Séminaire «Méthodes et outils pour l'Open Data»

A method of multi-video summarization

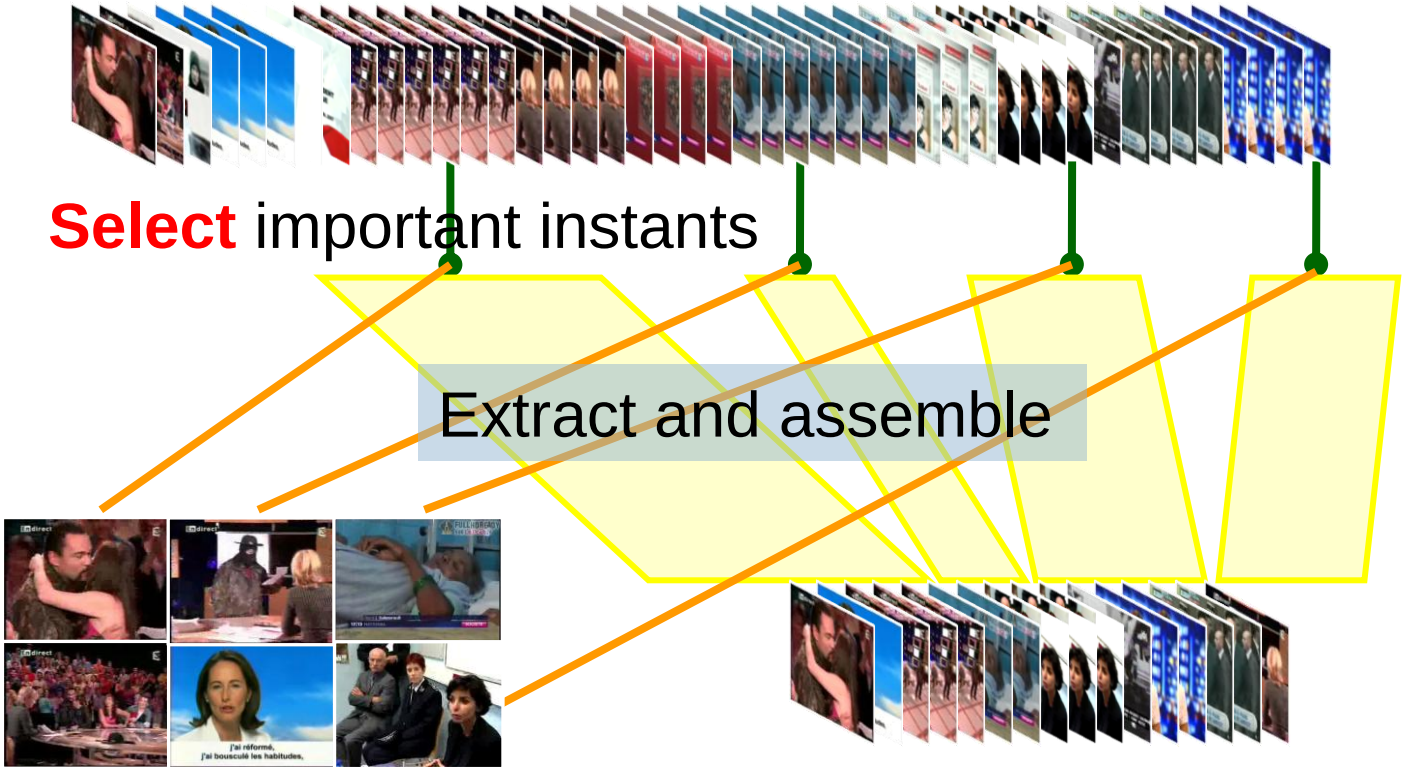
-*Multimedia Maximal Marginal Relevance*

Yingbo Li
Bioinformatics and Computational Biology, École normale supérieure, Paris, France
18 December 2014

