Semantic Indexing Using Deep CNNs and GMM Supervectors

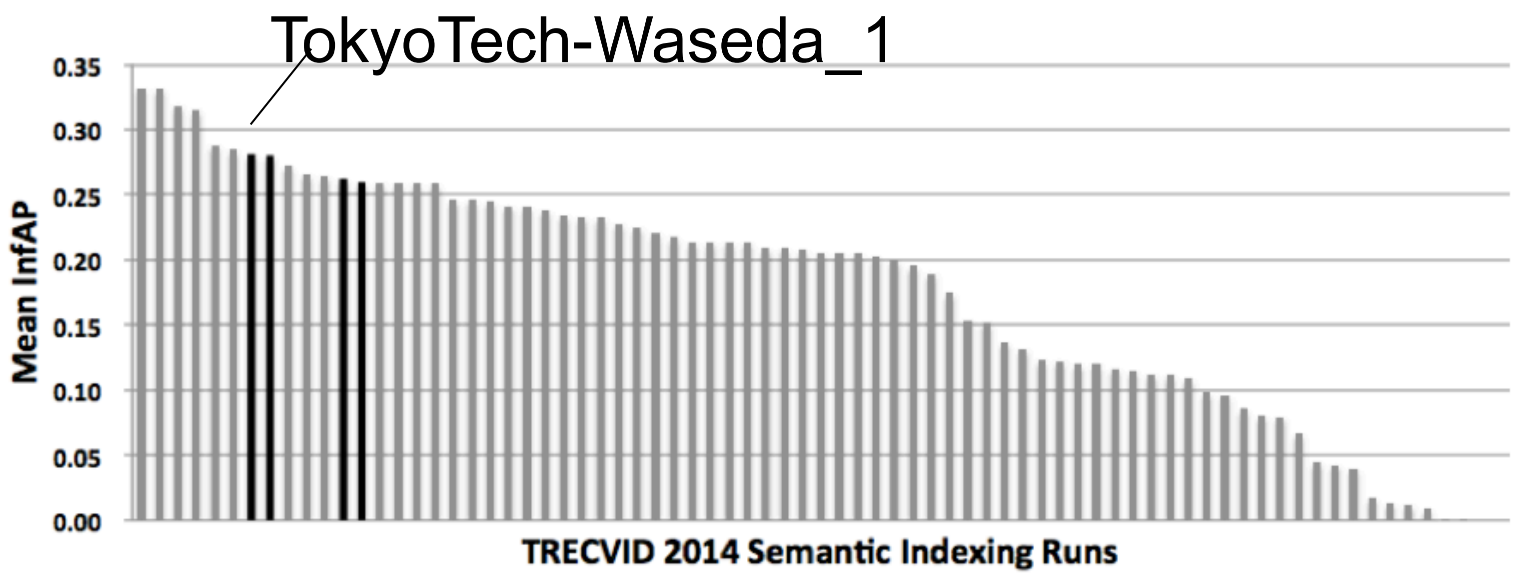
Nakamasa Inoue and Koichi Shinoda
Tokyo Institute of Technology

