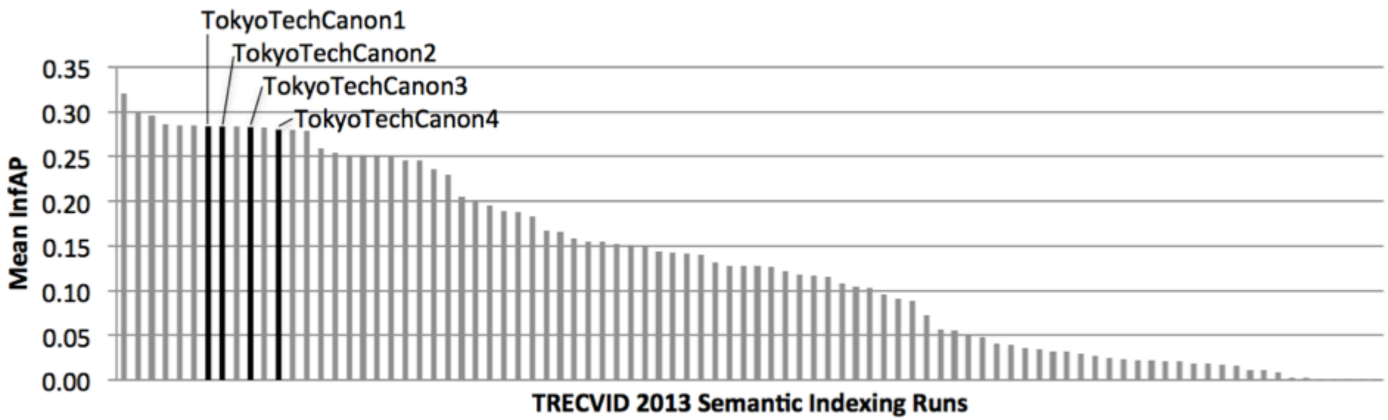
Semantic Indexing Using GMM Supervectors and Video-Clip Scores

Nakamasa Inoue, Kotaro Mori, and Koichi Shinoda,
Department of Computer Science,
Tokyo Institute of Technology

Outline

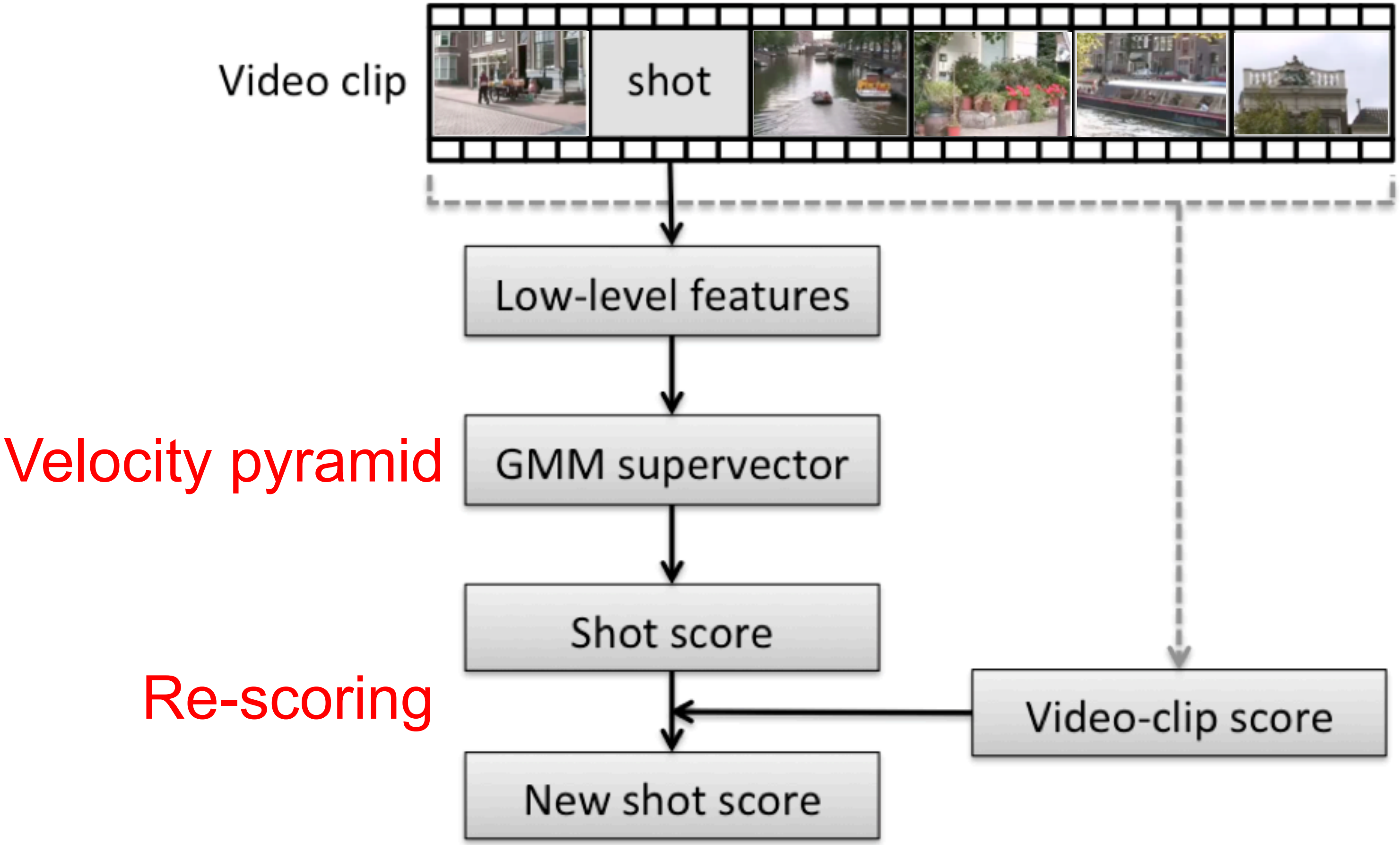
- System overview
- Baseline system
  - GMM spuervectors for 6 types of low-level features
- Spatial pyramid + Velocity pyramid*
- Re-scoring by video-clip scores
- Best result: Mean InfAP = 28.4%



