High-Level Feature Extraction Using SIFT GMMs, Audio Models, and MFoM

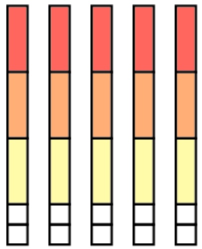
Nakamasa Inoue, Shanshan Hao,
Tatsuhiko Saito, Koichi Shinoda,
Department of Computer Science,
Tokyo Institute of Technology

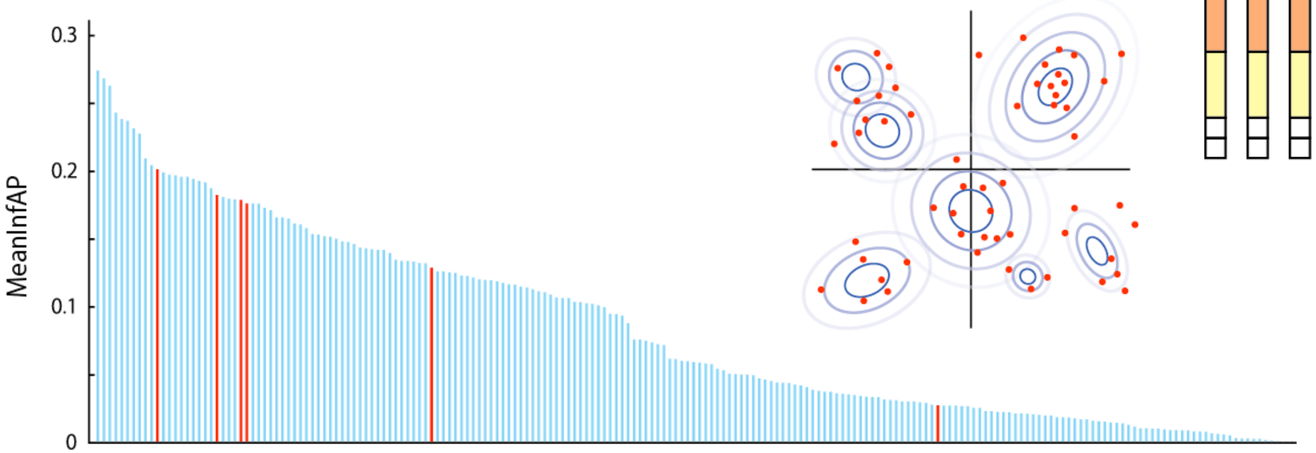
Ilseo Kim,
Chin-Hui Lee,
Department of Computer Science,
Georgia Institute of Technology

Outline

1. SIFT Gaussian mixture models (GMMs) and audio models
2. Text representation of images
3. Multi-Class Maximal Figure-of-Merit (MC MFoM) classifier to combine 1 & 2

Best result: Mean InfAP = 0.168

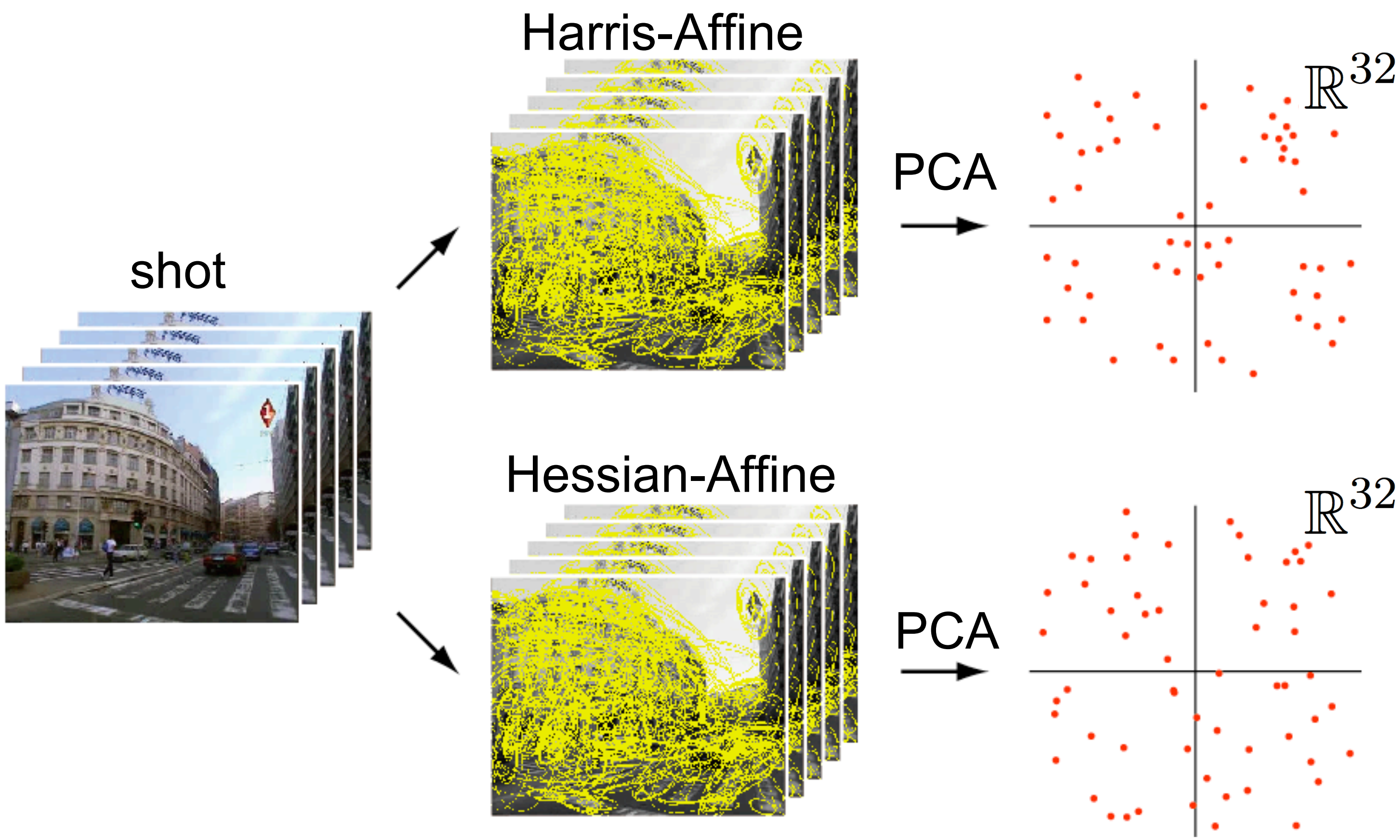




1. SIFT GMMs and Audio Models

SIFT Feature Extraction

- Extract SIFT features from *all the image frames* with Harris-Affine / Hessian-Affine regions.
- Apply PCA to reduce dimension [128dim → 32dim].