Most research in this presentation was conducted at EURECOM, France

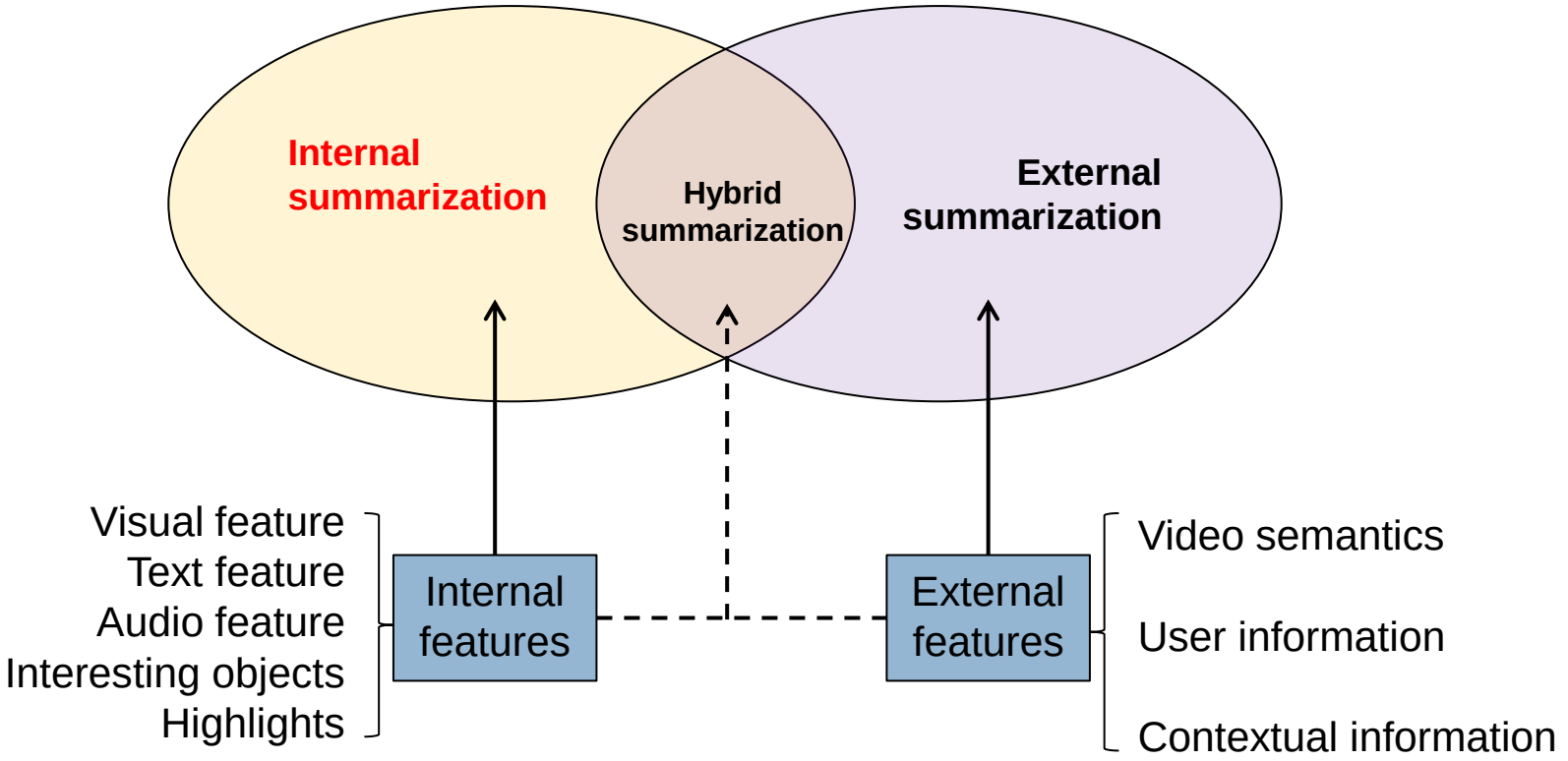
Video summarization

- Video summarization produces the condensed version of a full length video stream but keeps the important content



Video summarization

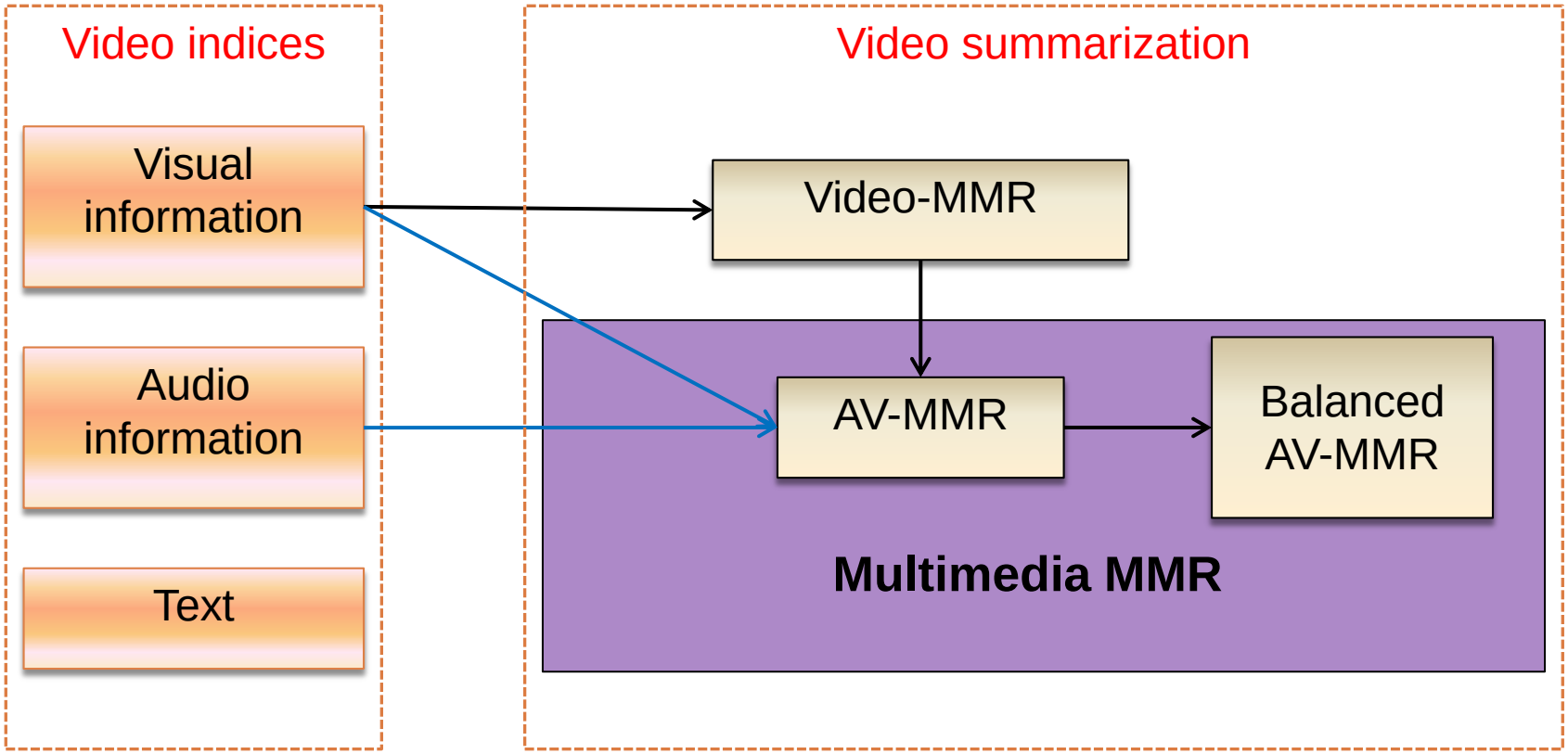
- Current techniques of video summarization[1]



[1] Money A G, Agius H. Video summarisation: A conceptual framework and survey of the state of the art[J]. Journal of Visual Communication and Image Representation, 2008, 19(2): 121-143.