Zhang Xuefeng and Kazuya Ueki
Waseda University

Outline

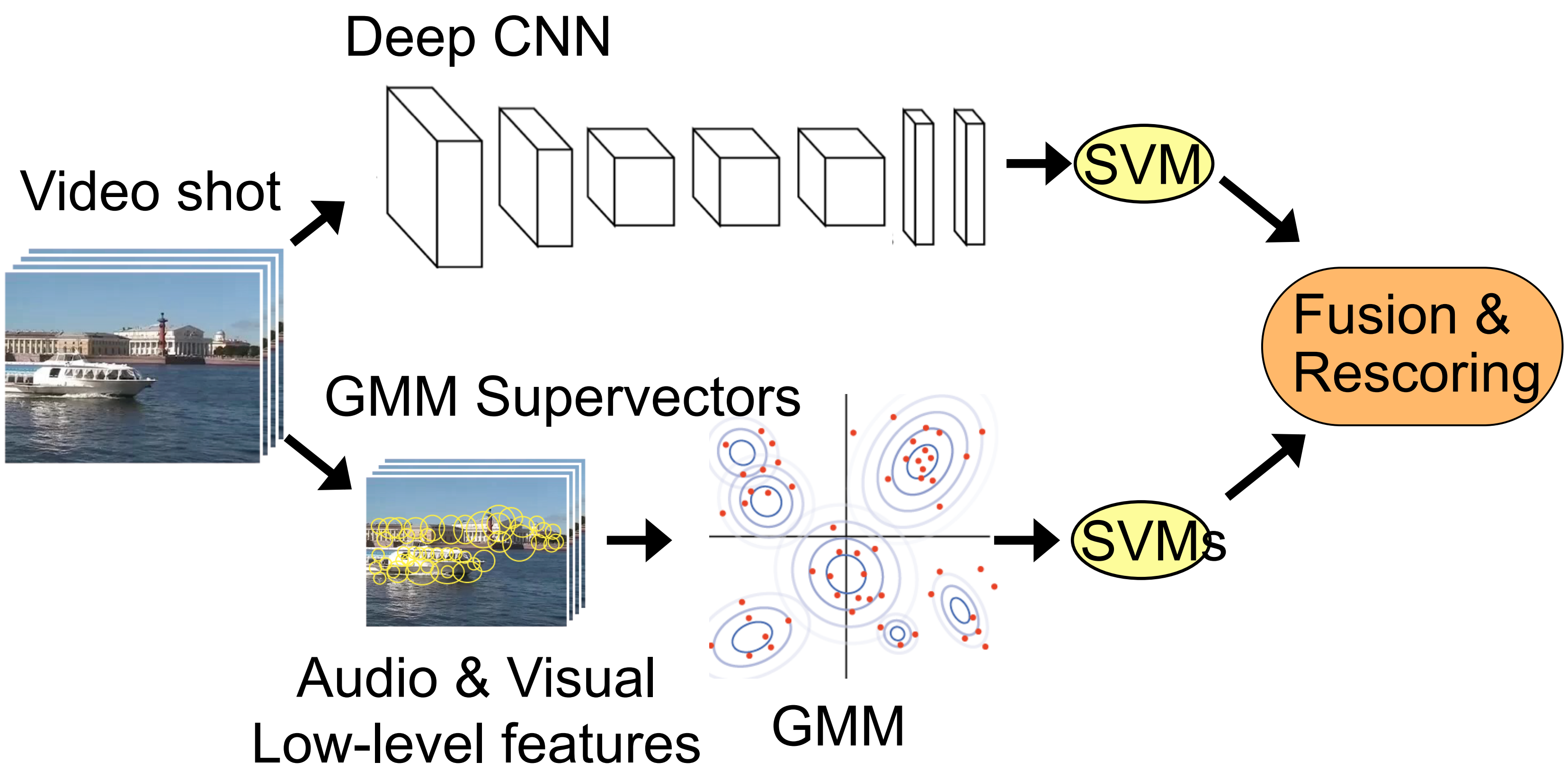
- Part 1: Our system at TRECVID 2014
  - Deep CNNs + GMM spuervectors
  - n-gram models for re-scoring

