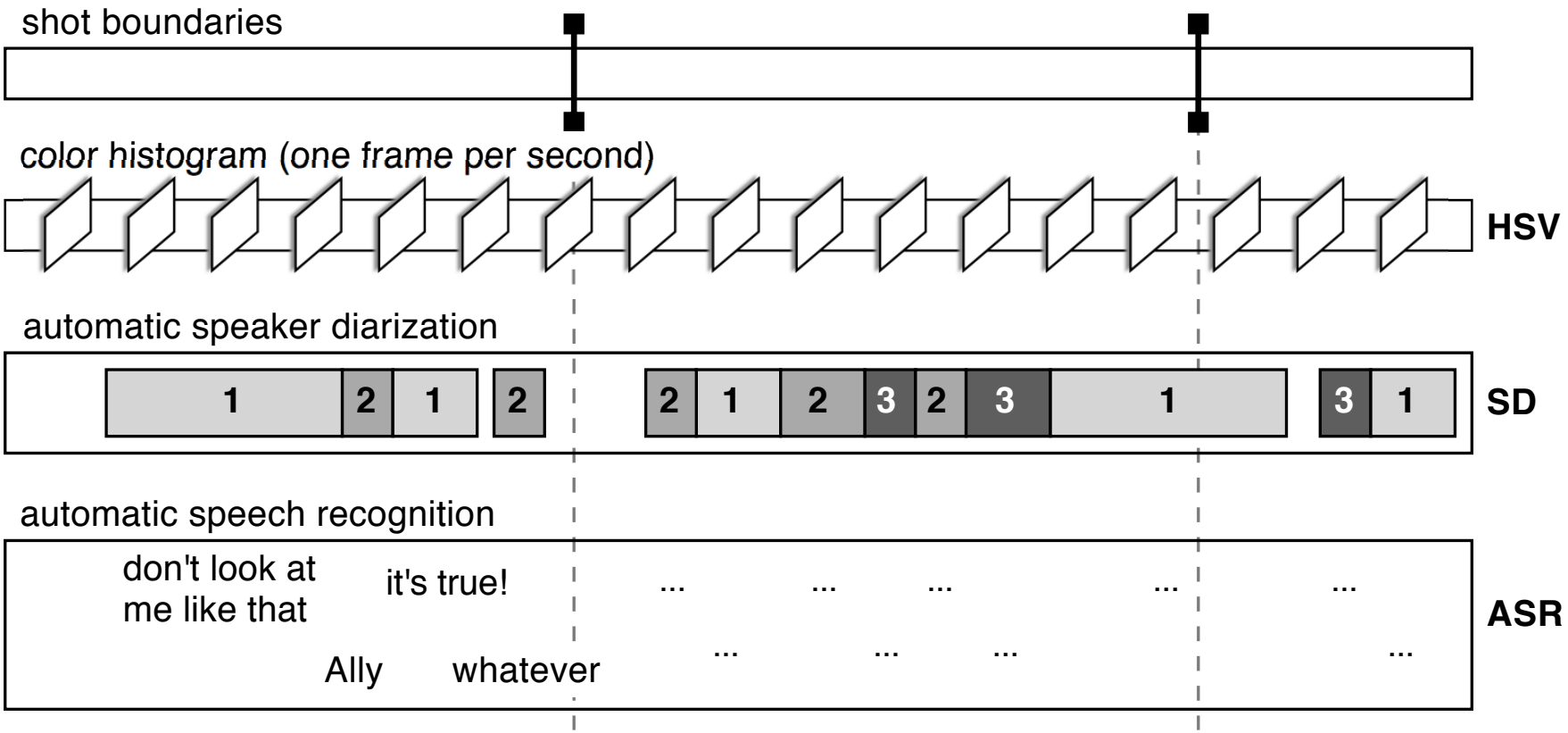


HSV/GSTG Results

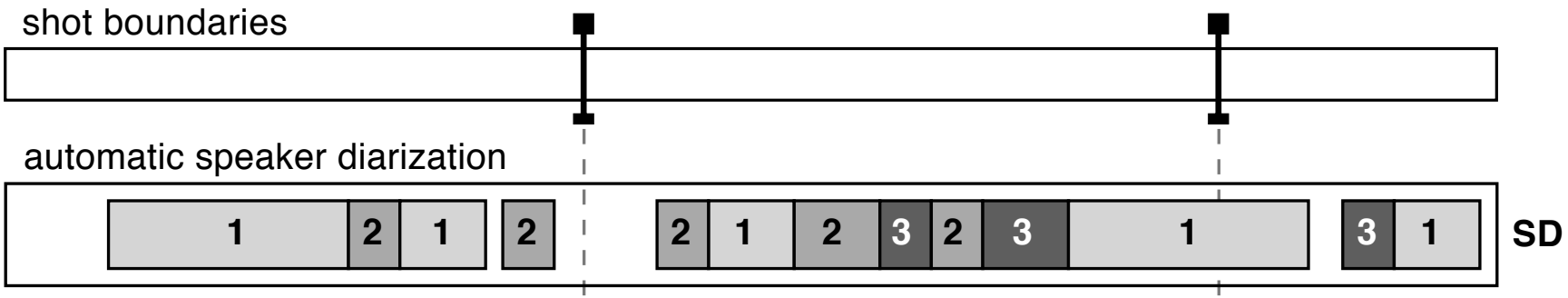
- Corpus
  - First eight episodes of *Ally McBeal* TV shows
  - 5 hours of videos, 5564 shots and 306 scenes
- Evaluation
  - Leave-one-episode-out cross validation

	Precision	Recall	F-Measure	# Scenes
HSV/STG	0.256	0.533	0.449	461