* Z. Liang, N. Inoue, and K. Shinoda, "Event Detection by Velocity Pyramid," Proc. Multimedia Modeling (MMM), accepted, 2014

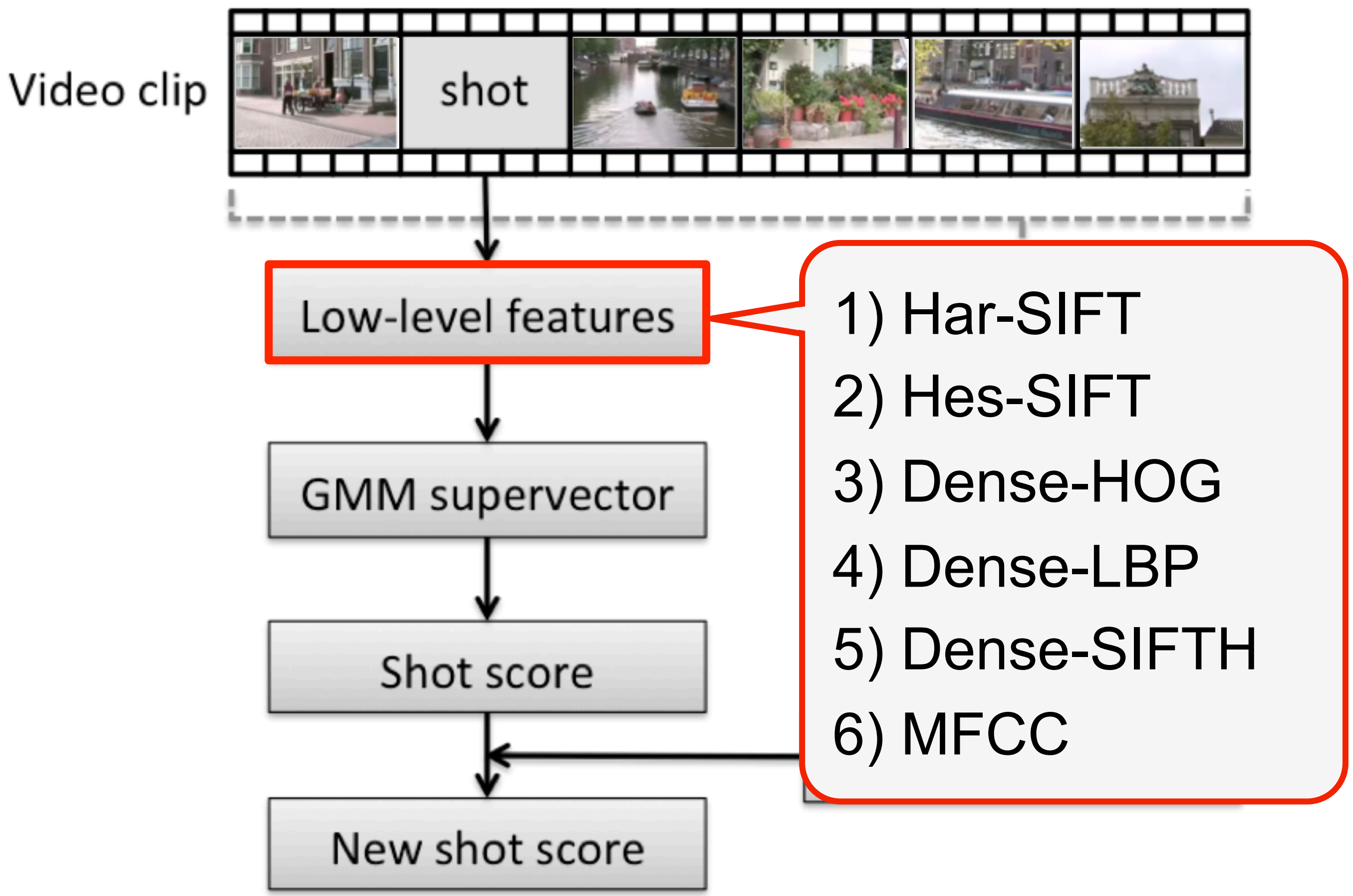
System Overview

- Extend Bag-of-Words to a probabilistic frame work



System Overview

- STEP1: low-level feature extraction



Low-Level Features (Visual)

1) Har-SIFT
- Harris-affine detector [Mikolajczyk, 2004]
- Multi-frame (every other frame)