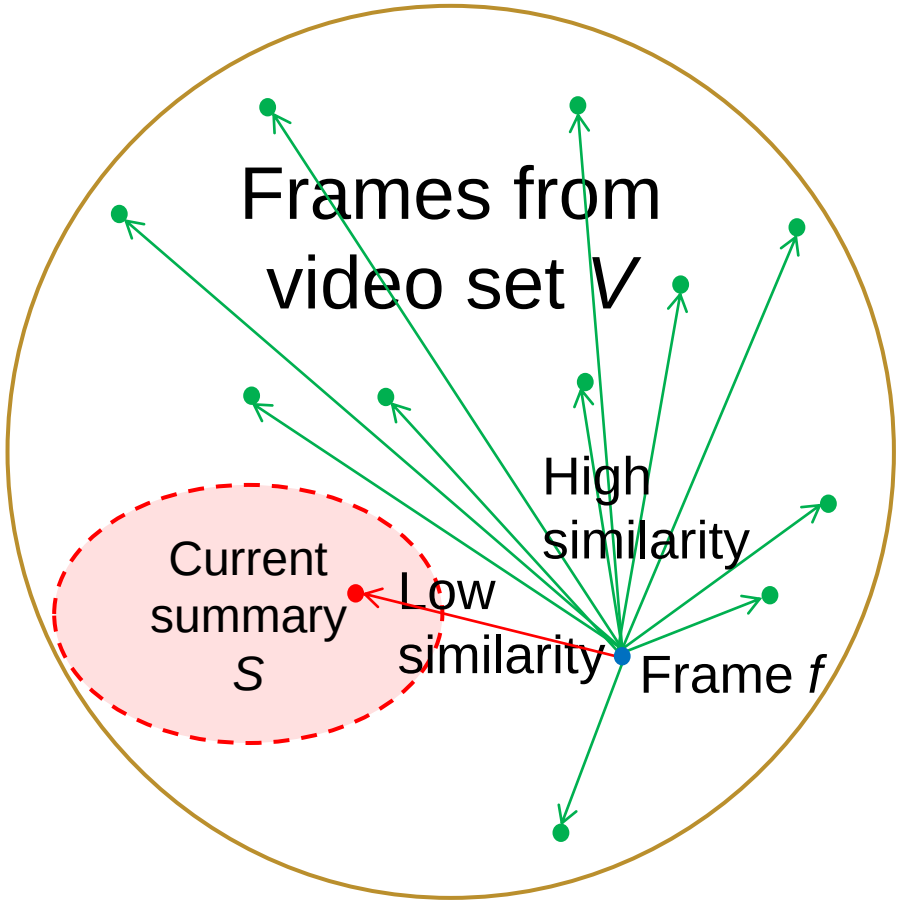
Multimedia Maximal Marginal Relevance(MMR)

- The proposed algorithms



1. Video-MMR

- Video-MMR[1] mimics MMR in text summarization



[1] Li, Y. and Merialdo, B., Multi-video summarization based on Video-MMR, WIAMIS 2010

1. Video-MMR

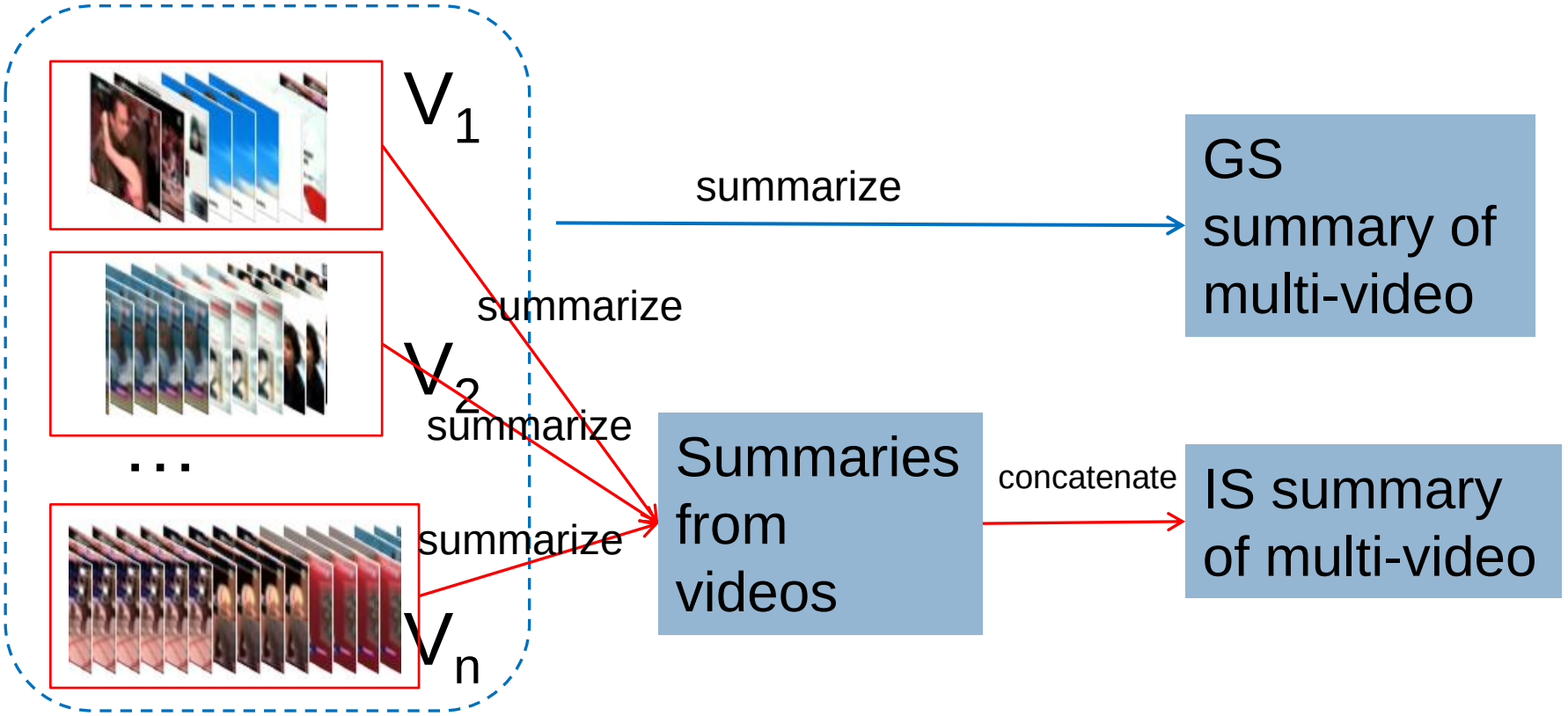
- Video-MMR selects summary keyframes based on two constraints:
  - reward relevant keyframes
  - penalize redundant keyframes

$$S_{k+1} = S_k \cup \arg\max_{f \in V \setminus S_k} [\lambda \cdot sim_1(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim_2(f, g)]$$

where $Sim_1(f_i, V \setminus S_k) = \frac{1}{|V \setminus (S_k \cup f_i)|} \sum_{f_j \in V \setminus (S_k \cup f_i)} sim(f_i, f_j)$

