# Multimedia Event Detection using GS-SVMs and Audio-HMMs

Nakamasa Inoue, Yusuke Kamishima,
Koichi Shinoda,
*Department of Computer Science,*
*Tokyo Institute of Technology*

Shunsuke Sato
*Canon Inc.*

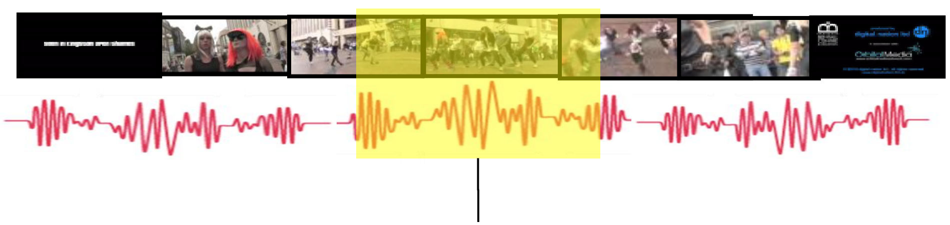
# Outline

- Motivation
- System Overview
- Method
  - Features extraction
  - GS-SVM
  - Audio HMMs
- Results
  - Best result: Minimum NDC = 0.525

# Motivation

- Two event feature categories:
  - Features that appear in every frame
  - Features that appear only in some frames
- Their combination can improve the detection performance.

ex.) Flash Mob Gathering clips



Every frame:
•Outdoor
•Dancers
•Road
•Crowd
•Crowd buzz
…

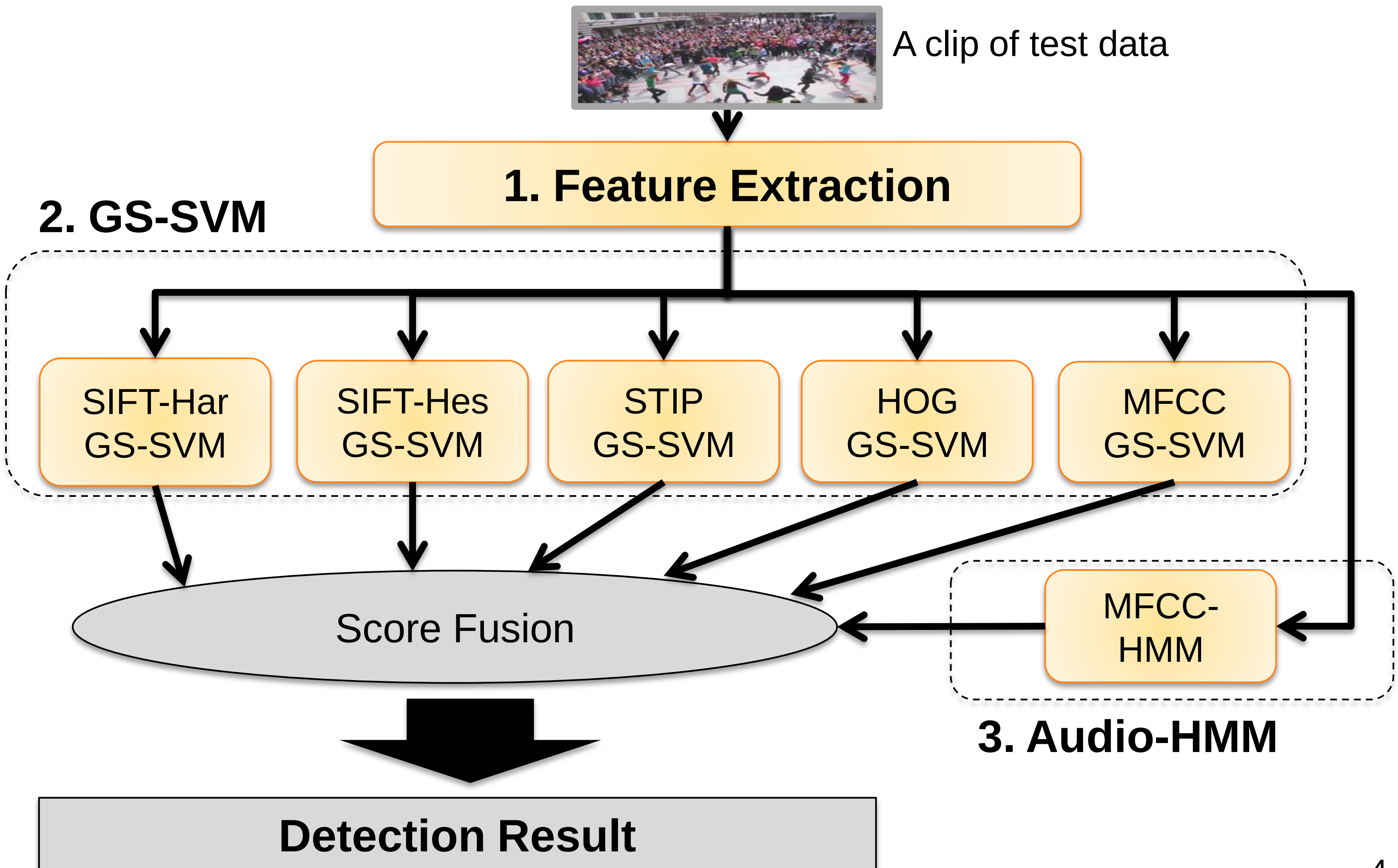
Some frames:
•Dancing
•Dance music
•Cheering voice