SIFT Gaussian Mixture Models

- Model SIFT features by a Gaussian Mixture Model (GMM).

  Robustness against quantization errors that occur in hard-assignment clustering in the BoW approach is expected.

- Probability density function (pdf) of SIFT GMM :

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

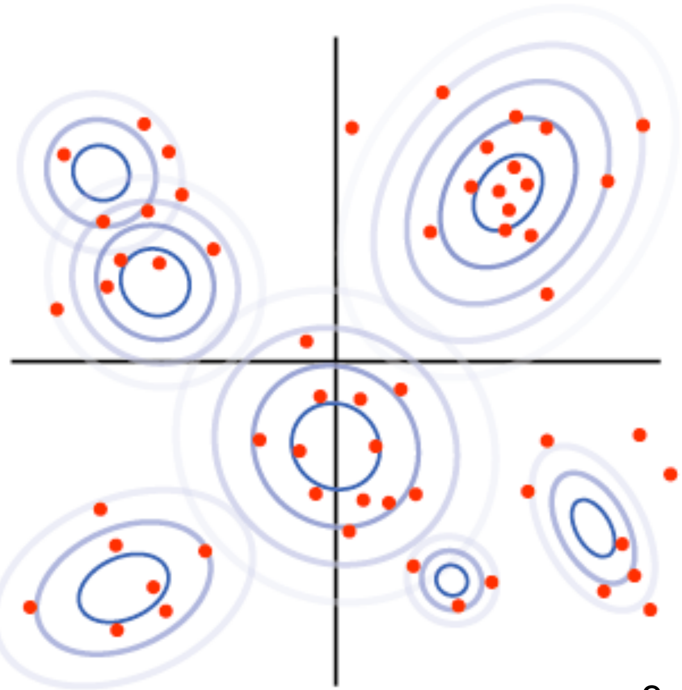
K : num. of mixtures (512)

w_k : mixing coefficient

$\mathcal{N}(x|\mu_k, \Sigma_k)$: pdf of Gaussian

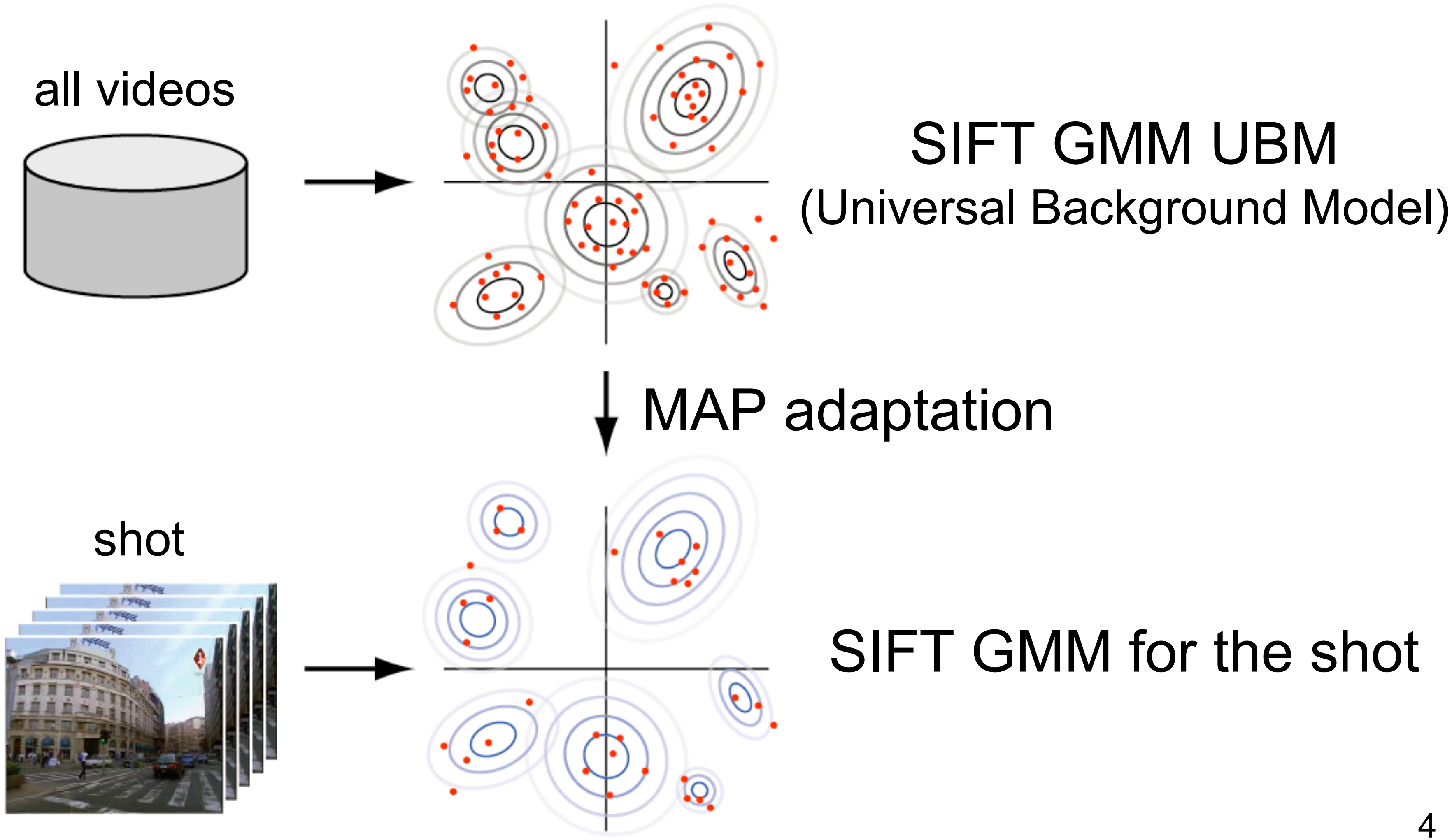
μ_k : mean vector

Σ_k : variance matrix



SIFT Gaussian Mixture Models

- Maximum A Posteriori (MAP) adaptation



Classification

- Distance between SIFT GMMs:

  Weighted sum of Mahalanobis distance

$$d(s,t) = \sum_{k=1}^{K} w_k^{(g)} (\mu_k^{(s)} - \mu_k^{(t)})^T (\Sigma_k^{(g)})^{-1} (\mu_k^{(s)} - \mu_k^{(t)})$$

  $\theta^{(g)}$: UBM, $\theta^{(s)}, \theta^{(t)}$: *s*-th and *t*-th shots

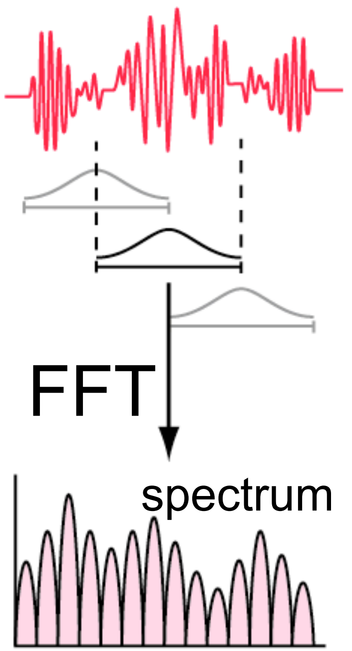
- SVM classification with probability outputs

  Kernel function : $K(s,t) = \exp(-\gamma d(s,t))$

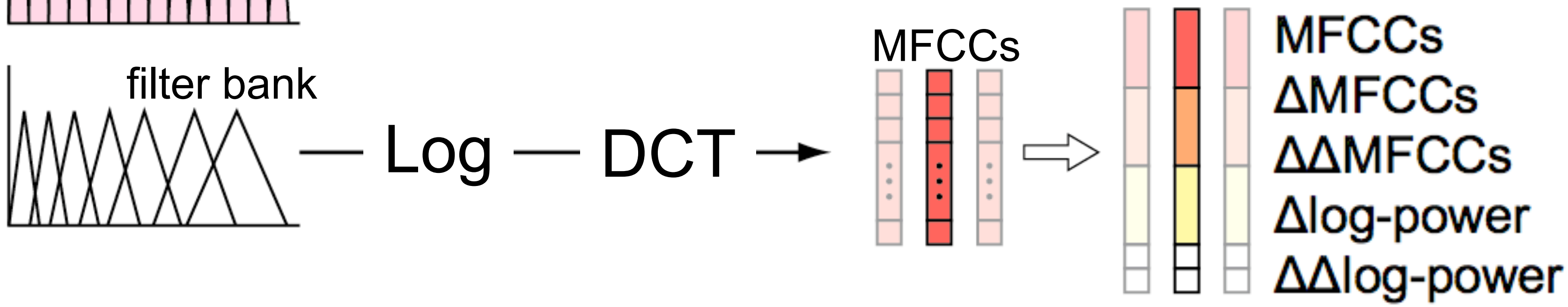
Finally, we obtain posteriori probability $p(h = +1 | X_s)$

Audio Models

- Features: Mel-Frequency Cepstral Coefficients (MFCCs)
- Models: Hidden Markov Models (HMMs)