1. Video-MMR

- Problem: summarize multiple videos at the same time?
  - Global Summarization (GS) VS Individual Summarization (IS) for multi-video



1. Video-MMR

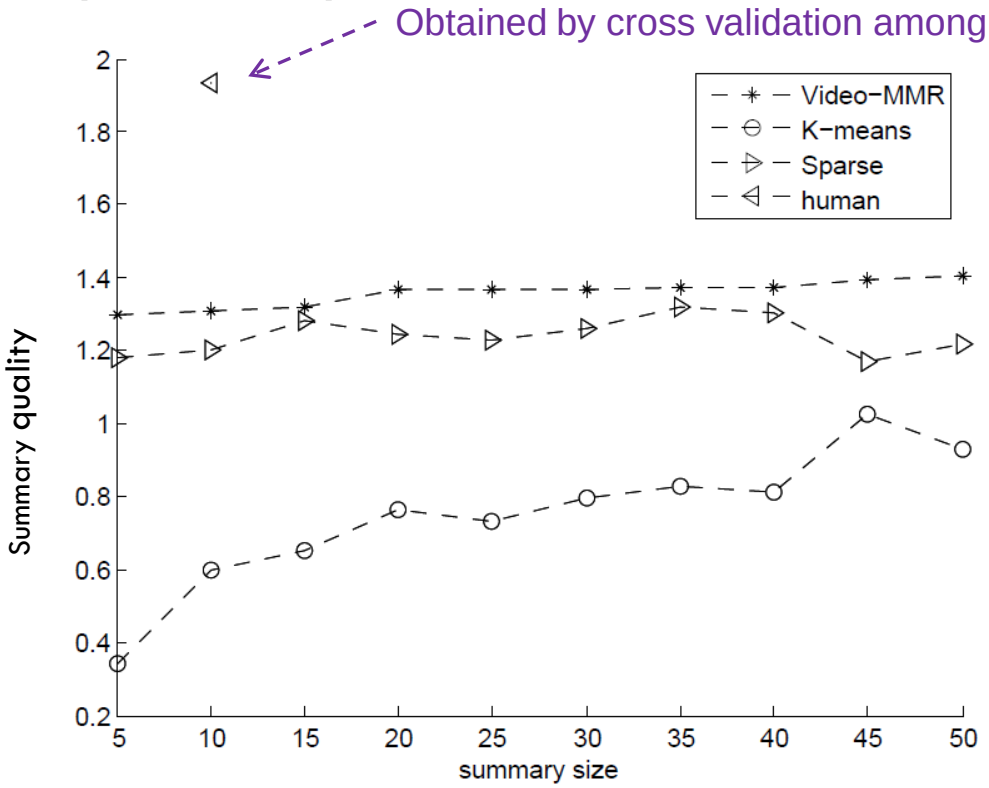
- Global Summarization vs Individual Summarization

- GS

- IS

1. Video-MMR summary VS human summary

- Evaluation compared to human ground truth:
  - Quality comparison of the summaries from Video-MMR, spares representation, K-means and human



1. Video-MMR

- Human assessment to Video-MMR

	Video-MMR	K-means	Sparse representation
Mean scores	8.06	3.96	6.24
Person 1	8.3	3.3	6.3
Person 2	8.1	5.3	5.3
Person 3	9.8	5.6	9.5
Person 4	7.1	2	5
Person 5	7	3.6	5.1

2.AV-MMR

- Video-MMR: visual information
- AV-MMR[1] (Audio Video MMR): visual and audio information

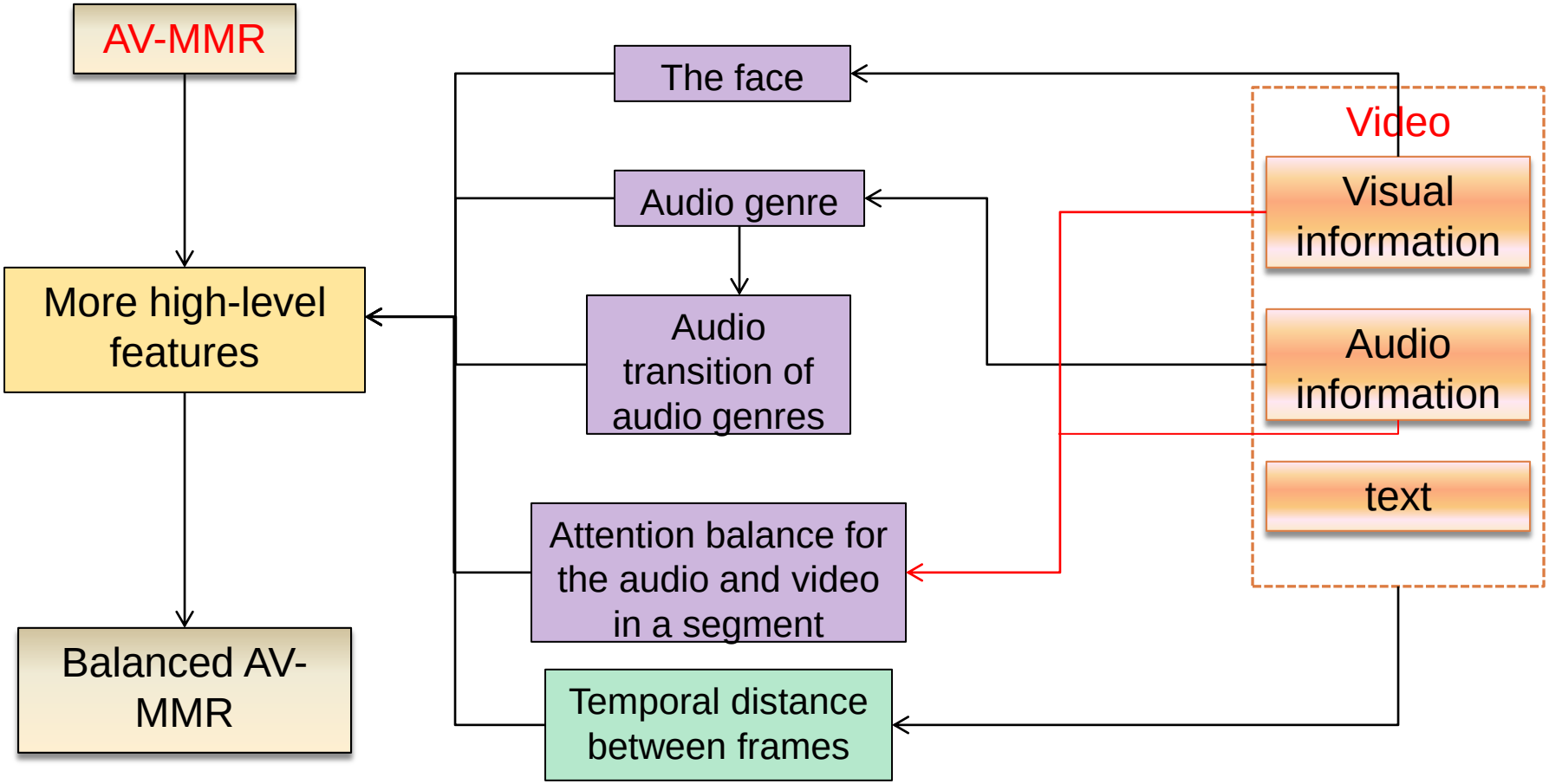
$$S_{k+1} = S_k \cup \arg\max_{f_1, f_2 \in V \setminus S_k} \{ [\lambda \cdot sim_{I1}(f_1, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim_{I2}(f_1, g)] + [\mu \cdot sim_{A1}(f_2, A \setminus S_k) - (1-\mu) \max_{g \in S_k} sim_{A2}(f_2, g)] \}$$