2) Hes-SIFT
- Hessian-affine detector
- Multi-frame (every other frame)

3) Dense HOG
- 32 dimensional HOG, 10,000 samples per frame
- up to 100 frames per shot

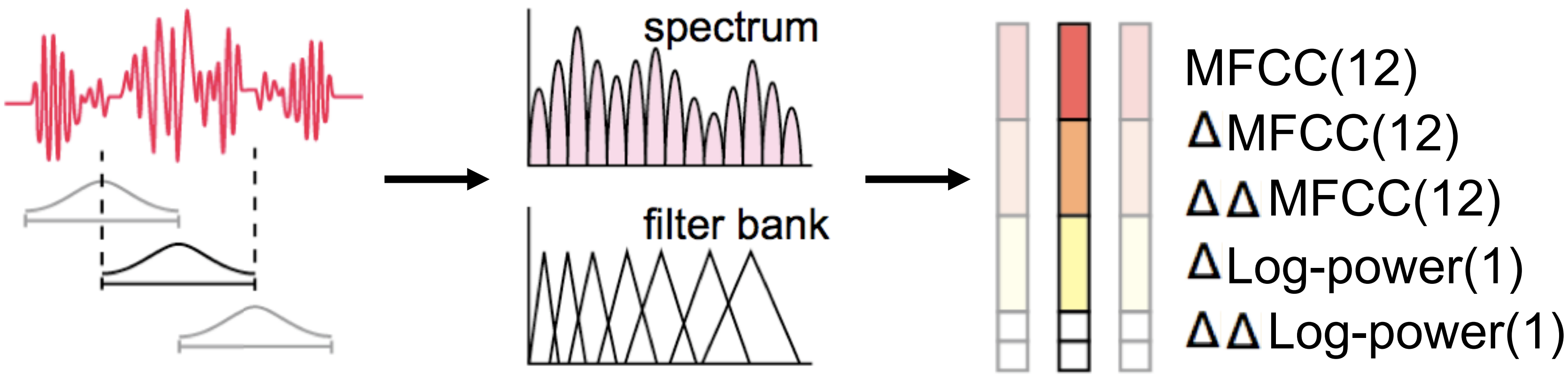
4) Dense LBP
- Local binary pattern, 10,000 samples per frame
- up to 100 frames per shot

5) Dense SIFTH
- SIFT + Hue histogram
- 30,000 samples from a key-frame

Low-Level Features (Audio)