HSV/GSTG	**0.447**	0.566	**0.487**	403

Multiple modalities, multiple d_{ij}

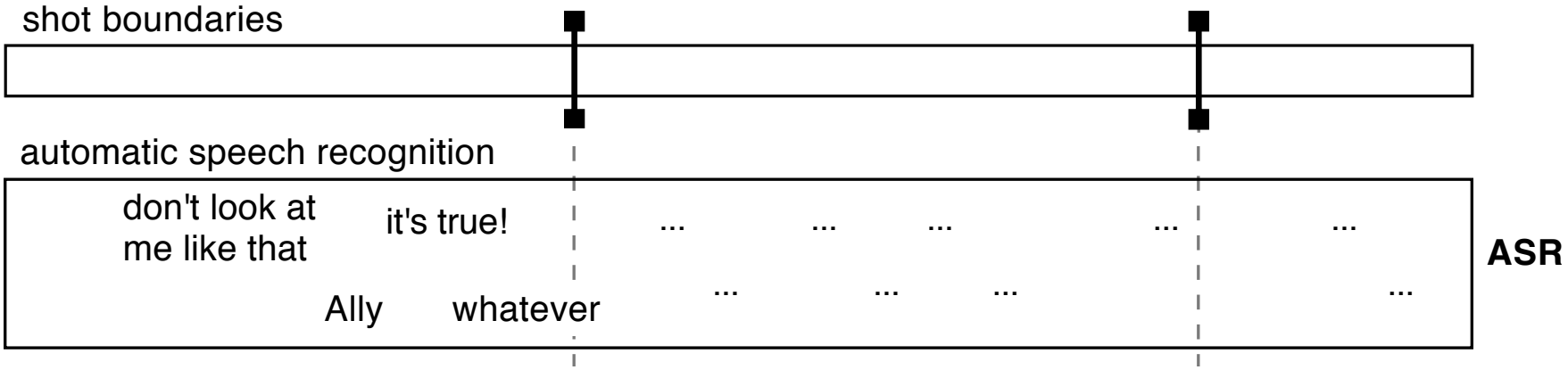


Speaker diarization



- Multiple speakers per shot, but only one descriptor
  - **Term** Frequency / Inverse **Document** Frequency
  - **Speaker** Frequency / Inverse **Shot** Frequency
- d_{ij}^{SD} : cosine distance between TF-IDF vector