# Method Overview

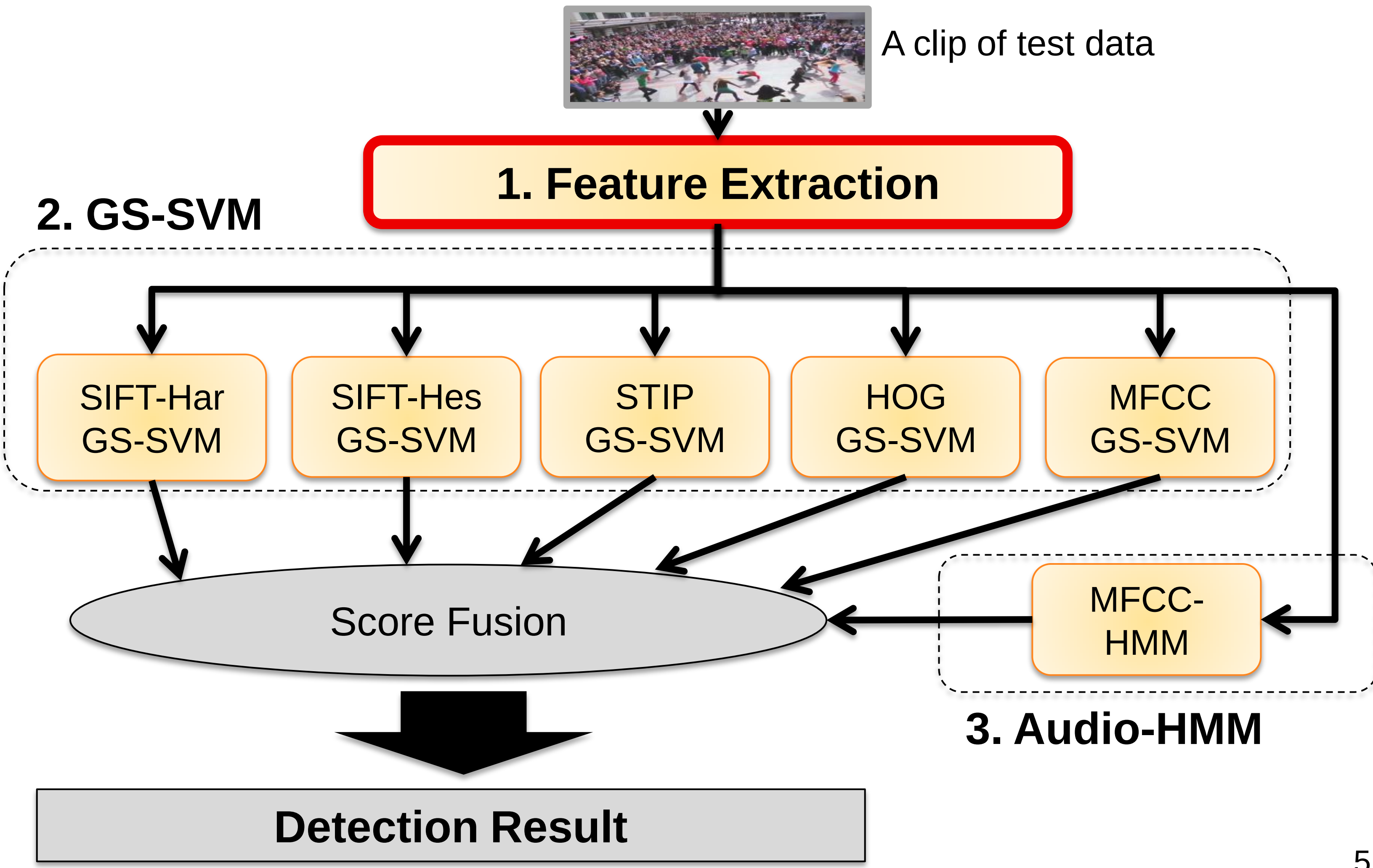
- For every-frame features: GS-SVM
  (GMM-Supervector Support Vector Machine)
  - Use several visual and audio features
  - Soft clustering - robust against quantization errors
  - Based on our system of TRECVID 2010 SIN task

- For some-frame features: HMM
  (Hidden Markov model)
  - Model temporal features in sound
  - Apply word-spotting in speech recognition
  - Use only audio, not video

# System Overview



# System Overview



# Feature Extraction

- 5 types of features, from 3 kinds of sources

clip

**Still images**
frames every
2 seconds

- SIFT(Harris)
- SIFT(Hessian)
- HOG

**Spatio-temporal image**


- STIP

**Audio**

- MFCC

# List of Features

<table>
<tr><th>source</th><th>feature</th><th>description</th></tr>
<tr><td rowspan="3">Still images</td><td>SIFT (Harris)</td><td rowspan="2">Scale-Invariant Feature Transform<br>with Harris-affine regions<br>and Hessian-affine regions<br>[Mikolajczyk, 2004]</td></tr>
<tr><td>SIFT (Hessian)</td></tr>
<tr><td>HOG</td><td>32 dimensional HOG<br>Dense sampling (every 4 pixels)</td></tr>
<tr><td>Spatio-temporal images</td><td>STIP</td><td>Space-Time Interest Points<br>HOG and HOF features extracted<br>[Laptev, 2005]</td></tr>
<tr><td>Audio</td><td>MFCC</td><td>Mel-frequency cepstral coefficients<br>Audio features for speech recognition</td></tr>
</table>