where $\mu = 0.5$

[1] Y. Li, and B. Merialdo, Multi-video summarization based on AV-MMR, CBMI 2010

3. Balanced AV-MMR

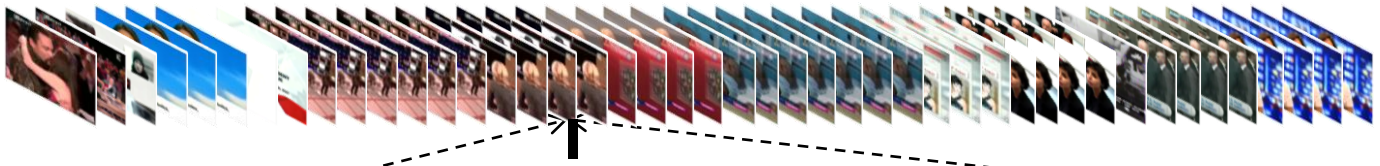
- Balanced AV-MMR[1] uses more features



[1] Li, Y. and Merialdo, B., Video Summarization Based on Balanced AV-MMR, MMM, 2012

Fundamental Balanced AV-MMR

- We propose to introduce the idea of balance between audio and visual information into AV-MMR



People naturally focus their attention but attention capacity is limited[1]. The separated attentions to the audio and the video for a video segment at the same time will vary

Visual component

$$S_{k+1} = S_k \bigcup \arg\max_{f \in V \setminus S_k} \{\rho(f)[\lambda \cdot sim_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim_{I2}(f, g)] +$$

$$(1-\rho(f))[\mu \cdot sim_{A1}(f, A \setminus S_k) - (1-\mu) \max_{g \in S_k} sim_{A2}(f, g)]\}$$

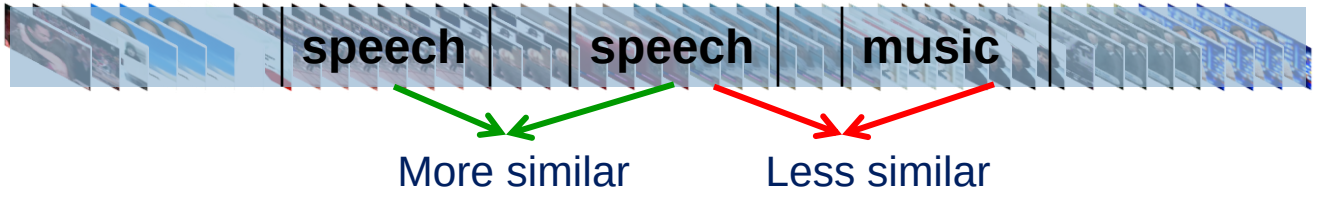
where $\rho(f) = 0.5$

Acoustic component

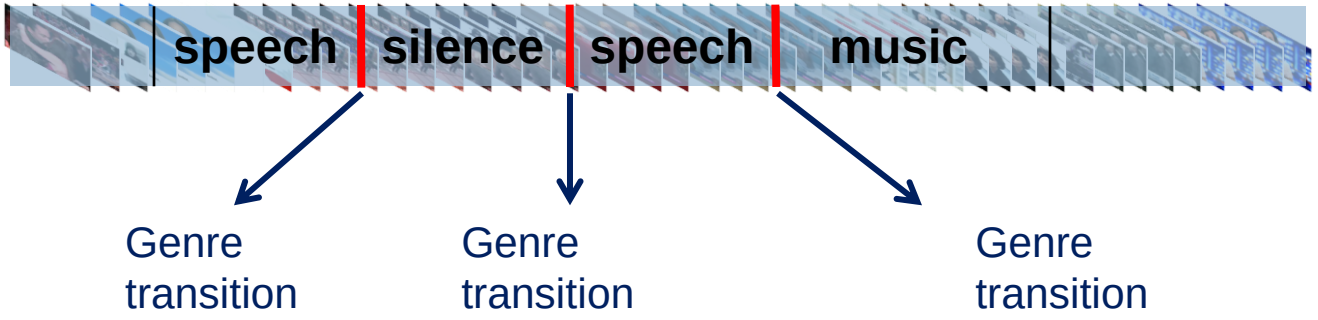
[1]: Marois, R. and Ivanoff, J., Capacity limits of information processing in the brain, Trends in Cognitive Sciences, 2005

Balanced AV-MMR V1

- audio genres (speech, music, silence…)
  1. Audio genres influence the similarity of consecutive audio segments



  2. Genre transitions incur the attention, so they increase the importance of audio



Balanced AV-MMR V1

1. Audio genre weights frame similarity: $\tau(f, f')$
2. Audio genre transitions emphasize audio information: $1-\rho(f) \rightarrow (1-\rho(f))(1+\varphi_{tr}(f))$

$$S_{k+1} = S_k \cup \arg\max_{f \in V \setminus S_k} \{\rho'(f)[\lambda \cdot sim_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim_{I2}(f, g)] + (1-\rho'(f))[\mu \cdot sim'_{A1}(f, A \setminus S_k) - (1-\mu) \max_{g \in S_k} sim'_{A2}(f, g)]\}$$