Automatic Speech Recognition



- Lemmatization (tree-tagger)
- d_{ij}^{ASR}: cosine distance between TF-IDF vector

Monomodal GSTG Results

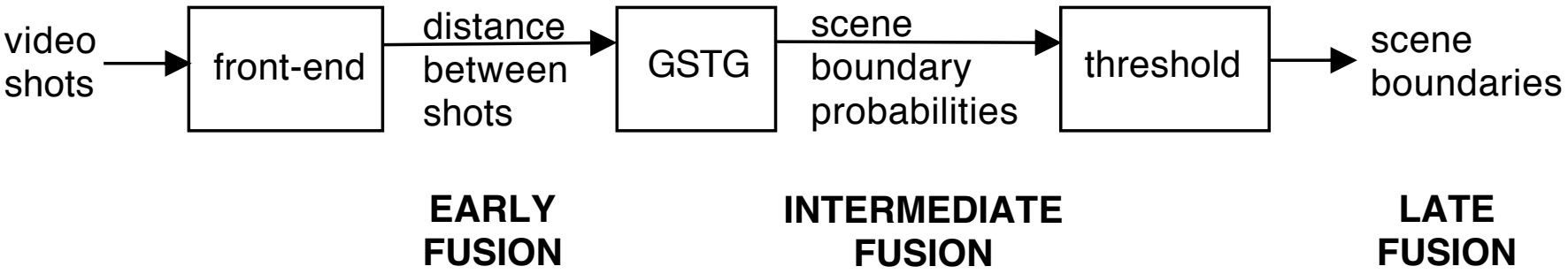
- Corpus
  - First eight episodes of *Ally McBeal* TV shows
  - 5 hours of videos, 5564 shots and 306 scenes
- Evaluation
  - Leave-one-episode-out cross validation

	Precision	Recall	F-Measure	# Scenes
HSV (STG)	0.256	0.533	0.449	461
HSV	**0.447**	0.566	**0.487**	403
SD	0.157	0.562	0.240	1136
ASR	0.105	0.572	0.175	1751

Monomodal GSTG approaches

- Limitation:
  - One modality cannot solve the problem on its own
- Proposition:
  - Multimodal fusion

Multimodal Fusion



- Late fusion
  - intersection $\cap$ or union $\cup$
- Early fusion
  - $d_{ij} = w_{\mathrm{HSV}} \cdot d_{ij}^{\mathrm{HSV}} + w_{\mathrm{SD}} \cdot d_{ij}^{\mathrm{SD}} + w_{\mathrm{ASR}} \cdot d_{ij}^{\mathrm{ASR}}$
- Intermediate fusion
  - $p = w_{\mathrm{HSV}} \cdot p_{\mathrm{HSV}} + w_{\mathrm{SD}} \cdot p_{\mathrm{SD}} + w_{\mathrm{ASR}} \cdot p_{\mathrm{ASR}}$

Multimodal Results

Fusion	Precision	Recall	F-Measure
HSV (baseline)	0.447	0.566	**0.487**
HSV ∩ SD	**0.598**	0.357	0.438
HSV ∩ SD ∩ ASR	**0.606**	0.242	0.341
HSV ∪ SD	0.180	**0.770**	0.288
HSV ∪ SD ∪ ASR	0.121	**0.851**	0.210
d(HSV) + d(SD)	0.445	0.599	0.499
d(HSV) + d(SD) + d(ASR)	0.445	0.599	0.499
p(HSV) + p(SD)	0.484	0.555	**0.510**
p(HSV) + p(SD) + p(ASR)	0.488	0.622	**0.539**

Conclusion & future work

- Graph-based approach to segmentation
- Using more modalities is better
  - Better use of ASR output
  - Add visual semantic concept detection

What's next?

- At the episode level
  - **Shot segmentation**
  - **Scene segmentation**
  - **Plot (or story) deinterlacing**
  - **Episode summarization**
    - « Previously, on Lost… »
    - Browse by plot
- At the collection level
  - Cross-episode plot
  - Episode summarization wrt. whole collection