  Best result: Mean InfAP = 0.281
- Part 2: Motion features & Future work



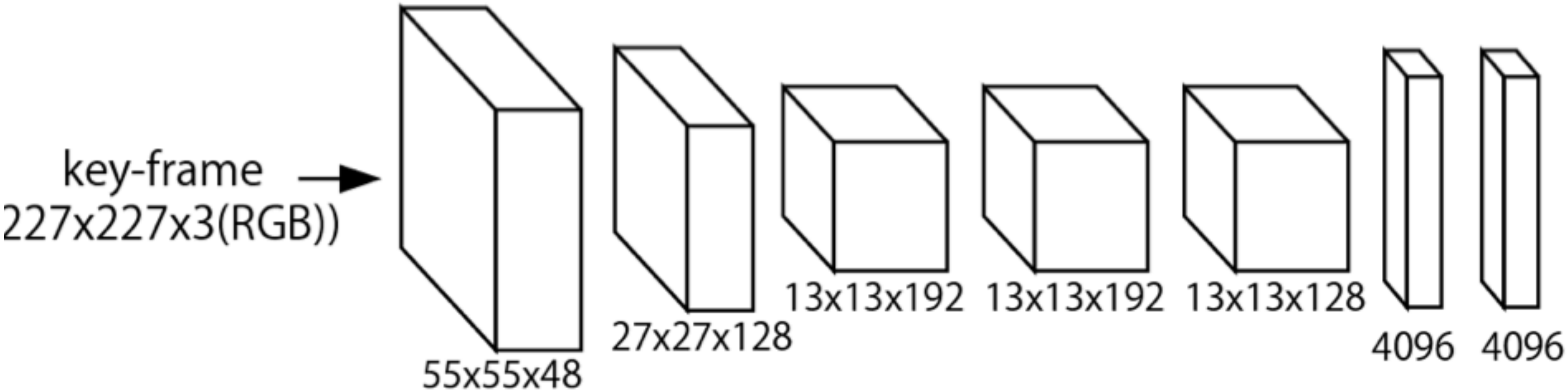
System Overview

- Deep CNN + GMM Supervectors



Deep CNN

- A 4096 dimensional feature vector at the sixth layer is extracted
- A pre-trained model on ImageNET 2012 [1]

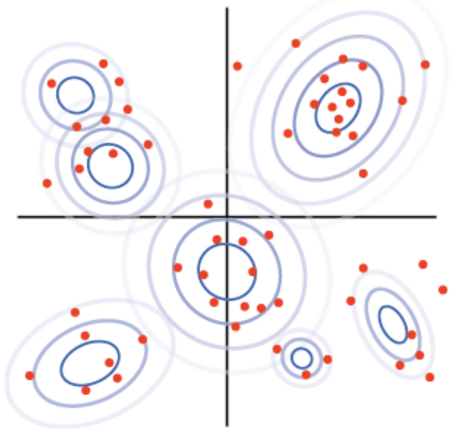


[1] Y. Jia, et al., Caffe: Convolutional Architecture for Fast Feature Embedding. Proc. ACM Multimedia Open Source Competition, 2014.

GMM Supervectors

- Extend BoW to a probabilistic framework

1) Extract 6 types of visual/audio features: Har-SIFT, Hes-SIFT, Dense HOG, Dense LBP, Dense SIFTH, and MFCC
2) Estimate GMM parameters for each shot
3) Combine normalized mean vectors



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

GMM supervector

Shot Scores

- Linear combination of SVM scores

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

where F is a feature type, α_{F} is a weight.