# System Overview

A clip of test data

1. Feature Extraction

2. GS-SVM

SIFT-Har GS-SVM

SIFT-Hes GS-SVM

STIP GS-SVM

HOG GS-SVM

MFCC GS-SVM

MFCC-HMM

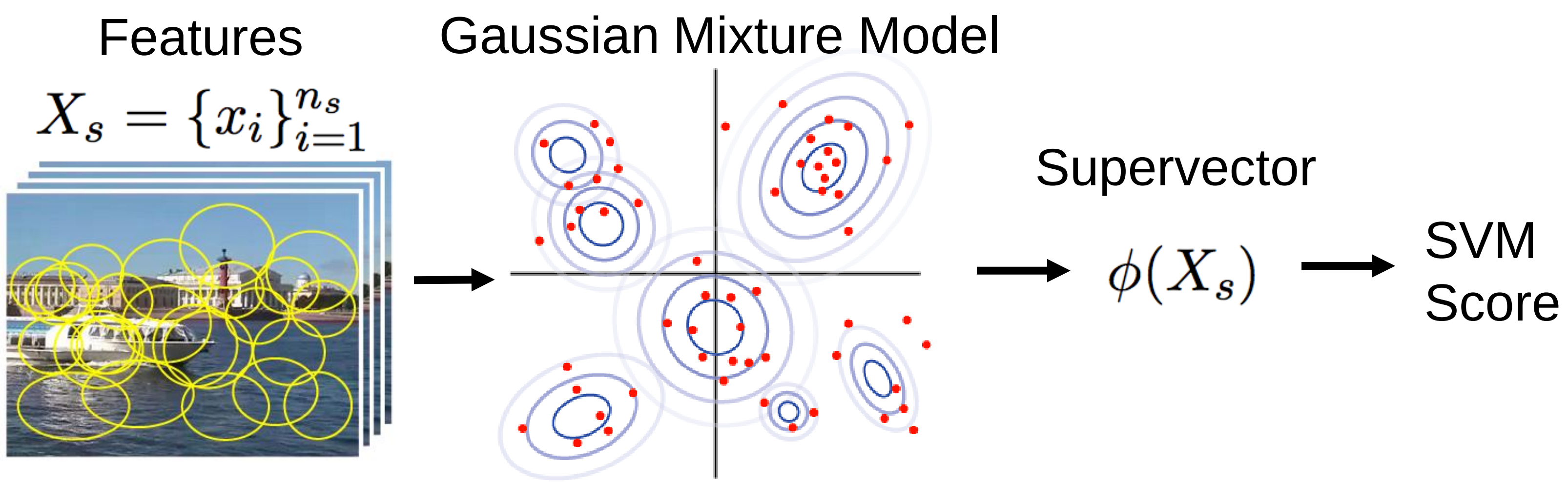
3. Audio-HMM

Score Fusion

Detection Result

# GMM Supervector SVM (GS-SVM)

- Represent the distribution of each feature
  - Each clip is modeled by a GMM (Gaussian Mixture Model)
  - Derive a supervector from the GMM parameters
  - Train SVM (Support Vector Machine) of the supervectors



# GMM Estimation

- Estimated by using maximum a posteriori (MAP) adaptation for mean vectors:

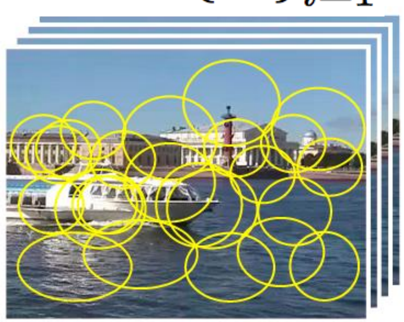
$$\hat{\mu}_k^{(s)} = \frac{\tau \mu_k^{(U)} + \sum_{i=1}^{n_s} c_{ik} x_i}{\tau + C_k}$$