With:

$$sim'_{A1}(f_i, A \setminus S_k) = \frac{1}{|A \setminus (S_k \cup f_i)|} \sum_{f_j \in A \setminus (S_k \cup f_i)} \tau(f_i, f_j) sim(f_i, f_j)$$

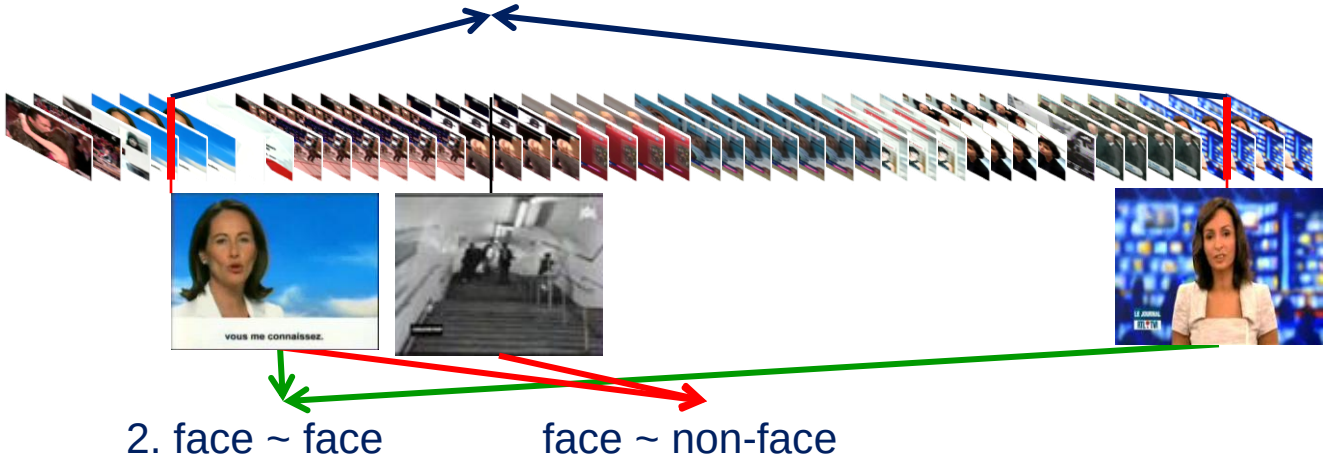
$$sim'_{A2}(f, g) = \tau(f, g) sim(f, g)$$

$$\rho'(f) = \frac{\rho(f)}{\rho(f) + (1-\rho(f)) \cdot (1+\varphi_{tr}(f))} = \frac{\rho(f)}{1+\varphi_{tr}(f) - \rho(f) \cdot \varphi_{tr}(f)}$$

Balanced AV-MMR V2

- Face occurrence is used to weight the balance factor between audio and video, and frame similarity

1. The face increases the importance of visual information in the balance between audio and visual information



Balanced AV-MMR V2

- Use face information for the balance and $\beta_{face}(f)$ similarity: $\beta'_{face}(f,f')$

$$S_{k+1} = S_k \cup \arg\max_{f \in V \setminus S_k} \{\rho''(f)[\lambda \cdot sim'_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim'_{I2}(f,g)] + (1-\rho''(f))[\mu \cdot sim'_{A1}(f, A \setminus S_k) - (1-\mu) \max_{g \in S_k} sim'_{A2}(f,g)]\}$$

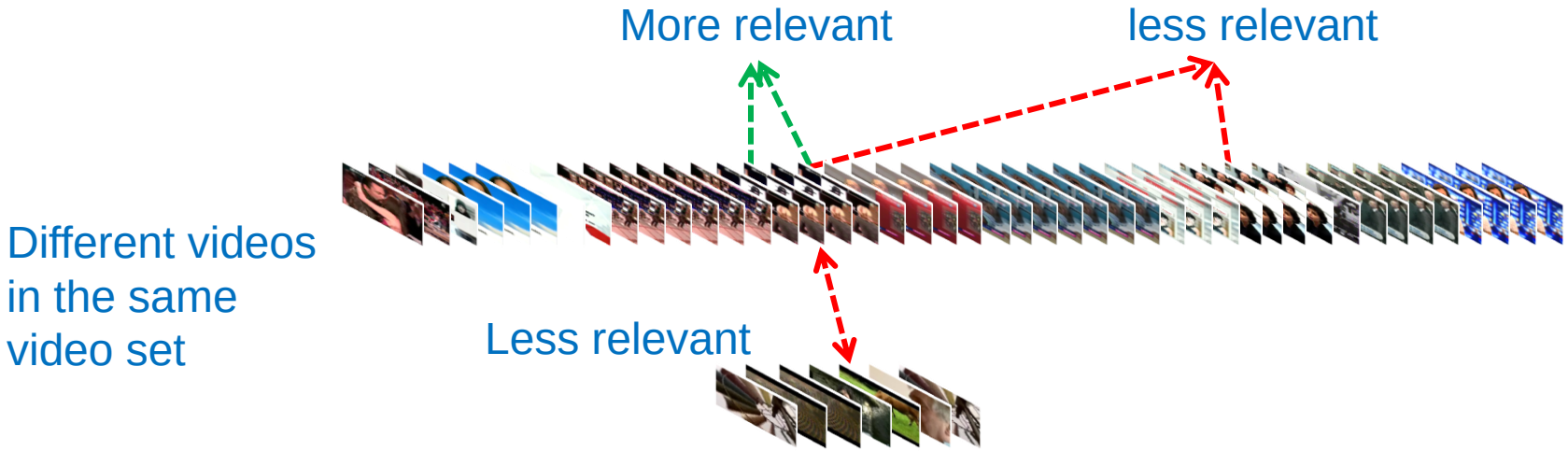
With: $$sim'_{I1}(f_i, V \setminus S_k) = \frac{1}{|V \setminus (S_k \cup f_i)|} \sum_{f_j \in V \setminus (S_k \cup f_i)} \beta'_{face}(f_i, f_j) sim(f_i, f_j)$$

$$sim'_{I2}(f,g) = \beta'_{face}(f,g) sim(f,g)$$

$$\rho''(f) = \frac{\rho(f)(1+\beta_{face}(f))}{1+\varphi_{tr}(f) - \rho(f) \cdot (\beta_{face}(f) - \varphi_{tr}(f))}$$