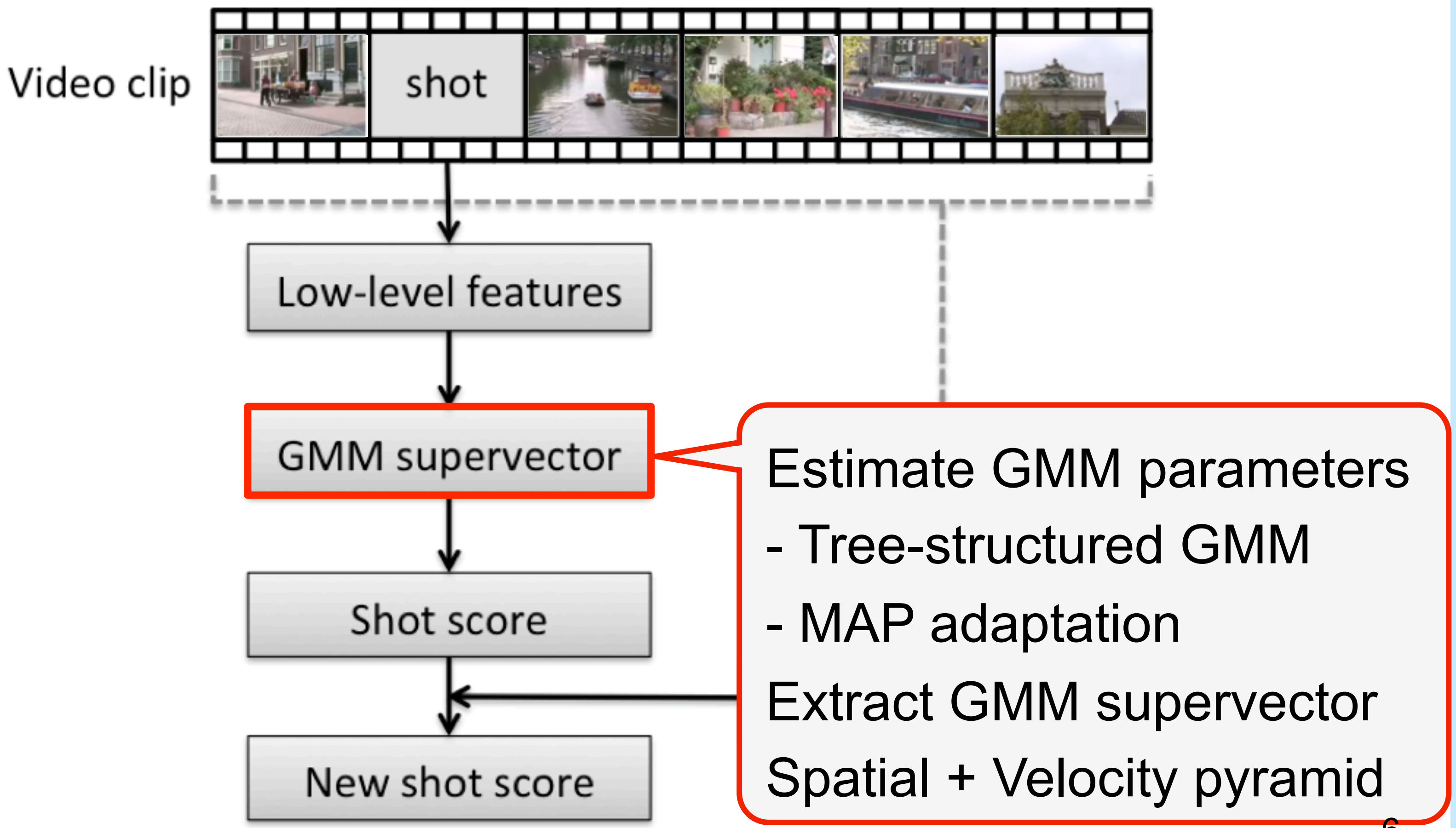
6) MFCC

- Mel-frequency cepstrum coefficients (MFCC)
- Audio features for speech recognition
- Targets: Speaking, Singing etc.



System Overview

- STEP2: GMM supervector extraction



Gaussian Mixture Models (GMMs)

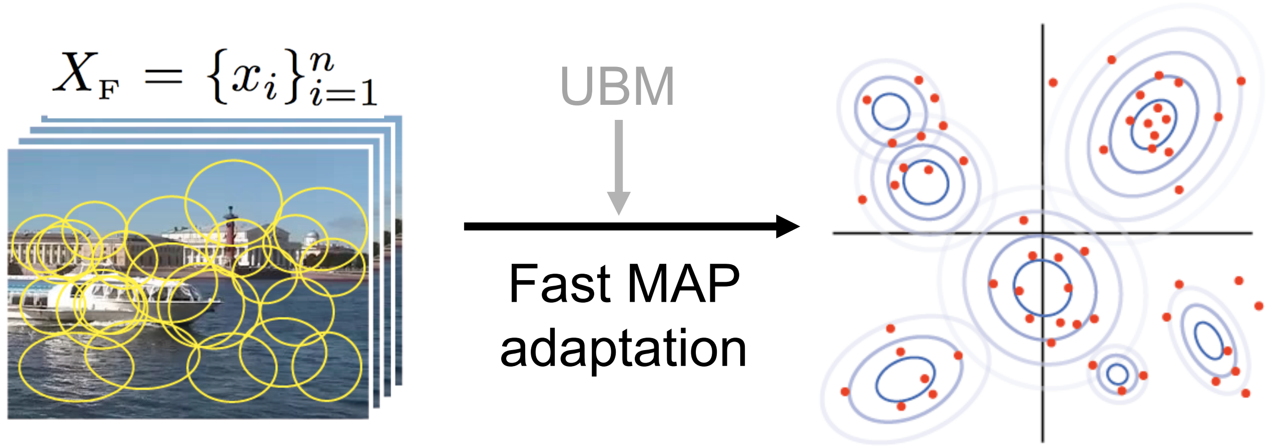
- Each shot is model by a GMM

  $X_{\mathrm{F}} = \{x_i\}_{i=1}^{n}$: local features

  $\theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$: GMM parameters

- GMM parameters are estimated by using maximum a posteriori (MAP) adaptation

$X_{\mathrm{F}} = \{x_i\}_{i=1}^{n}$



Universal background model (UBM): a prior GMM which is estimated by using all video data.

Gaussian Mixture Models (GMMs)

- MAP adaptation for mean vectors:

$$\hat{\mu}_k = \frac{\tau \hat{\mu}_k^{(\mathrm{U})} + \sum_{i=1}^{n} c_{ik} x_i}{\tau + C_k}$$

where

$$c_{ik} = \frac{w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}, \quad C_k = \sum_{i=1}^{n_s} c_{ik}$$

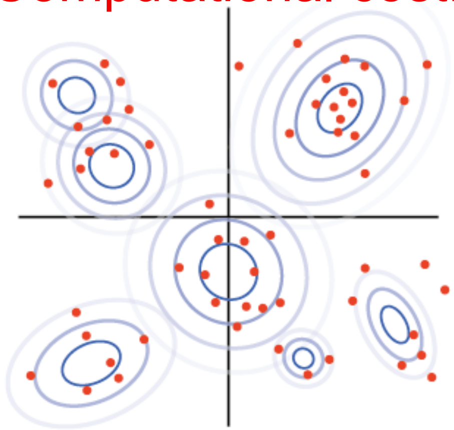
responsibility of component k for x_i

Computational cost: high

$$X_{\mathrm{F}} = \{x_i\}_{i=1}^{n}$$

UBM

Fast MAP adaptation*



* N. Inoue and K. Shinoda, "A Fast and Accurate Video Semantic-Indexing System Using Fast MAP Adaptation and GMM Supervectors," IEEE Trans. on Multimedia, vol.14, no.4, pp. 1196-1205, 2012.