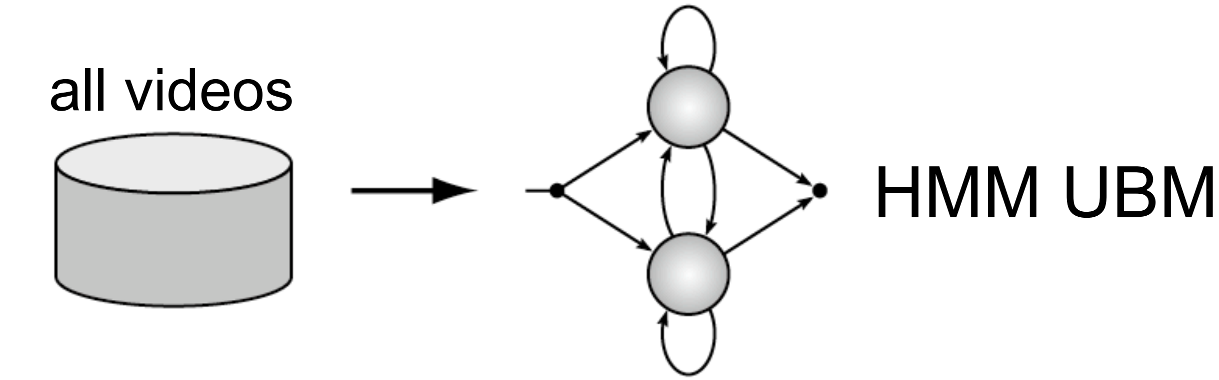


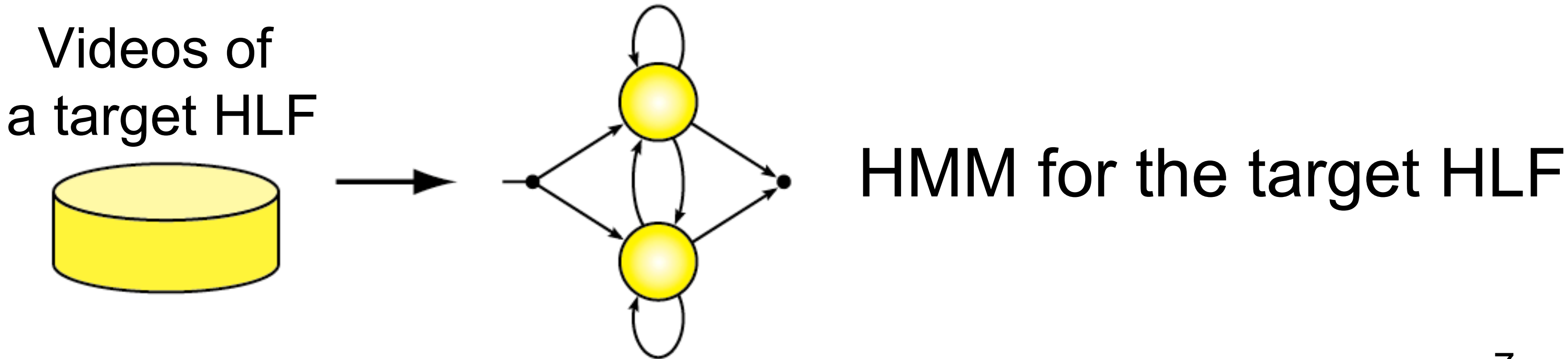
Feature extraction process
1. Frame extraction
2. Windowing [Hamming window]
3. Fast Fourier transform (FFT)
4. Mel scale filter bank
5. Logarithmic transform
6. Discrete cosine transform (DCT)



Hidden Markov Models

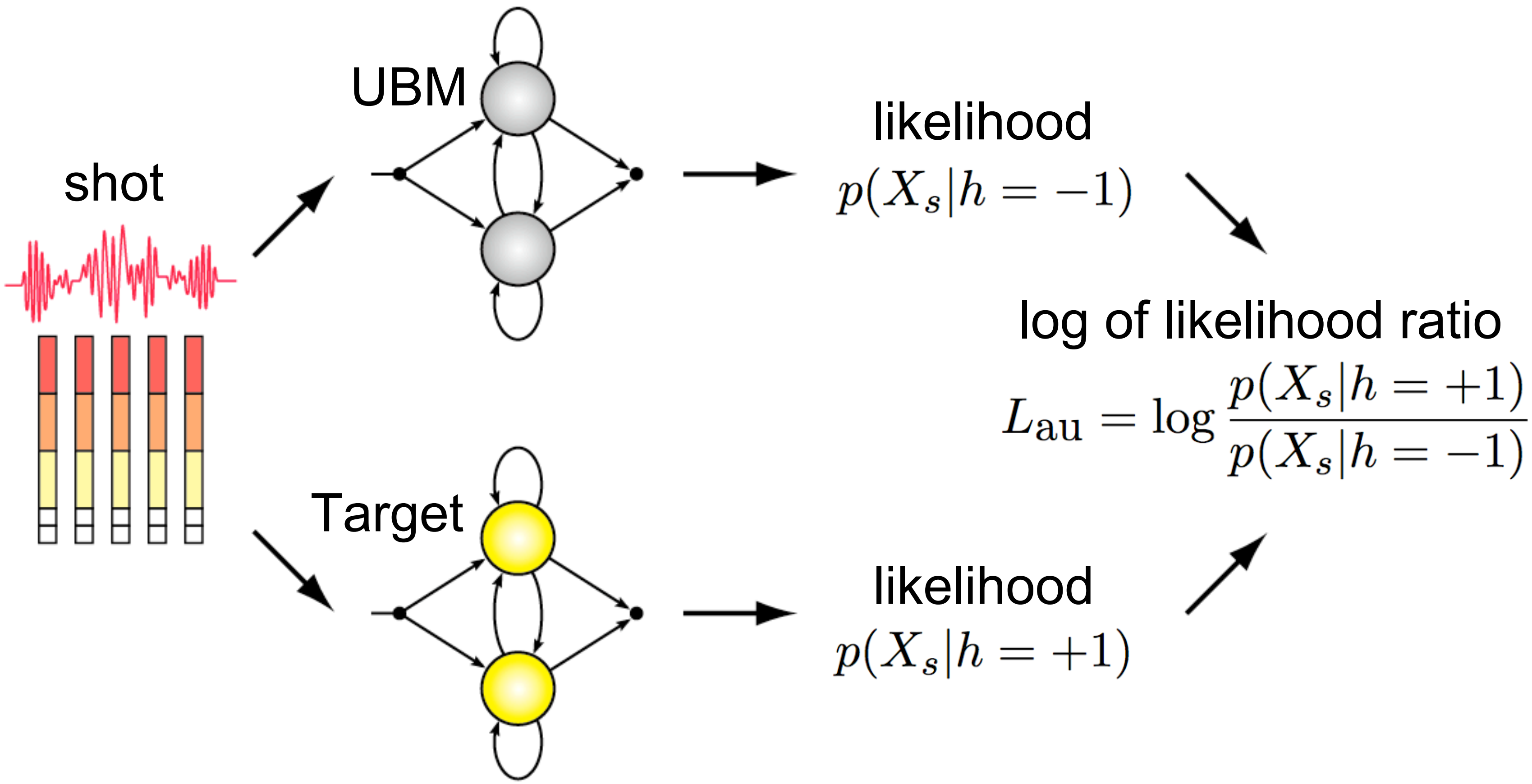
- Ergodic HMMs (2 states, GMMs with 512 mixtures)
- Log of likelihood ratio