Balanced AV-MMR V3

- Temporal distance between two frames in one video or two different videos:
  - Closer frames in timeline are more likely to represent similar content
  - Two frames from two individual videos are less similar



Balanced AV-MMR V3

- Include temporal distance weight $\alpha_{time}(f, f')$ in audio and video similarity:

$$S_{k+1} = S_k \cup \arg\max_{f \in V \setminus S_k} \{\rho''(f)[\lambda \cdot sim''_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} sim''_{I2}(f, g)] + (1-\rho''(f))[\mu \cdot sim''_{A1}(f, A \setminus S_k) - (1-\mu) \max_{g \in S_k} sim''_{A2}(f, g)]\}$$

With:

$$sim''_{I1}(f_i, V \setminus S_k) = \frac{1}{|V \setminus (S_k \cup f_i)|} \sum_{f_j \in V \setminus (S_k \cup f_i)} \beta'_{face}(f_i, f_j) \alpha_{time}(f_i, f_j) sim_{I1}(f_i, f_j)$$

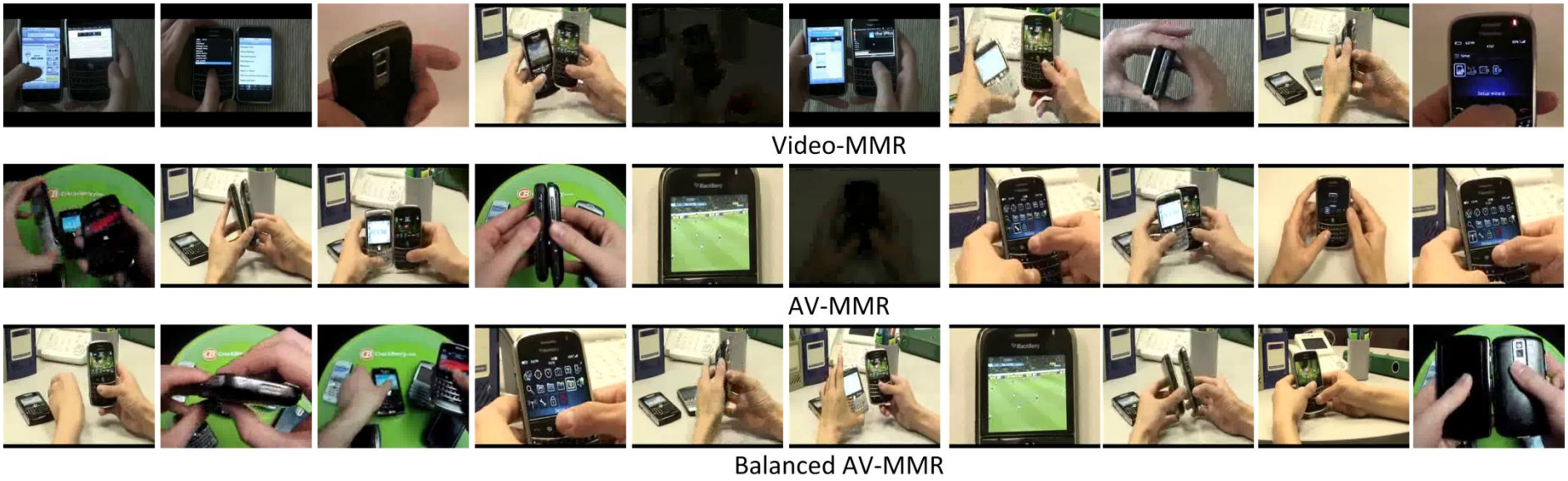
$$sim''_{A1}(f_i, V \setminus S_k) = \frac{1}{|A \setminus (S_k \cup f_i)|} \sum_{f_j \in A \setminus (S_k \cup f_i)} \tau(f_i, f_j) \alpha_{time}(f_i, f_j) sim_{A1}(f_i, f_j)$$

Multimedia MMR for multi-video

- A multi-video

Multimedia MMR for multi-video

- The static summary of 10 frames in multi-video



Multimedia MMR for multi-video

- 10 seconds' dynamic summary (video skims)
  1. 5 video segments with each 2 seconds: Video-MMR,AV-MMR, BAV-MMR