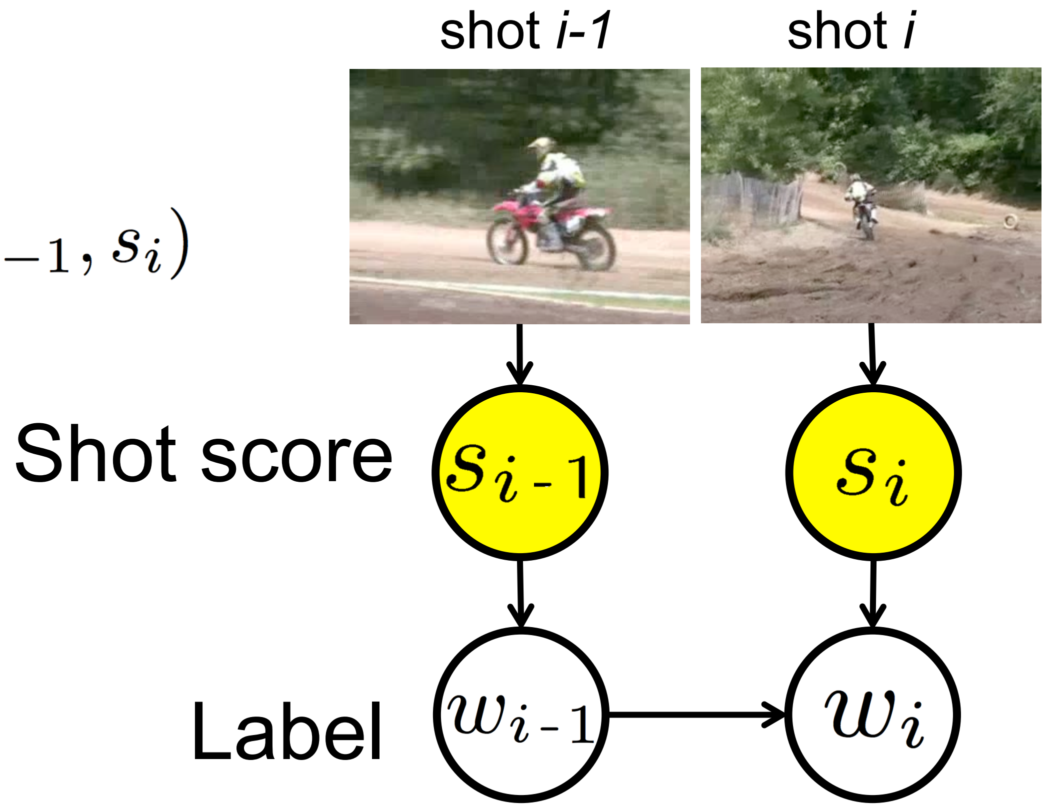
n-Gram Models

- n-consecutive video shots are dependent
- Bigram (n=2)

Re-scoring by

$$p(w_i = +1 | s_{i-1}, s_i)$$

Label (+1 or -1)