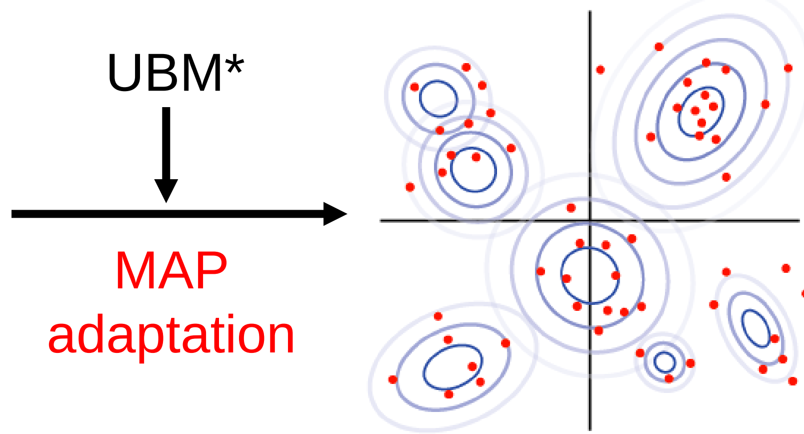
adapted mean

UBM's mean

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

$$X_s = \{x_i\}_{i=1}^{n_s}$$






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

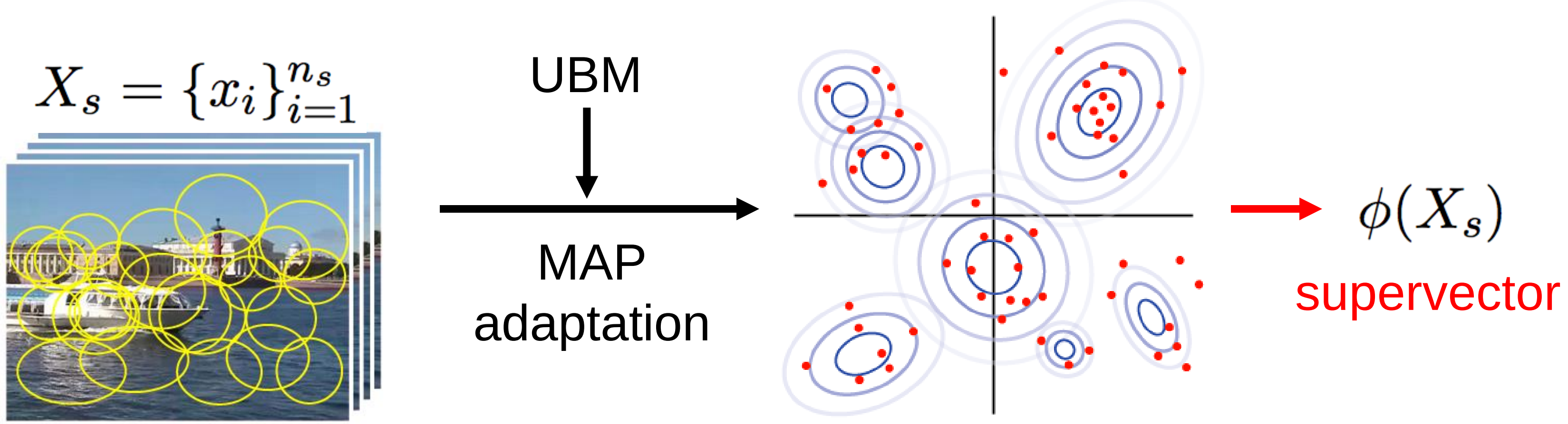
# GMM Supervector

- GMM Supervector: combination of the mean vectors.

$$\phi(X_s) = \begin{pmatrix} \tilde{\mu}_1^{(s)} \\ \tilde{\mu}_2^{(s)} \\ \vdots \\ \tilde{\mu}_K^{(s)} \end{pmatrix}$$

where

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



# Score Fusion in GS-SVM

- GS-SVMs use RBF-kernels:

$$k(X_{\mathrm{F}}, X'_{\mathrm{F}}) = \exp\left(-\gamma \|\phi(X_{\mathrm{F}}) - \phi(X'_{\mathrm{F}})\|_2^2\right),$$

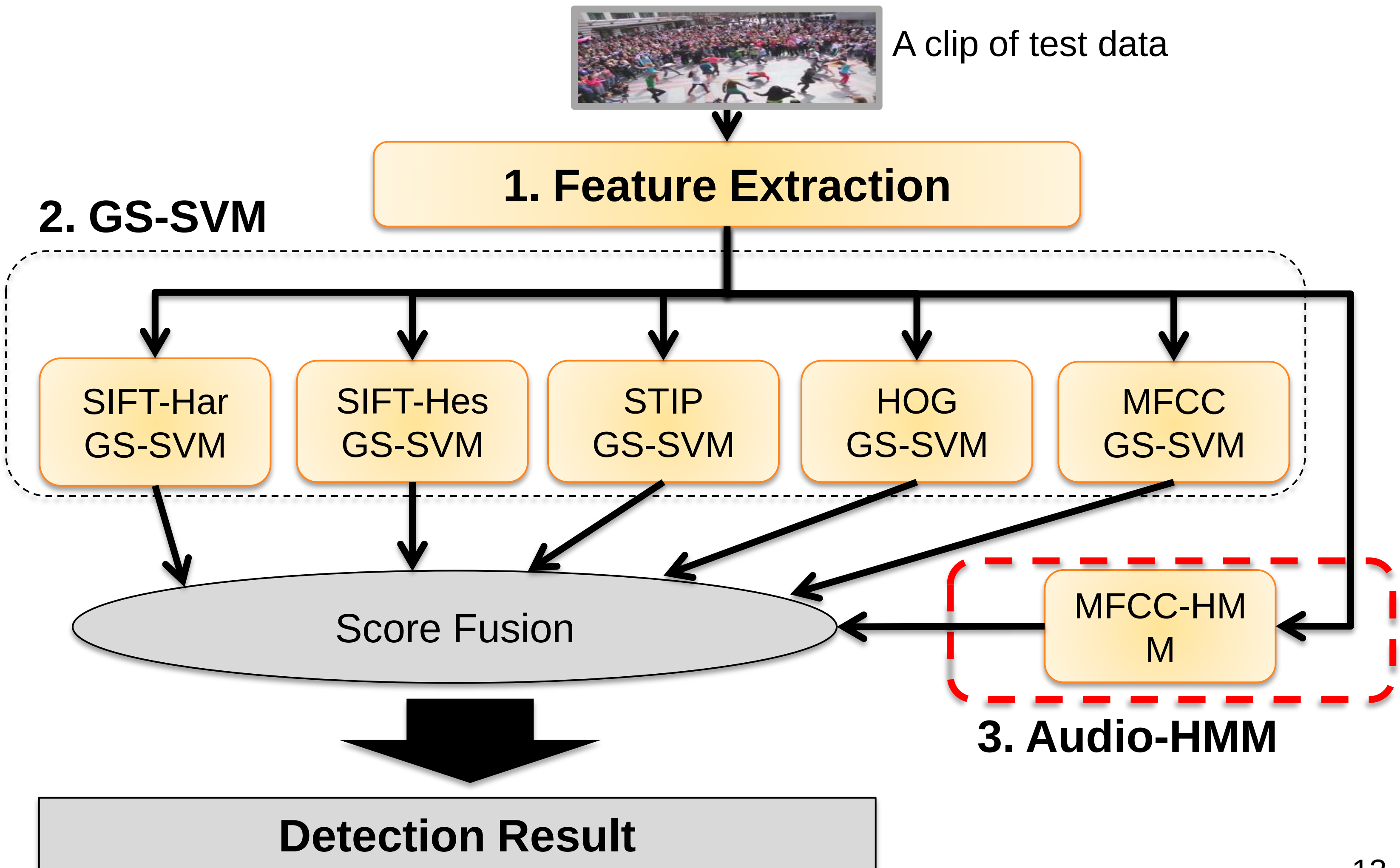
- Score: Weighted Average of SVM outputs:

$$f(X) = \sum_{\mathrm{F} \in \mathcal{F}} \alpha_{\mathrm{F}} f_{\mathrm{F}}(X_{\mathrm{F}}), \quad 0 \le \alpha_{\mathrm{F}} \le 1, \quad \sum_{\mathrm{F}} \alpha_{\mathrm{F}} = 1$$

where $\mathcal{F}$ = {SIFT-Her, SIFT-Hes, HOG, STIP, MFCC}

- $\alpha_{\mathrm{F}}$ are decided by 2-fold cross validation based on
  - Minimum Normalized Detection Cost - Run 1 & Run 2
  - Average Precision - Run 3
  - In Run 4, $\alpha_{\mathrm{F}}$ is equal for all features

# System Overview



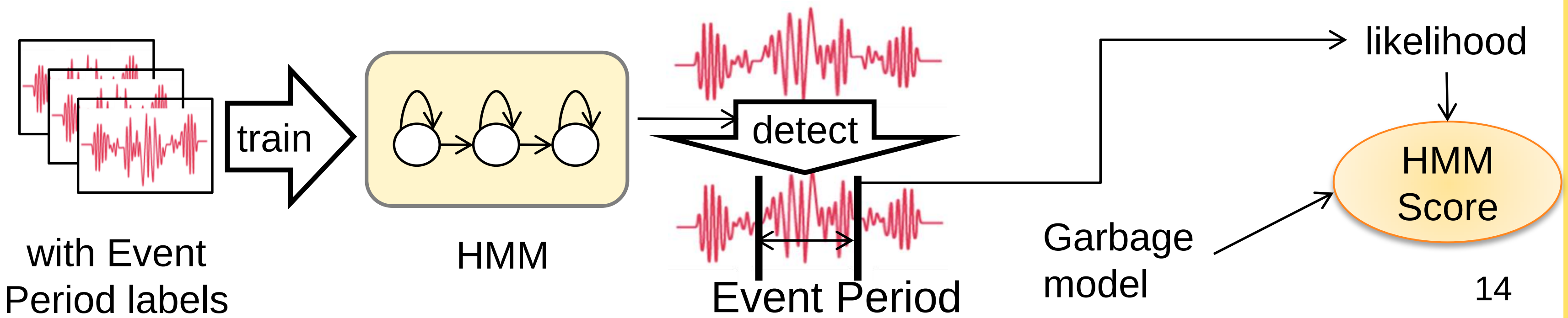
# Audio HMM

Training:

1. Label an event period manually for each event clip
2. Train an event HMM using MFCC

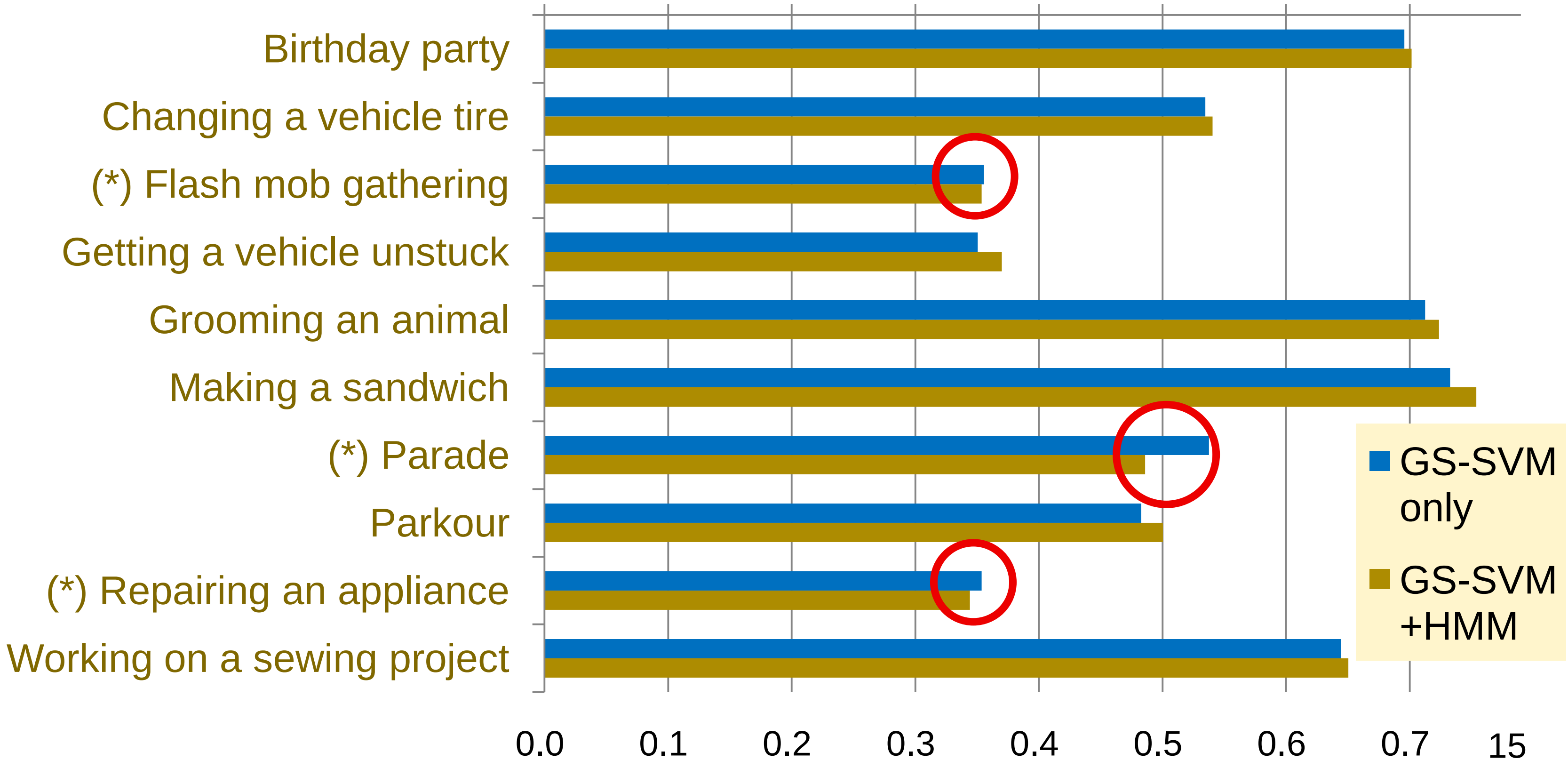
Test:

1. Find likelihood $L_E$ of the event period by word-spotting
2. Find likelihood $L_G$ of the event period for a garbage model estimated from all video data
3. Calculate likelihood ratio $L_E/L_G$ as the detection score

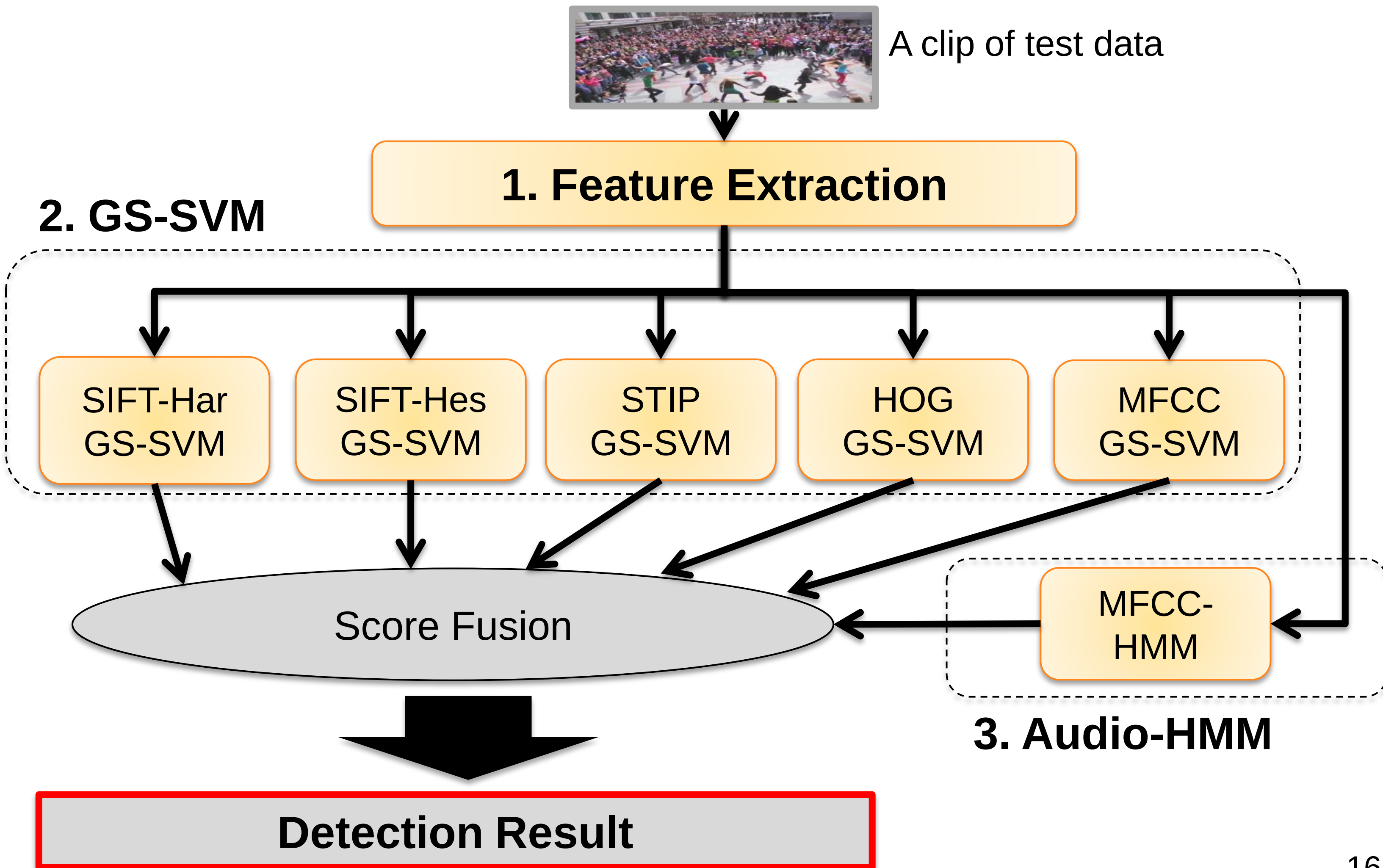


# Preliminary result of Audio HMMs

- Fuse HMM score with GS-SVM by weighted average.
- Audio HMMs are effective in 3 events – Use them in Run1.