Hidden Markov Models

- Ergodic HMMs (2 states, GMMs with 512 mixtures)
- Log of likelihood ratio



Combination of SIFT GMMs and Audio Models

- Outputs from

  audio models L_{au}

  SIFT GMMs with Harris-Affine regions $p_{\text{har}}(h=+1|X_s)$

  SIFT GMMs with Hessian-Affine regions $p_{\text{hes}}(h=+1|X_s)$

- Log of likelihood ratio and posteriori probability

- Combined log of likelihood ratio

$$L = w_{\text{au}}L_{\text{au}} + w_{\text{har}}H(p_{\text{har}}(h=+1|X_s)) + w_{\text{hes}}H(p_{\text{hes}}(h=+1|X_s))$$

where $H(p) = \log \frac{p}{1-p}$

Optimize weight parameters by 2-fold cross validation

Combination of SIFT GMMs and Audio Models

- Outputs from

  audio models L_{au}

  SIFT GMMs with Harris-Affine regions $p_{\mathrm{har}}(h=+1|X_s)$

  SIFT GMMs with Hessian-Affine regions $p_{\mathrm{hes}}(h=+1|X_s)$

- Log of likelihood ratio and posteriori probability

$$
\begin{aligned}
L_{\mathrm{har}} &= \log \frac{p_{\mathrm{har}}(X_s|h=+1)}{p_{\mathrm{har}}(X_s|h=-1)} \\
&= \log \frac{p_{\mathrm{har}}(h=+1|X_s)}{p_{\mathrm{har}}(h=-1|X_s)} \cdot \frac{p_{\mathrm{har}}(h=-1)}{p_{\mathrm{har}}(h=+1)} \\
&= H(p_{\mathrm{har}}(h=+1|X_s)) + \underbrace{H(p_{\mathrm{har}}(h=+1))^{-1}}_{\text{const.}}
\end{aligned}
$$

where $H(p) = \log \frac{p}{1-p}$

Combination of SIFT GMMs and Audio Models

- Outputs from

  audio models L_{au}

  SIFT GMMs with Harris-Affine regions $p_{\mathrm{har}}(h=+1|X_s)$

  SIFT GMMs with Hessian-Affine regions $p_{\mathrm{hes}}(h=+1|X_s)$

- Log of likelihood ratio and posteriori probability
- Combined log of likelihood ratio

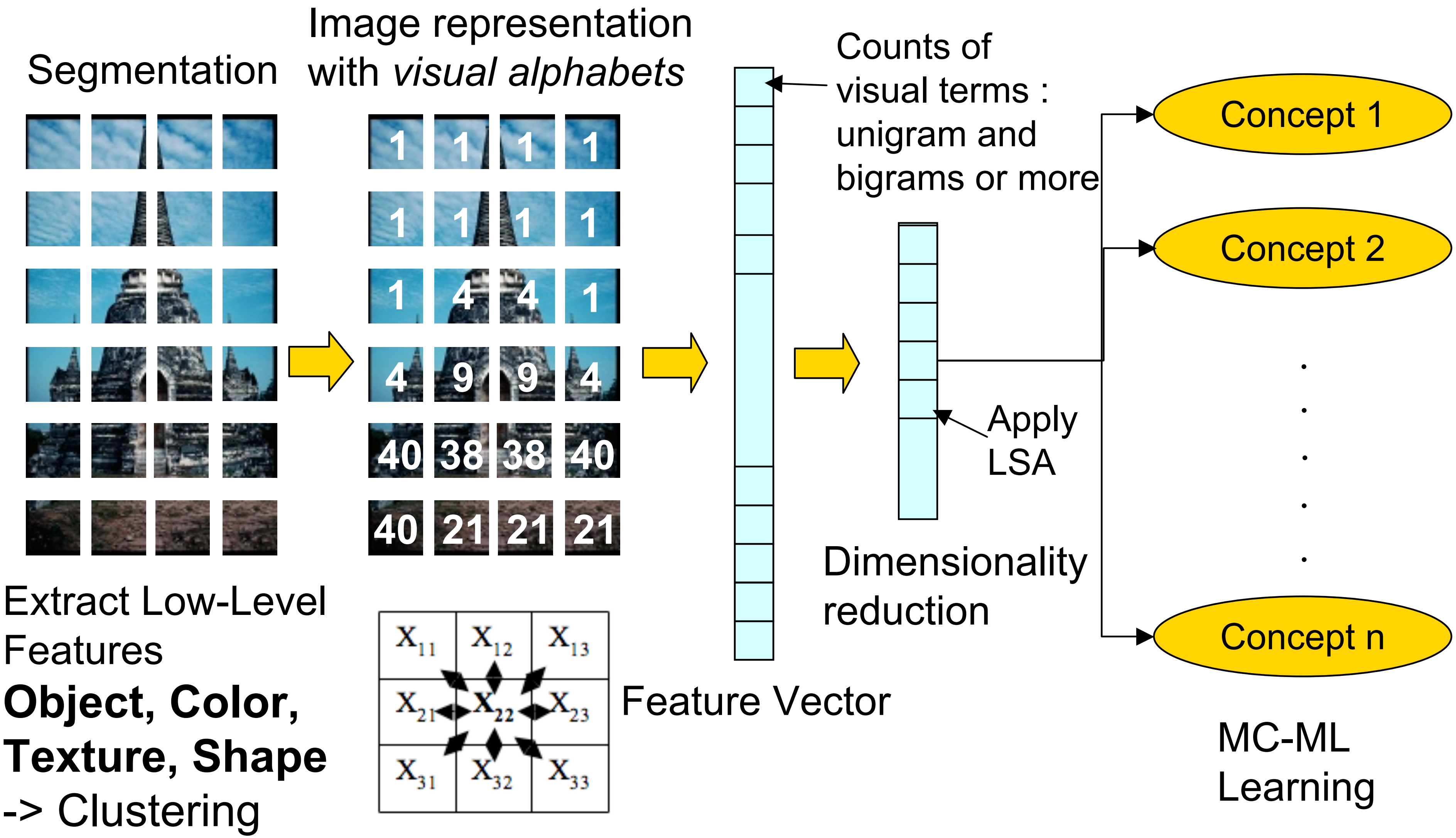
$$L = w_{\mathrm{au}} L_{\mathrm{au}} + w_{\mathrm{har}} H(p_{\mathrm{har}}(h=+1|X_s)) + w_{\mathrm{hes}} H(p_{\mathrm{hes}}(h=+1|X_s))$$

where $H(p) = \log \frac{p}{1-p}$

Optimize weight parameters by 2-fold cross validation

2. Text Representation of Images and MC MFoM Classifier

Text Representation of Images



MC MFoM Classifier

- Multi-Class (MC) learning approach

  MC learning approach can learn a classifier even if there are not enough positive samples like the case of the HLF extraction task in TRECVID2009.