N. Inoue and K. Shinoda, “n-gram models for video semantic indexing,” ACM MM 2014.

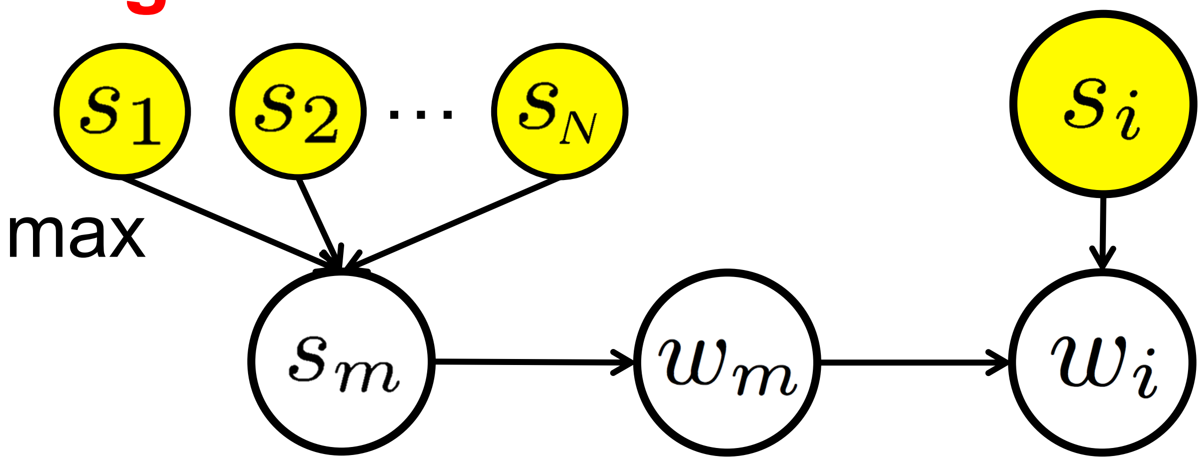
A Full-Gram Model

- n-consecutive video shots are dependent
- Full-gram

- we simply add the maximum shot score in a video clip

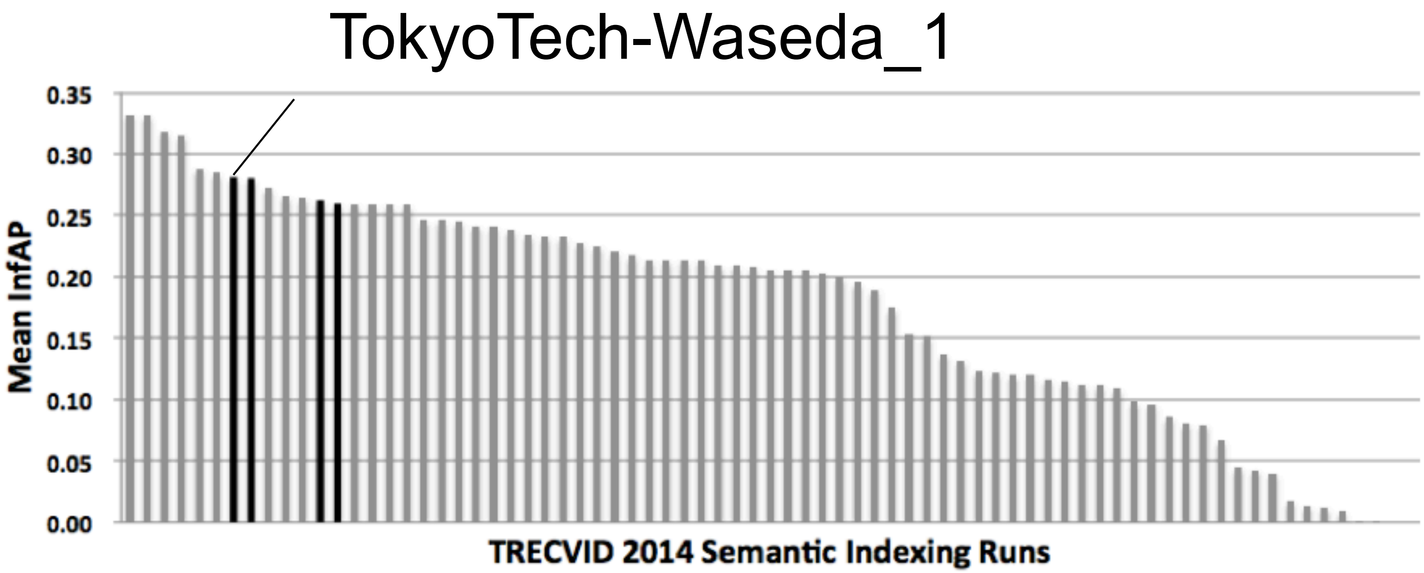
$$s'_i = (1-p)s_i + ps_{\max} \quad p = r\left\langle \frac{\#(\text{positive shots in a video clip})}{\#(\text{shots in a video clip})} \right\rangle$$