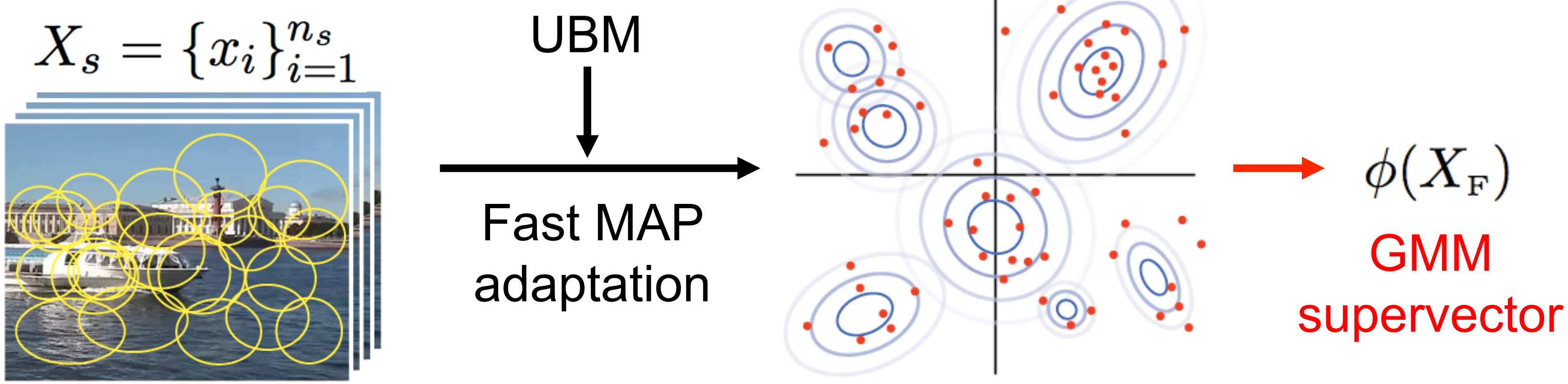
GMM Supervector

- Combine normalized mean vectors.

$$\phi(X_{\mathrm{F}}) = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_K \end{pmatrix}$$

where

$$\tilde{\mu}_k = \underbrace{\sqrt{w_k^{(U)}}(\Sigma_k^{(U)})^{-\frac{1}{2}}}_{\text{normalized}} \underbrace{\hat{\mu}_k}_{\text{mean}}$$