- Maximal Figure-of-Merit (MFoM) Classifier

  MFoM classifier can directly optimize *any objective performance metric* such as m-F1 and MAP by approximating discrete functions to continuous functions, and the GPD algorithm.

MC MFoM Learning Scheme

- The parameter set, $\Lambda = \{\Lambda_j, 1 \leq j \leq N\}$ is estimated by directly optimizing an objective performance metric with a linear classifier, $g_j(X;\Lambda_j) = W_j \cdot X + b_j$.

- Given N concepts, $C = \{C_j, 1 \leq j \leq N\}$ and D-dimensional image representation, $X \in R^D$, the decision rule is

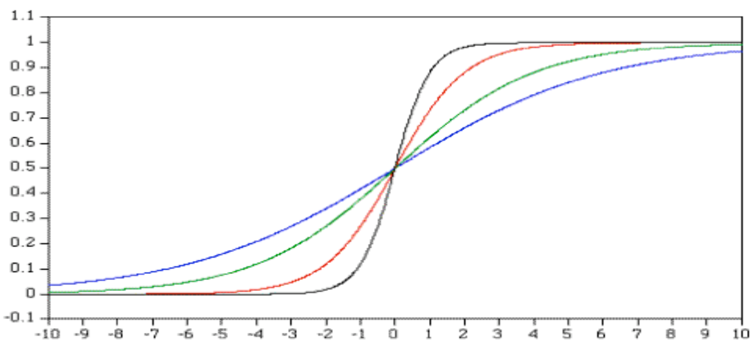
$$\begin{cases} Accept & X \in C_j, if \quad g_j(X;\Lambda_j) - g_j^-(X;\Lambda^-) > 0 \\ Reject & X \notin C_j, Otherwise \end{cases} \qquad 1 \leq j \leq N$$

where $g_j^-(X;\Lambda^-)$ indicates a geometric average for scores of all competing concepts to the concept j.

MC MFoM Learning Scheme

- Misclassification function, $d_j(X;\Lambda) = -g_j(X;\Lambda_j) + g_j^-(X;\Lambda^-)$ is defined where a correct decision is made when $d_j(X;\Lambda) < 0$.
- Approximation of discrete functions to continuous functions by introducing a sigmoid function

$$l_j(X;\Lambda) = \frac{1}{1 + exp\Big(-\alpha\big(d_j(X;\Lambda)+\beta\big)\Big)}$$



$$\begin{cases} TP_j \approx \sum_{X\in T} \big(1 - l_j(X;\Lambda)\big)\cdot 1\big(X \in C_j\big) \\ FP_j \approx \sum_{X\in T} \big(1 - l_j(X;\Lambda)\big)\cdot 1\big(X \notin C_j\big) \\ FN_j \approx \sum_{X\in T} l_j(X;\Lambda)\cdot 1\big(X \in C_j\big) \end{cases}$$

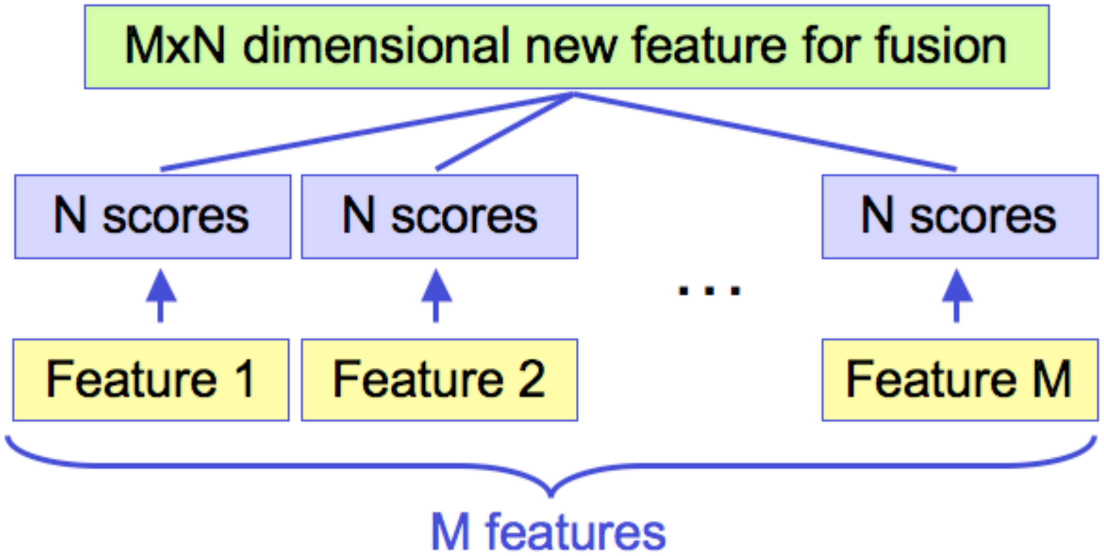
- Now, most commonly used metrics could be represented with the above approximations, and directly optimized with GPD algorithm.

3. MFoM Fusion

Discriminant Fusion Scheme

- Model Based Transformation (MBT) fusion

Given N concepts, N score functions are learned by an MC MFoM classifier. Taking the N score functions as the basis for the transformation, we can obtain a new N-dimensional feature.



A new MC-MFoM classifier can be trained using MxN-dimensional features.