Full-gram

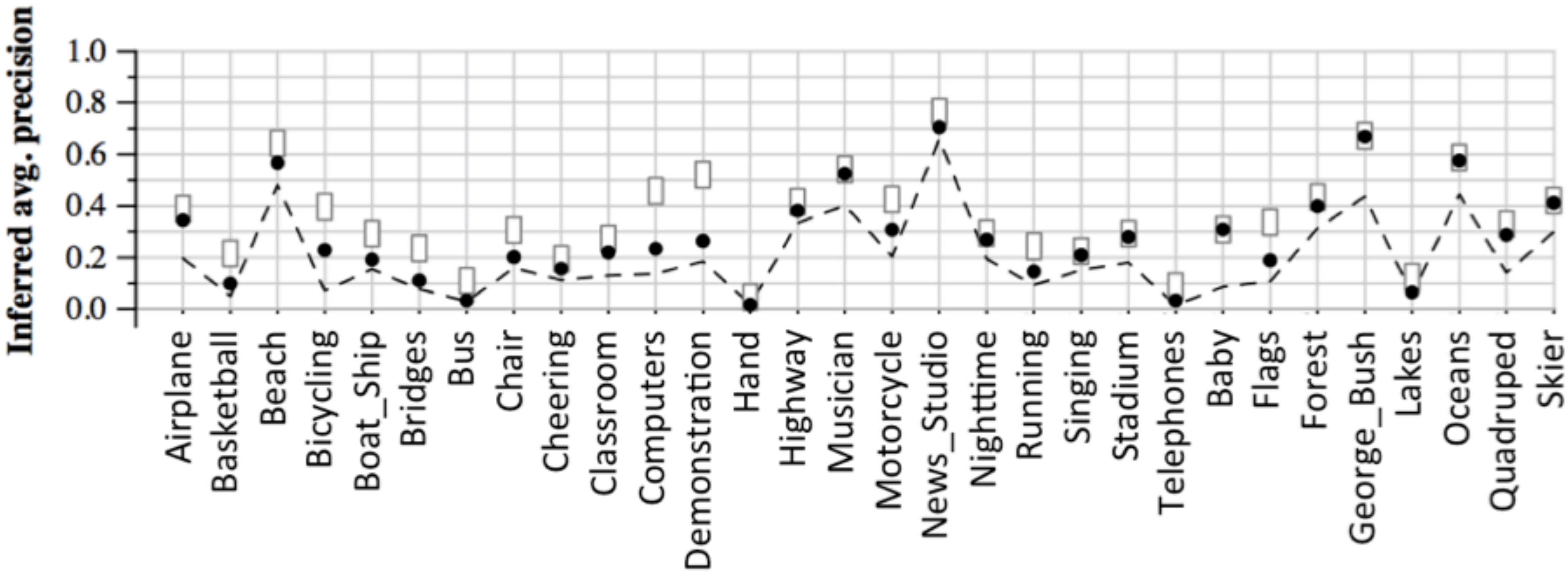


Results

Run ID	Method	Mean InfAP
TokyoTech-Waseda_4	baseline: GMM Supervectors + Full-gram re-scoring	0.260
TokyoTech-Waseda_3	+ sampling	0.262
TokyoTech-Waseda_2	+ Deep CNN	0.280
TokyoTech-Waseda_1	+ Deep CNN (optimized weight)	**0.281**



InfAP by Semantic Concepts



Evaluation of n-Gram Models

- Mean AP on SIN 2012

Method	MeanAP SIN 2012
Baseline	0.306
Bi-gram(n=2)	0.312
Tri-gram(n=3)	0.312
Full-gram	**0.321**

Conclusion (Part 1)

- Deep CNN + GMM Supervector
- n-gram models for re-scoring
- Experimental Results
  - Mean InfAP: 0.281
- Future work
  - Improving audio analysis
  - Introducing motion features for object tracking with deep CNNs

Motion features

- Our baseline system did not include any motion information

  - 5 visual (Har-SIFT, Hes-SIFT, Dense HOG, Dense LBP, and Dense SIFTH) + 1 audio features

- Tried to introduce Dense trajectories into our system

  - Probably effective for some actions / movements. ex.) "Running", "Swimming", "Throwing" and etc.

  - But unfortunately, we could not finish before the submission deadline.

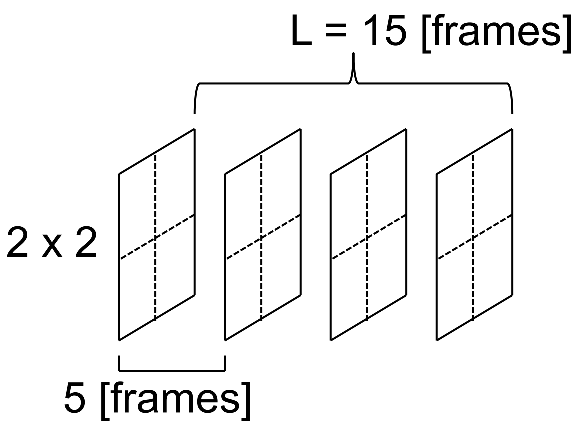
Dense trajectories

- 4 types of features were extracted from each shot
  - Trajectory (a sequence of displacement vectors)
  - HOG (Histogram of Oriented Gradient)
  - HOF (Histogram of Optical Flow)
  - MBH (Motion Boundary Histogram)