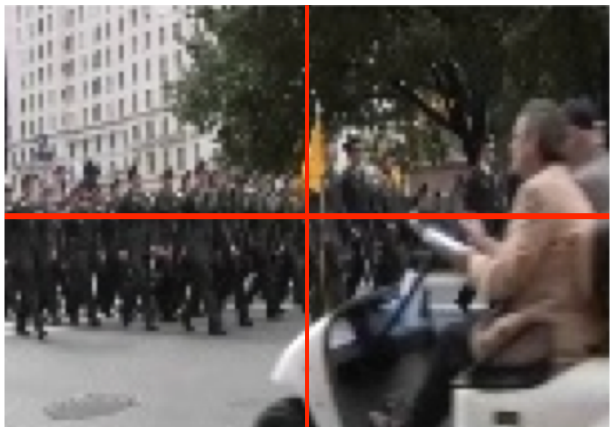


Velocity Pyramid

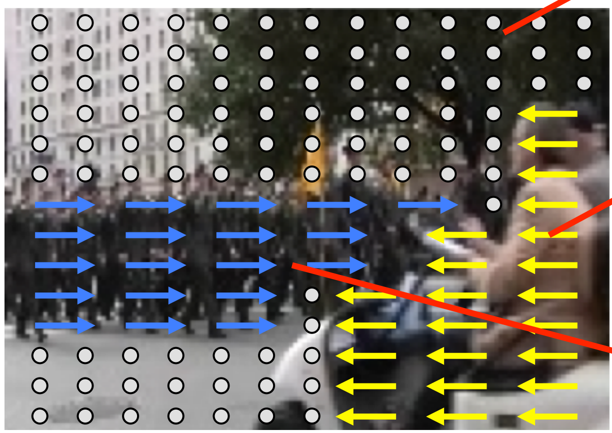
- Extend spatial pyramid to motion

- extract optical flow, quantize *velocity vectors*

- concatenate GMM supervectors



Spatial



Velocity

BoW/GMM sv