Reference experiment to MFoM fusion

- Rank fusion

  The rank numbers from different systems are combined to get a new rank number:

$$N(x) = \sum_i P_i R_i(x)$$

$R_i(x)$: the rank number of shot x in the ranked output of classification system i

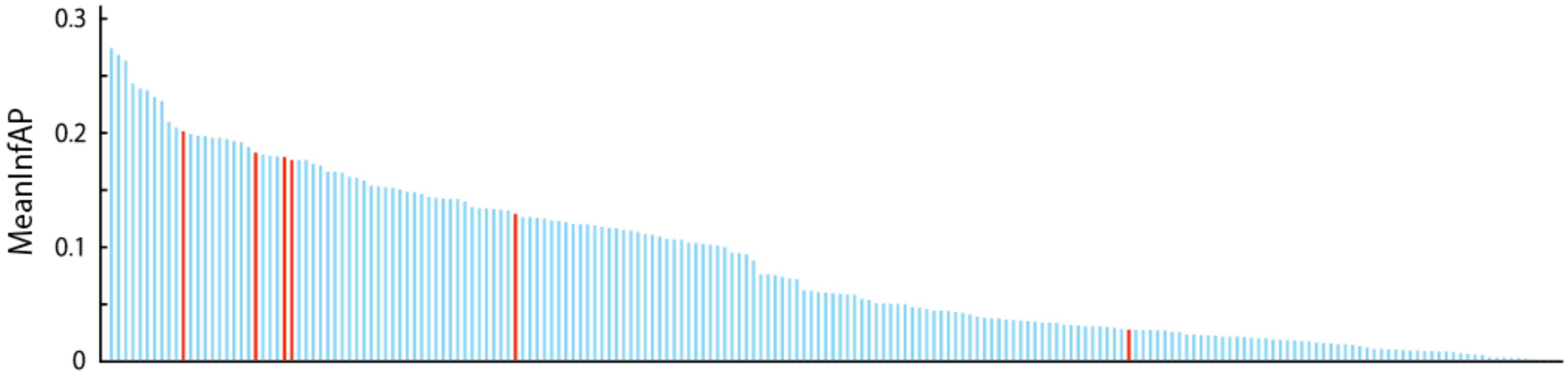
P_i : the weight assignment to system i

2-fold cross validation is used to determine the weight parameters

4. Experiment

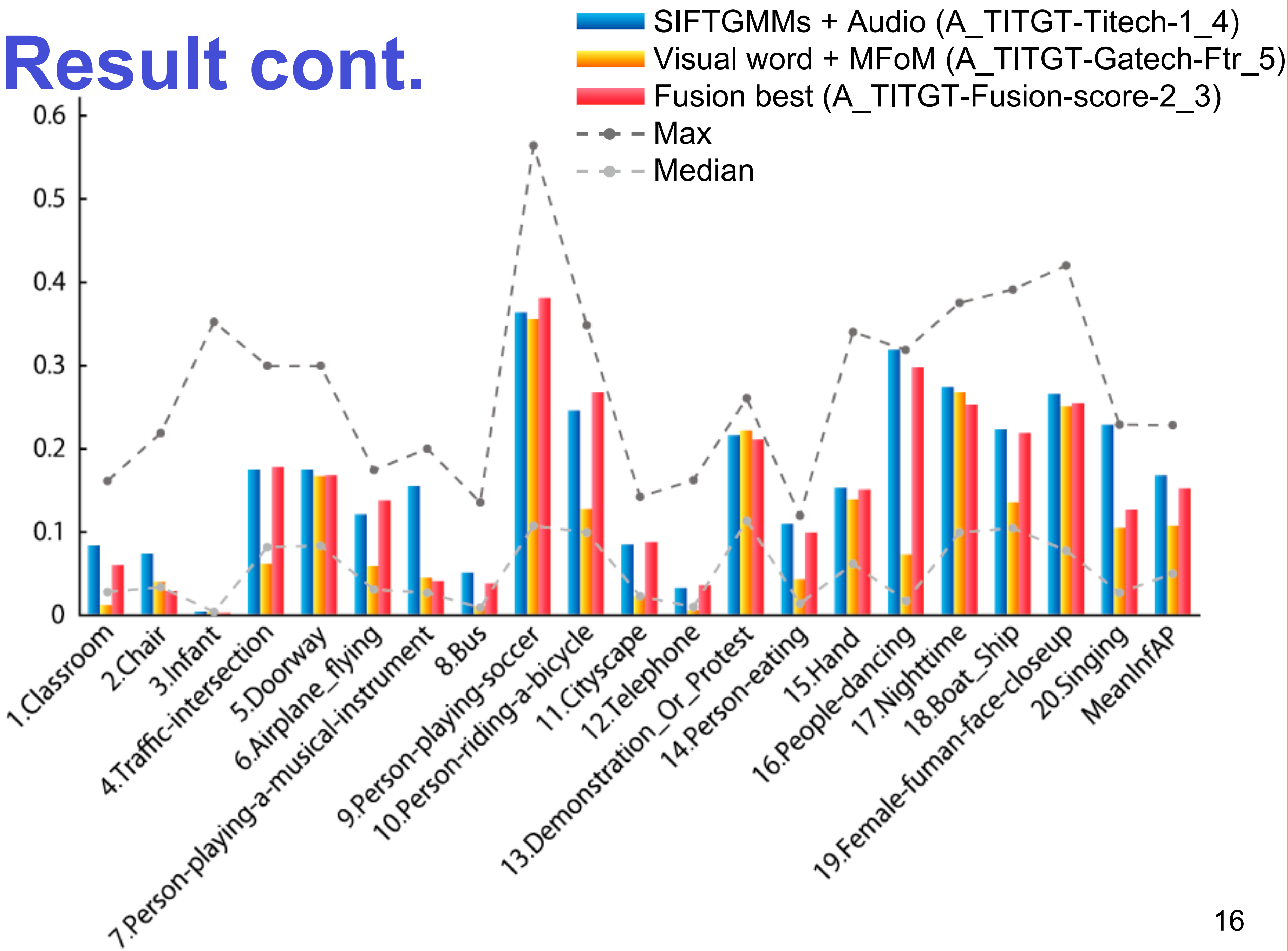
Result

Run name		MInfAP
A_TITGT-Titech-1_4	SIFT GMMs + Audio models (no fusion)	0.168
A_TITGT-Fusion-score-2_3	MFoM (MBT fusion) 1	0.152
A_TITGT-Fusion-score-1_2	MFoM (MBT fusion) 2	0.149
A_TITGT-Fusion-rank_1	Rank fusion	0.147
A_TITGT-Gatech-Ftr_5	Visual word + MFoM (no fusion)	0.108
A_TITGT-Titech-1_6	Local + Global features (no fusion)	0.023