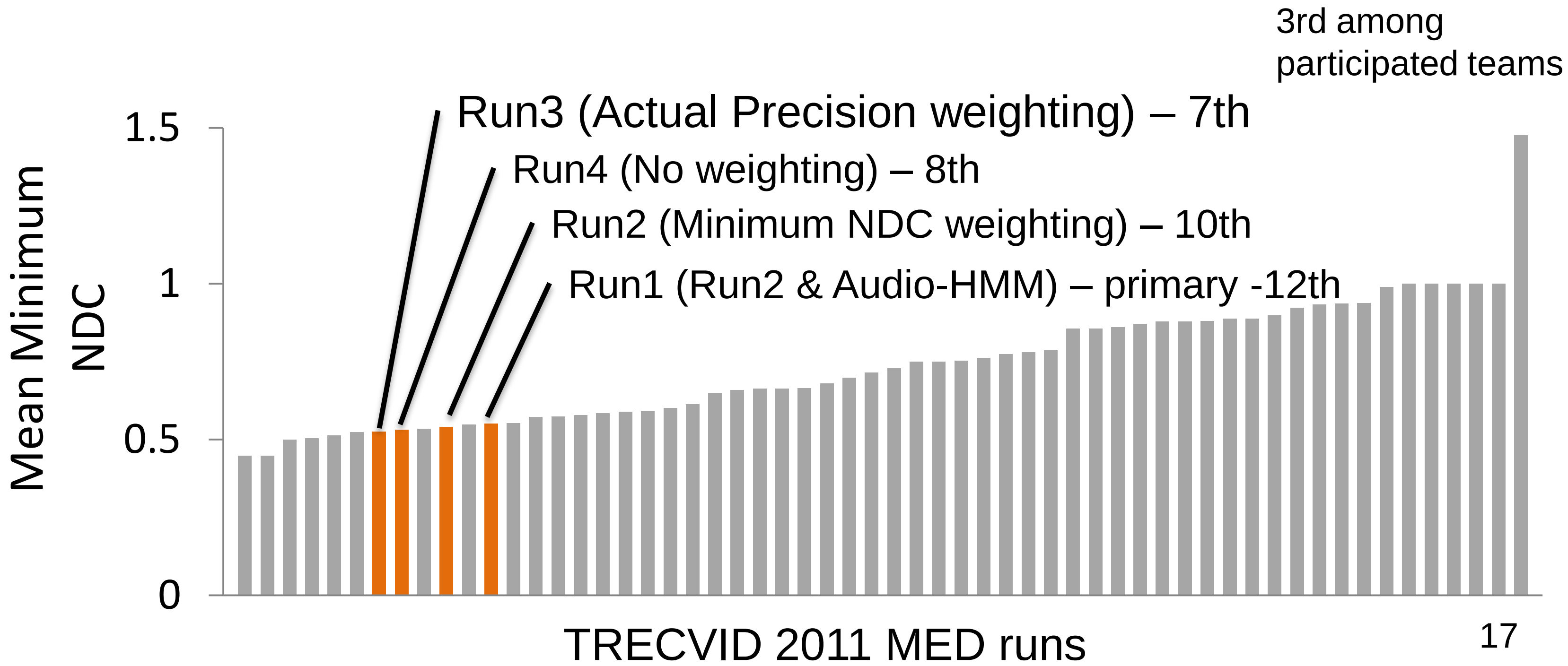


# System Overview



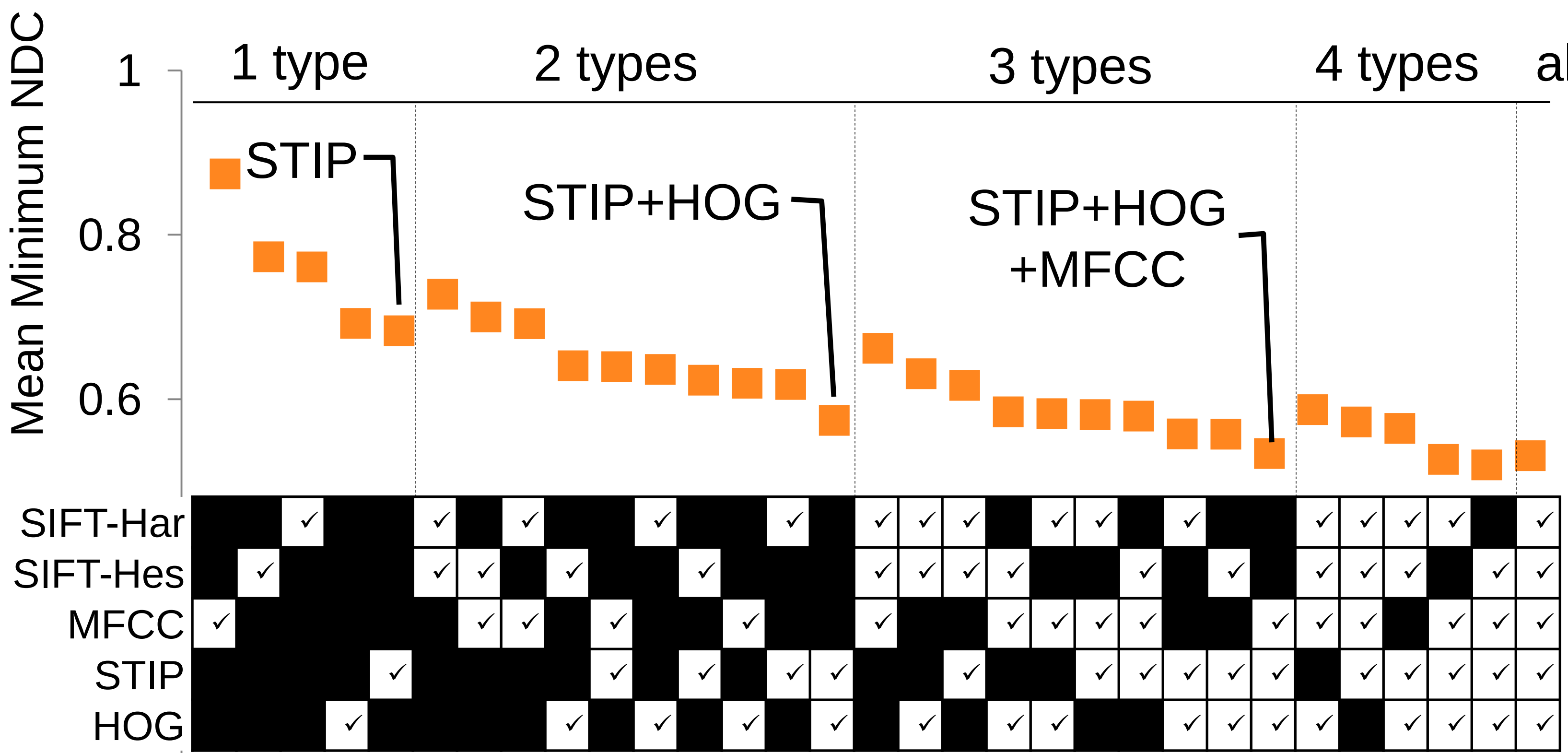
# Experiments

- Run3 was the best. GS-SVM was effective
- Run1 (Audio-HMM) did not show good performance
- Run2, weights decided by Minimum NDC, is not good
  - Simple cross validation may have failed.



# Effect of each feature in GS-SVM

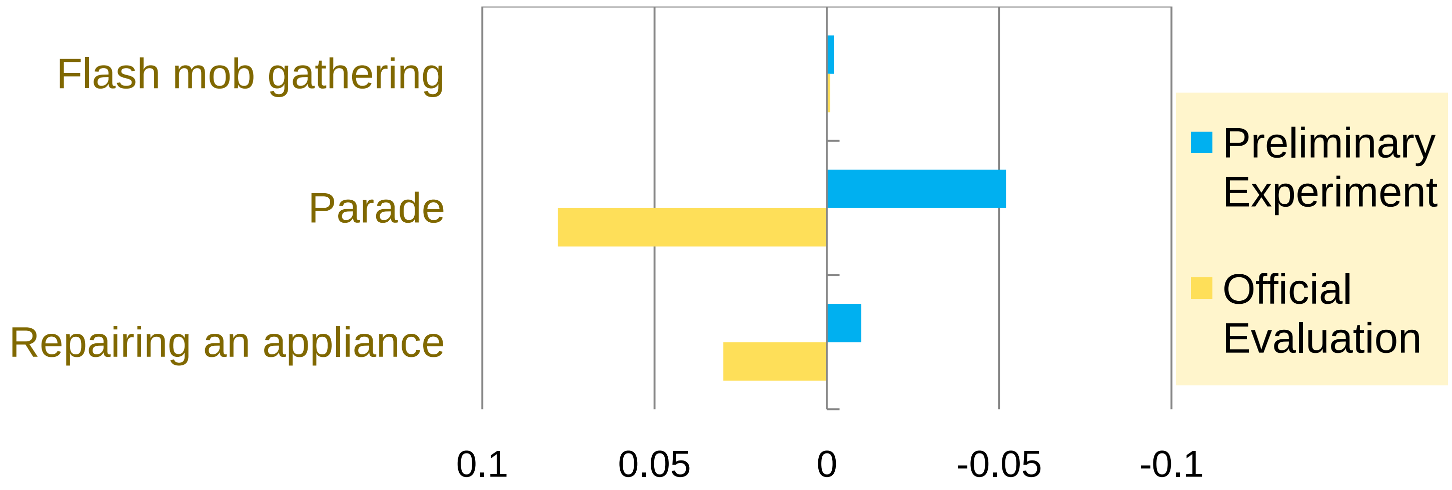
- STIP and HOG had better performance.
- MFCC was effective when combined with STIP and HOG.




Checked: used Black: not used

# Why Audio HMM did not work?

- It failed to capture temporal features
  - Each state represents a specific sound such as drum, cheering, which may appear in non-event and/or at random.
- Test data include many sounds not appear in training and development data




**Difference of Minimum NDC between with and without Audio HMMs**

# Conclusion

- We combine GS-SVM and Audio HMM
- GS-SVMs are effective for MED.
  - STIP, HOG, and MFCC are important
- Audio HMMs are not effective
  - It cannot capture temporal features
  - Variety of sounds are larger than expected

- Future works
  - Include other features, such as Dense SIFT
  - Improve the HMM-based sound detection
    - Model event subclasses and their relationship