[ ] no motion

[ ] left

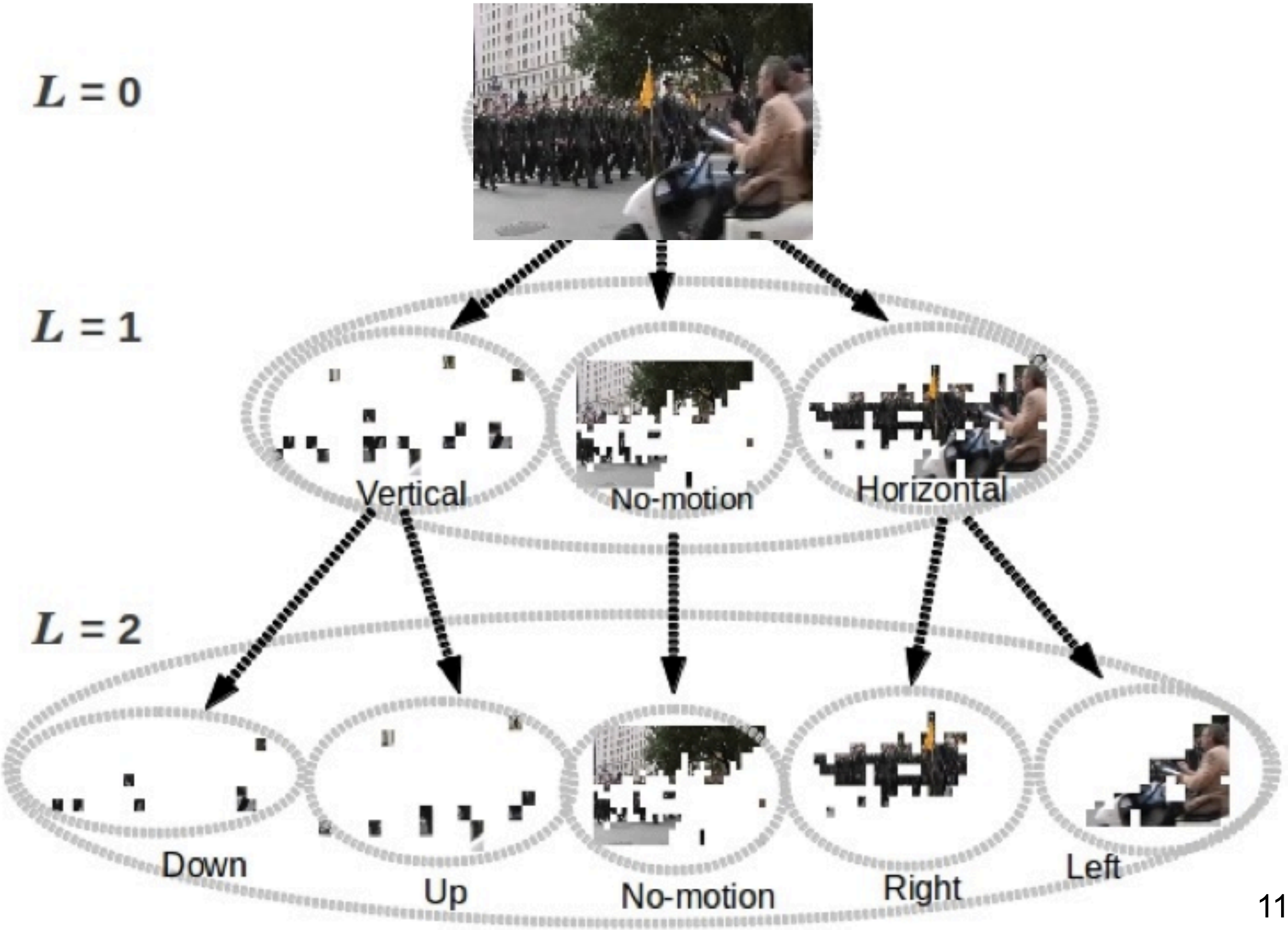
[ ] right

[ ] up

[ ] down

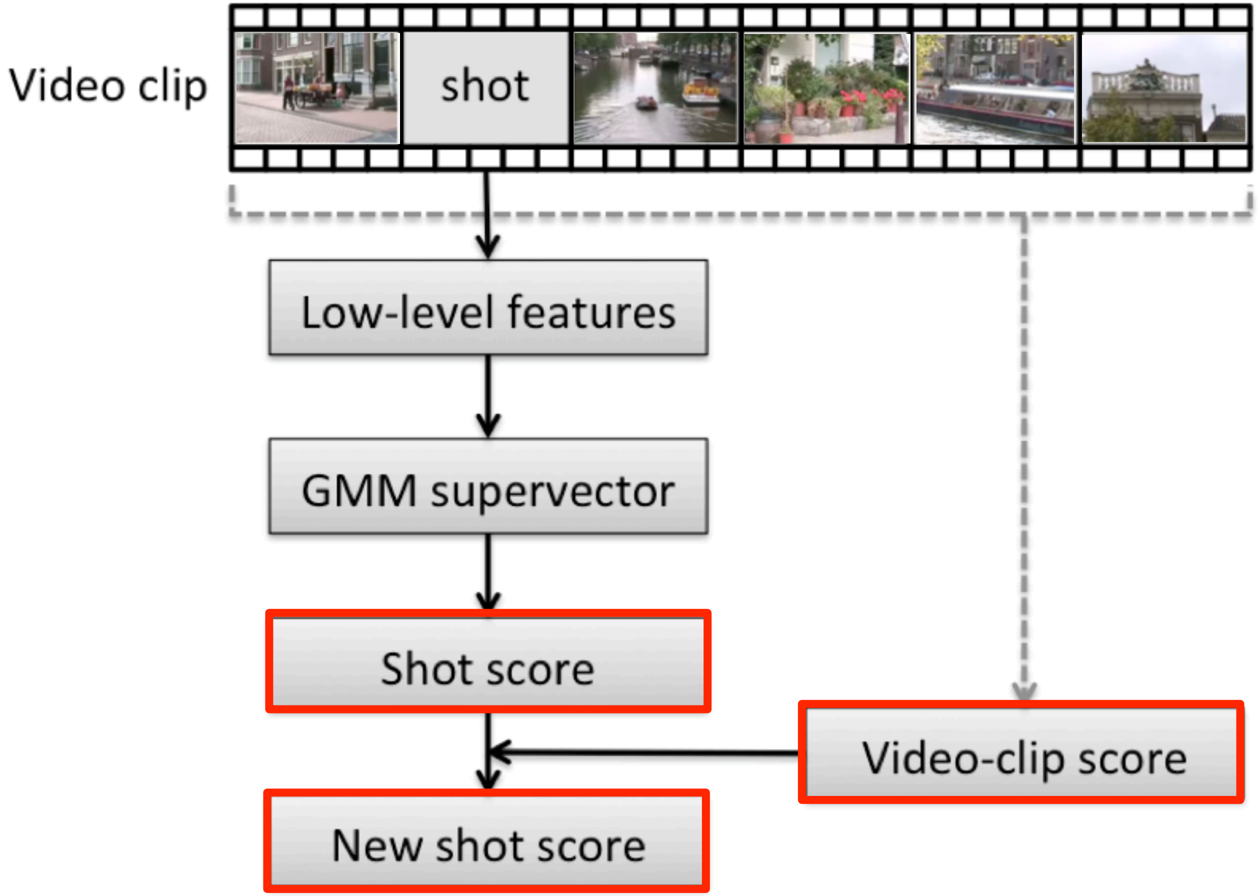
Z. Liang, N. Inoue, and K. Shinoda, "Event Detection by Velocity Pyramid," Proc. Multimedia Modeling (MMM), accepted, 2014