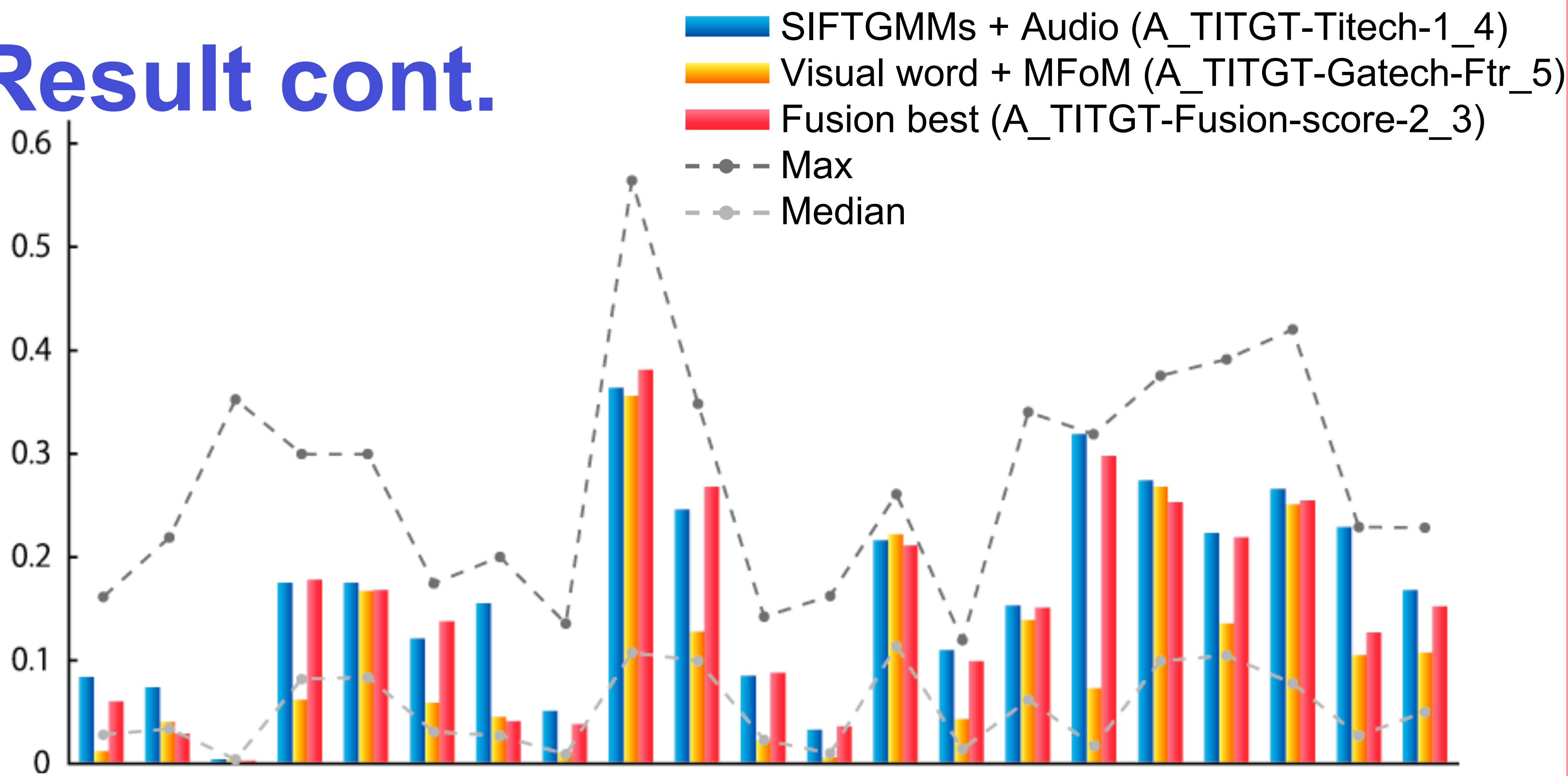


- MeanInfAP of SIFT GMMs + Audio models was 0.168, which is ranked 11th of all A-type runs and 4th among all participating teams.
- The MFoM fusion works better than the rank fusion.

Result cont.


SIFTGMMs + Audio (A_TITGT-Titech-1_4)
Visual word + MFoM (A_TITGT-Gatech-Ftr_5)
Fusion best (A_TITGT-Fusion-score-2_3)
Max
Median
0.6
0.5
0.4
0.3
0.2
0.1
0
1.Classroom
2.Chair
3.Infant
4.Traffic-intersection
5.Doorway
6.Airplane_flying
7.Person-playing-a-musical-instrument
8.Bus
9.Person-playing-soccer
10.Person-riding-a-bicycle
11.Cityscape
12.Telephone
13.Demonstration_Or_Protest
14.Person-eating
15.Hand
16.People-dancing
17.Nighttime
18.Boat_Ship
19.Female-fuman-face-closeup
20.Singing
MeanInfAP

Result cont.



- Combination with audio is effective for the HLF extraction.

  Good : Singing (0.229), People-dancing (0.319),
  People-playing-a-musical-instruments (0.155),
  Female-human-face-closeup (0.266).

- SIFT GMMs represent HLFs with the background.

  Good : Airplane_flying (0.138), Boat_Ship (0.250).

Conclusion

- Combination of SIFT GMMs and audio models is effective for the HLF extraction (Mean InfAP = 0.168).
  - SIFT GMMs work well for various HLFs.
  - Audio models can detect HLFs complementary.
- It is difficult to make a fusion of different systems.

Future work

- More improved collaboration work
- Using time/spatial region information