Dense trajectories

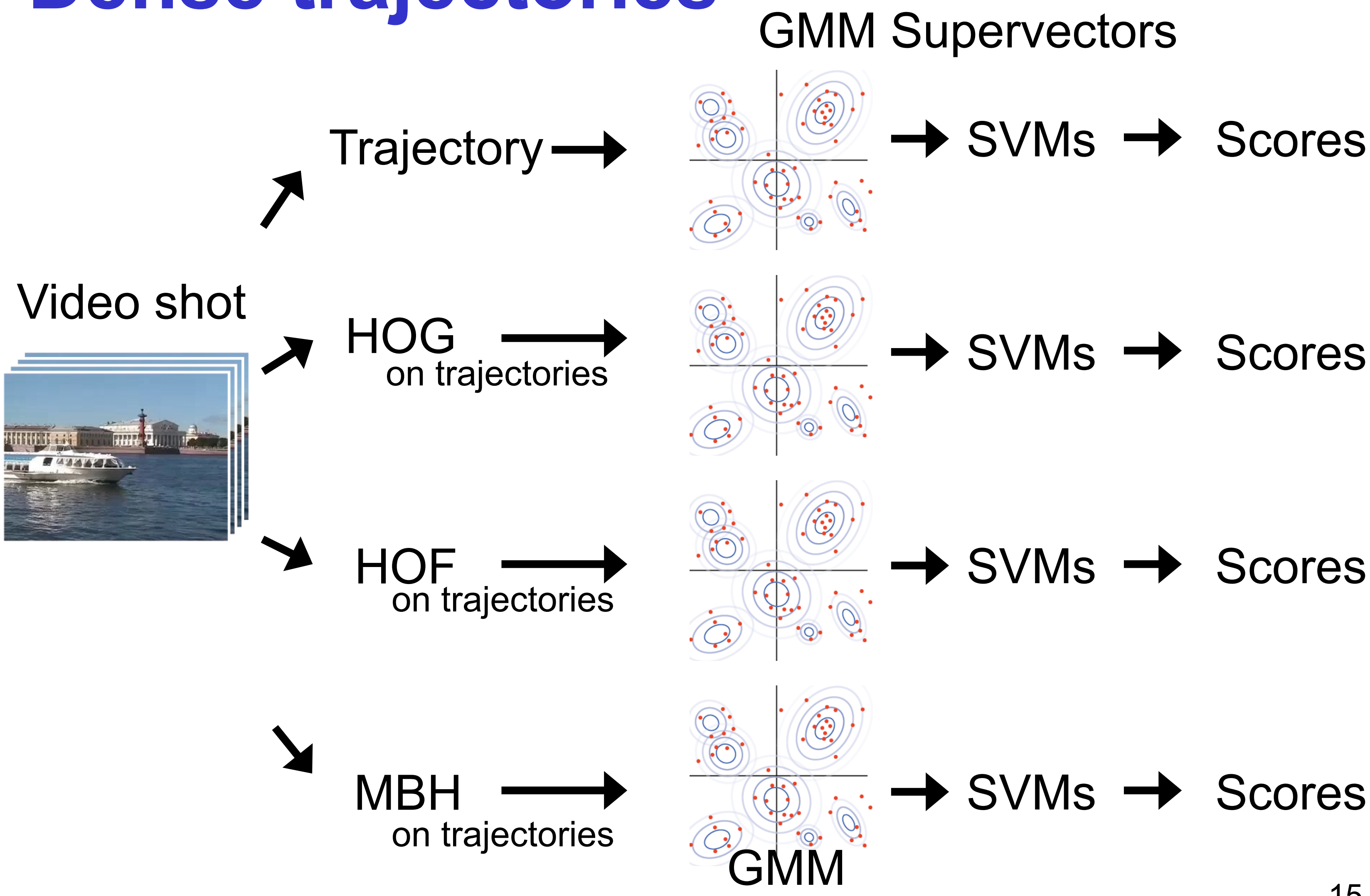
- Setting
  - Use every other frames
  - Trajectory length L=15

    → More than 30 frames are needed to extract features, but about 40% of shots have less than 30 frames…
  - Volume is subdivided into a spatio-temporal grid of size 2 x 2 x 3
  - Orientations are quantized into 8 (or 9) bins.



HOG: 96 dim	PCA →	32 dim
HOF: 108 dim	PCA →	32 dim
MBH: 108x2 dim	PCA →	64 dim

Dense trajectories



Performance of dense trajectories

Mean AP on SIN 2010

Method	MeanAP(%)
Baseline (6 features)	14.07
Trajectory	1.28
HOG on trajectories	8.30
HOF on trajectories	4.79
MBH on trajectories	7.14

Complementarity

Mean AP (%) on SIN 2010

Dense HOG + HOG on trajectories → 9.82

10.90 8.30

Late fusion

- We have not tried the fusion weight optimization, but Dense HOG and HOG on trajectories is not so complementary.