  2. 10 video segments with each 1 second: Video-MMR,AV-MMR, BAV-MMR

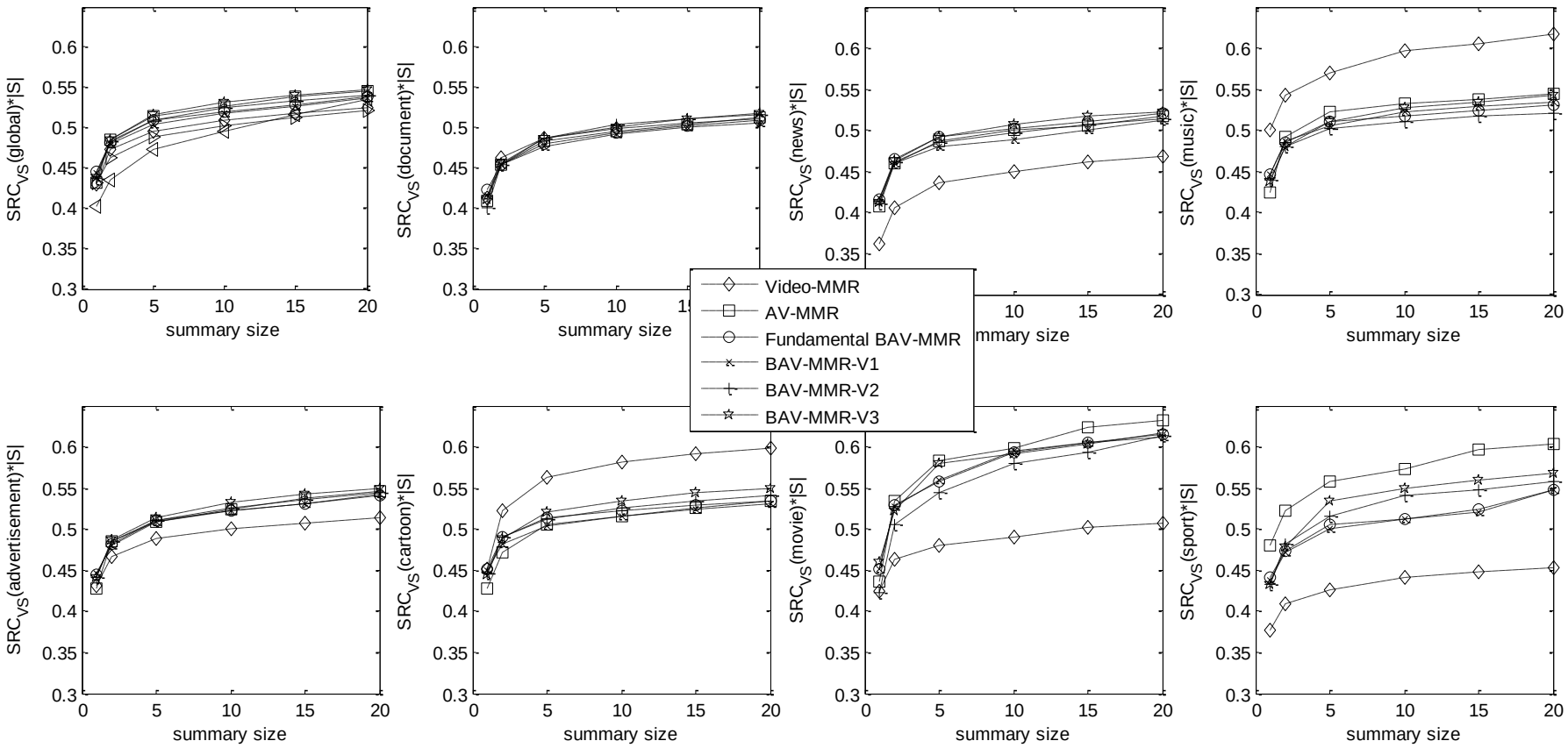
Multimedia MMR for multi-video

- The human evaluation to video skims composed of 2 seconds segment and 1 second segment

	Comfort			Information coverage		
	Video-MMR	AV-MMR	Balanced AV-MMR	Video-MMR	AV-MMR	Balanced AV-MMR
Mean scores of 2 seconds segment	7	8	8.25	6	7	7.25
Mean scores of 1 second segment	4.5	7	7	5.5	7	8

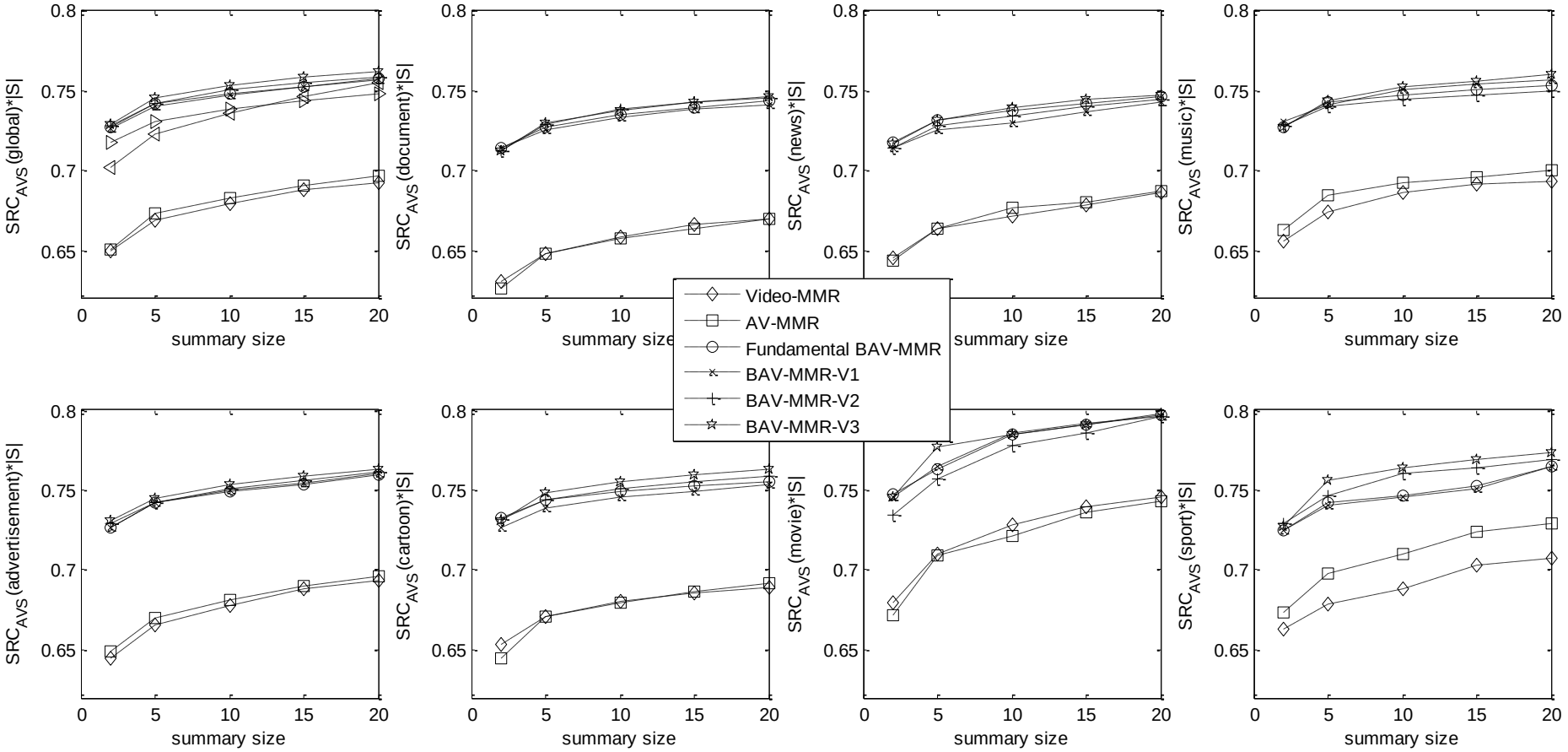
Large scale objective evaluation

1. Objective evaluation: MMR summaries compared to original videos by visual information



Large scale objective evaluation

2. Objective evaluation: MMR summaries compared to original videos by visual and audio information



Q&A

You have Questions

We have Answers