Velocity Pyramid



System Overview

- STEP3: compute shot scores



Shot Scores

- Linear combination of SVM scores

$$s = \sum_{\mathrm{F} \in \mathcal{F}} \alpha_{\mathrm{F}} f_{\mathrm{F}}(X_{\mathrm{F}}), \quad 0 \leqq \alpha_{\mathrm{F}} \leqq 1, \quad \sum_{\mathrm{F}} \alpha_{\mathrm{F}} = 1$$

where

$\mathcal{F} = \{\text{SIFT-Har}, \text{SIFT-Hes}, \text{SIFTH-Dense}, \text{HOG-Dense}, \text{LBP-Dense}, \text{MFCC}\}$

α_{F} : optimized for each semantic concept (on IACC_1_B)

s_1

s_2

s_3

s_4

s_5

Video-Clip Score

- A semantic concept often reappears in a video clip
- Problem: occlusion, closed-up etc.

Video-Clip Score

- Video-clip score: the maximum shot score in a clip
- Re-scoring:

$$s'_i = (1-p)s_i + p\underbrace{s_{\max}}_{\text{Video-clip score}} \qquad p = r\left\langle \frac{\#(\text{positive shots in a video clip})}{\#(\text{shots in a video clip})} \right\rangle$$

Shot score s_1 s_2 s_3 $\boxed{s_4}$ max s_5

Re-scoring

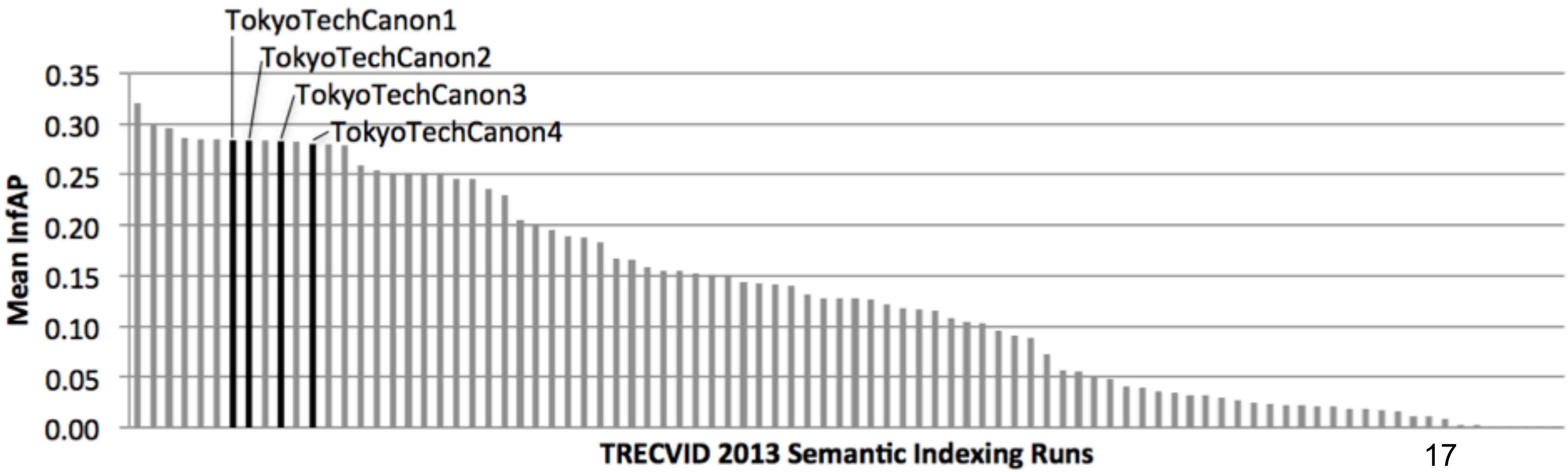
s'_1 s'_2 s'_3 s'_4 s'_5

Experimental Condition

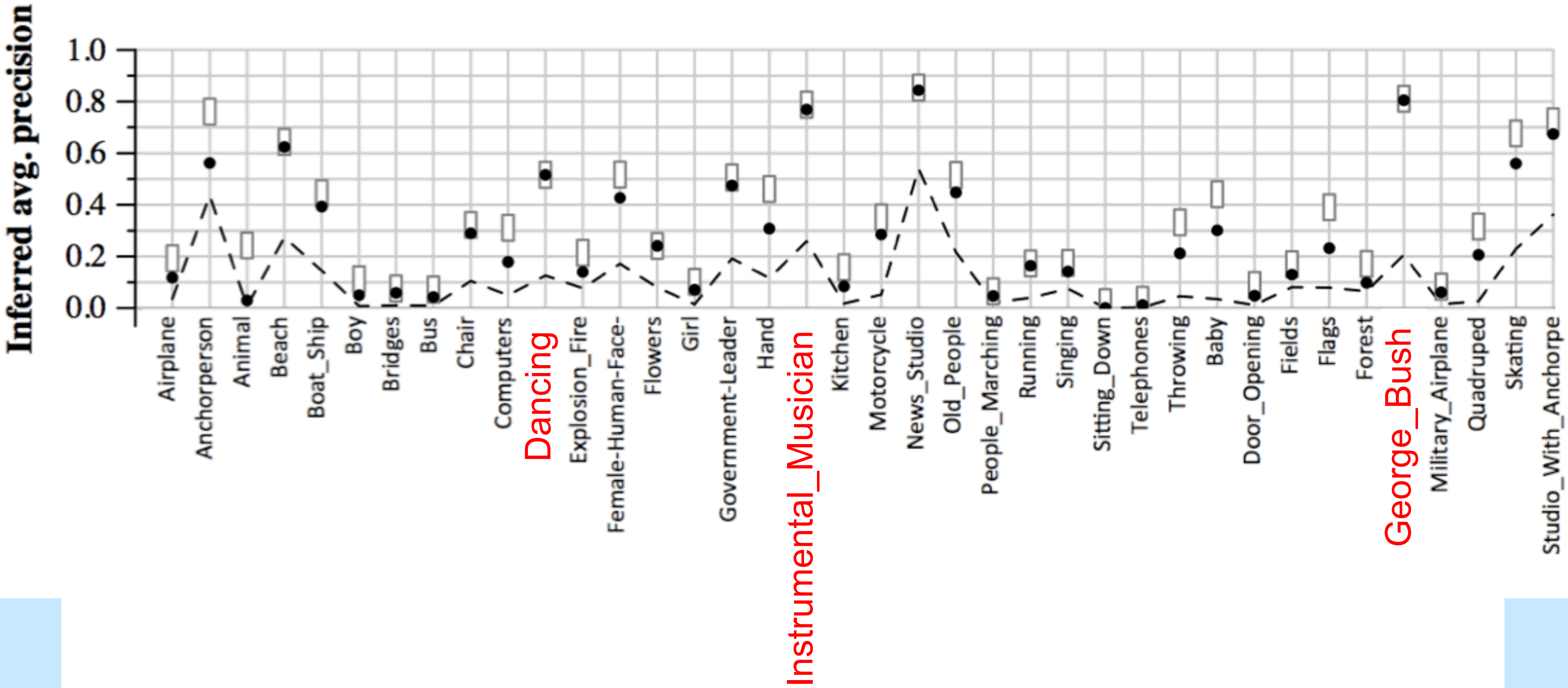
- TokyoTech_Canon_4
  - 6 types of GMM supervectors
  - Video-clip score (r=1.0)
- TokyoTech_Canon_3
  - \+ Spatial and velocity pyramid for HOG
- TokyoTech_Canon_2
  - set r=0.9 for video-clip scores
- TokyoTech_Canon_1
  - set r=0.8 for video-clip scores

Results

Run ID	Method	Mean InfAP
TokyoTech_Canon_4	6 types of GMM sv + video-clip scores	0.280
TokyoTech_Canon_3	+ Spatial and velocity pyramid	0.283
TokyoTech_Canon_2	set r = 0.9	**0.284**
TokyoTech_Canon_1	set r = 0.8	**0.284**



InfAP by Semantic Concepts



Evaluation of Velocity Pyramid

- Mean NDC on the MED task (HOG features)

	MED 10	MED 11
No pyramid	0.661	0.688
Spatial pyramid (SP)	0.635	0.617
Velocity pyramid (VP)	0.617	0.620
SP+VP	0.607	0.600

- Mean AP on the SIN task

	SIN 12 (HOG)	SIN 12 (Fusion)	SIN 13 (Fusion)
No pyramid	0.236	0.321	0.280
SV+VP	0.245	0.323	0.283

* Fusion: fusion of 6 types of visual and audio features,
but SV+VP is applied to only HOG

Evaluation of Video-clip Scores

- Mean AP on SIN 2012

Feature Type	Video-Clip Score	
	No	Yes
Har-SIFT	0.183	0.208
Hes-SIFT	0.179	0.207
Dense-SIFTH	0.202	0.224
Dense-HOG	0.236	0.259
Dense-LBP	0.235	0.260
MFCC	0.079	0.086
Fusion	0.306	0.321
Fusion (r=0.9)	0.306	0.324

Conclusion

- 6 types of audio and visual GMM supervectors
  + Velocity pyramid
  + Re-scoring by video-clip scores
- Experimental Results
  - Mean InfAP: 0.284
- Future work
  Improve audio analysis
  Audio-visual localization