Complementarity

- HOF and MBH are different from other features.
- Finally, we could slightly improve mean AP by combining MBH with our baseline method.

Mean AP (%) on SIN 2010

6 features + MBH on trajectories → **14.29**

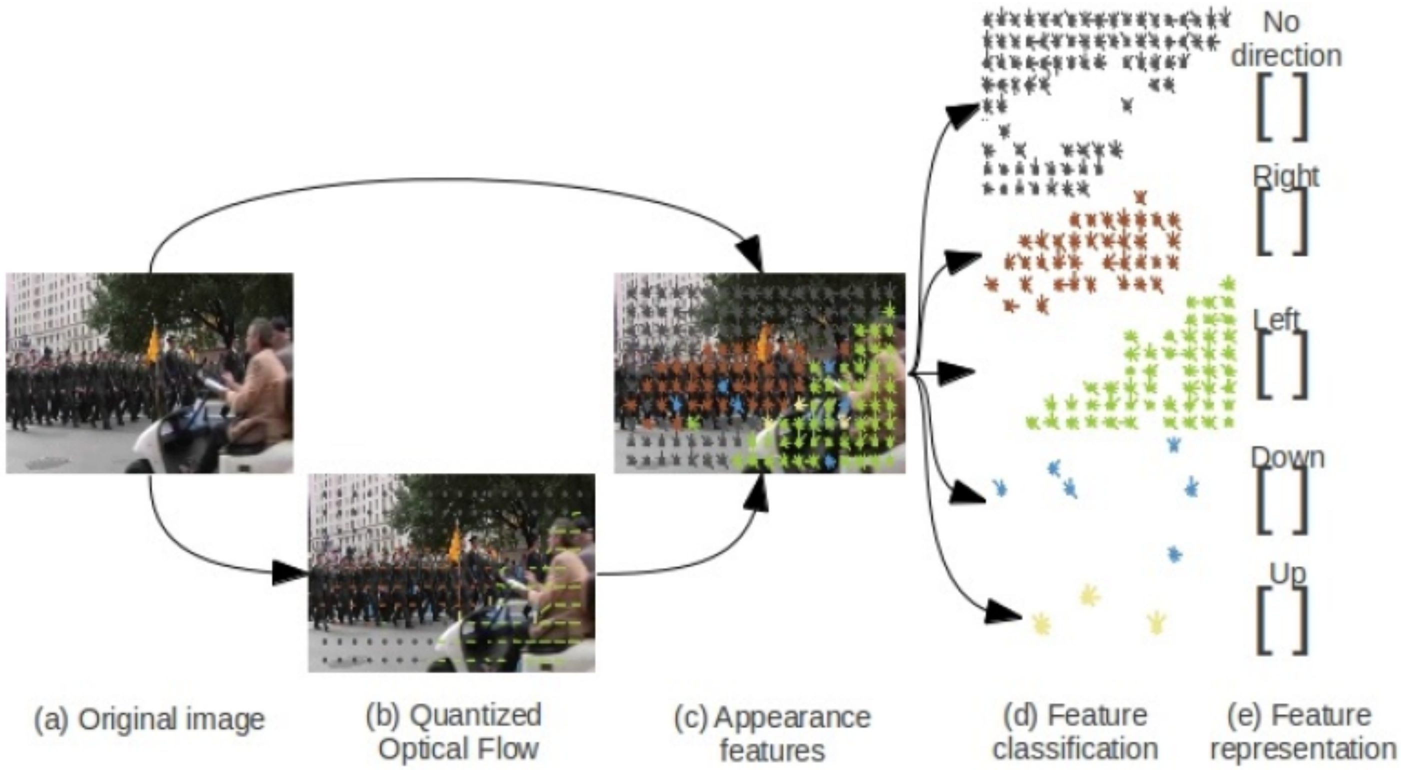
14.07 7.14

Late fusion

(*) no fusion weight optimization

Future work

- Adapt velocity pyramid to dense SIFT/HOG/LBP …



- Motion features with deep CNN