Multimedia Event Detection Using GMM Supervectors and Camera Motion Cancelled Features

Yusuke Kamishima, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology

Outline

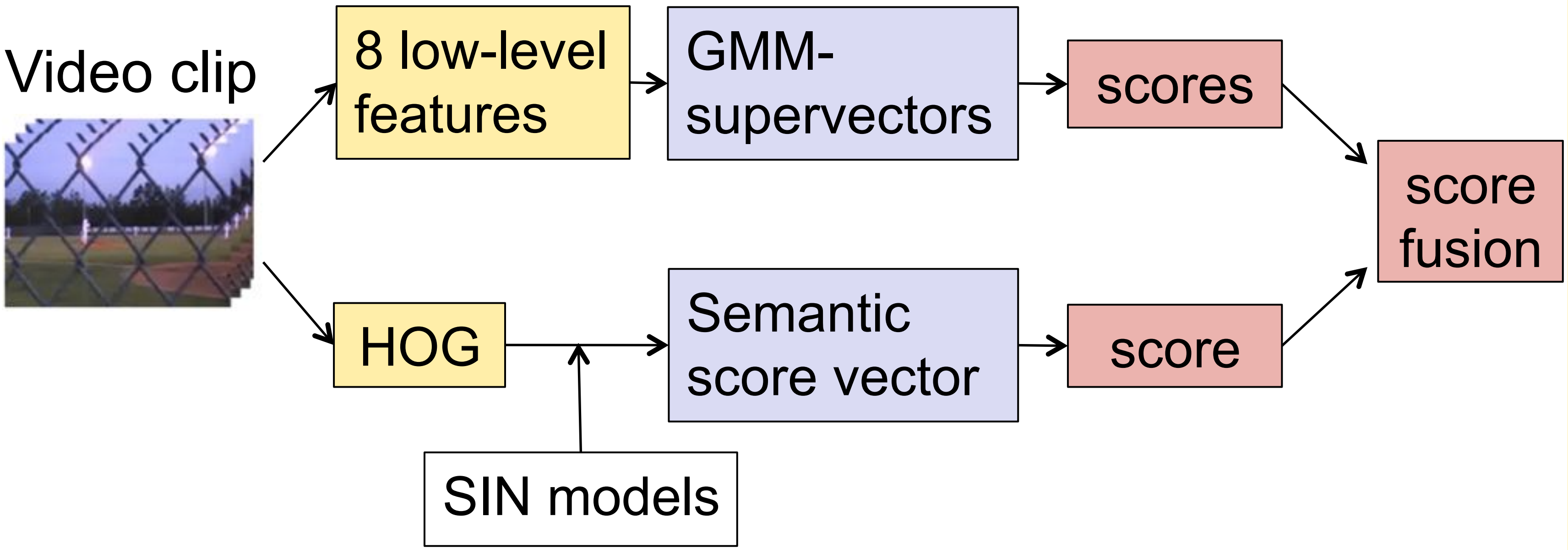
- System Overview
- Detection Method
  - **Camera motion cancellation for STIP features**

    + 7 low-level features (Motion, Appearance, Audio)
  - Gaussian mixture model (GMM) supervectors

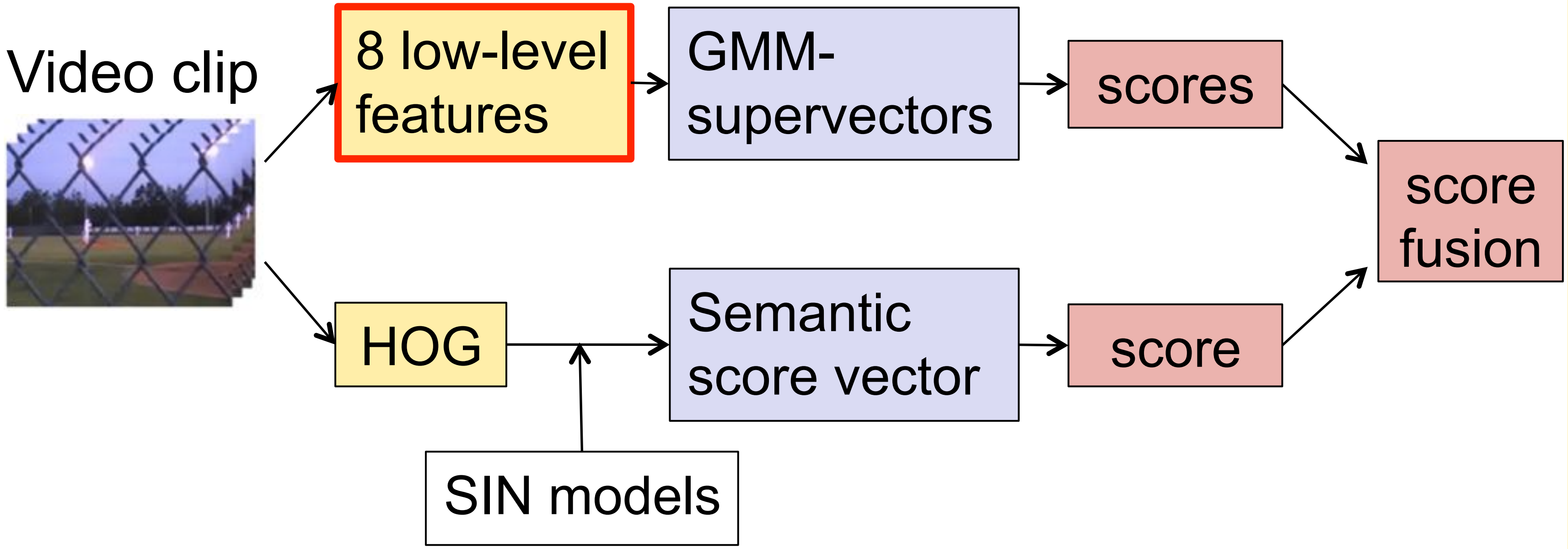
    + Spatial pyramids + SVM
  - Semantic score vector: 346 concepts from SIN task
- Experimental results
- Conclusion

Method	MANDC
Ours in MED 11	0.550
+ 3 feature types	**0.530**
+ semantic score	0.533

System Overview



System Overview



Low-Level Features

- Motion features

1) **Camera-motion-cancelled dense STIP (CC-DSTIP)**

2*) STIP

- Appearance features

3*) SIFT-Har, 4*) SIFT-Hes, 5) SURF,

6*) HOG, 7) RGB-SIFT,

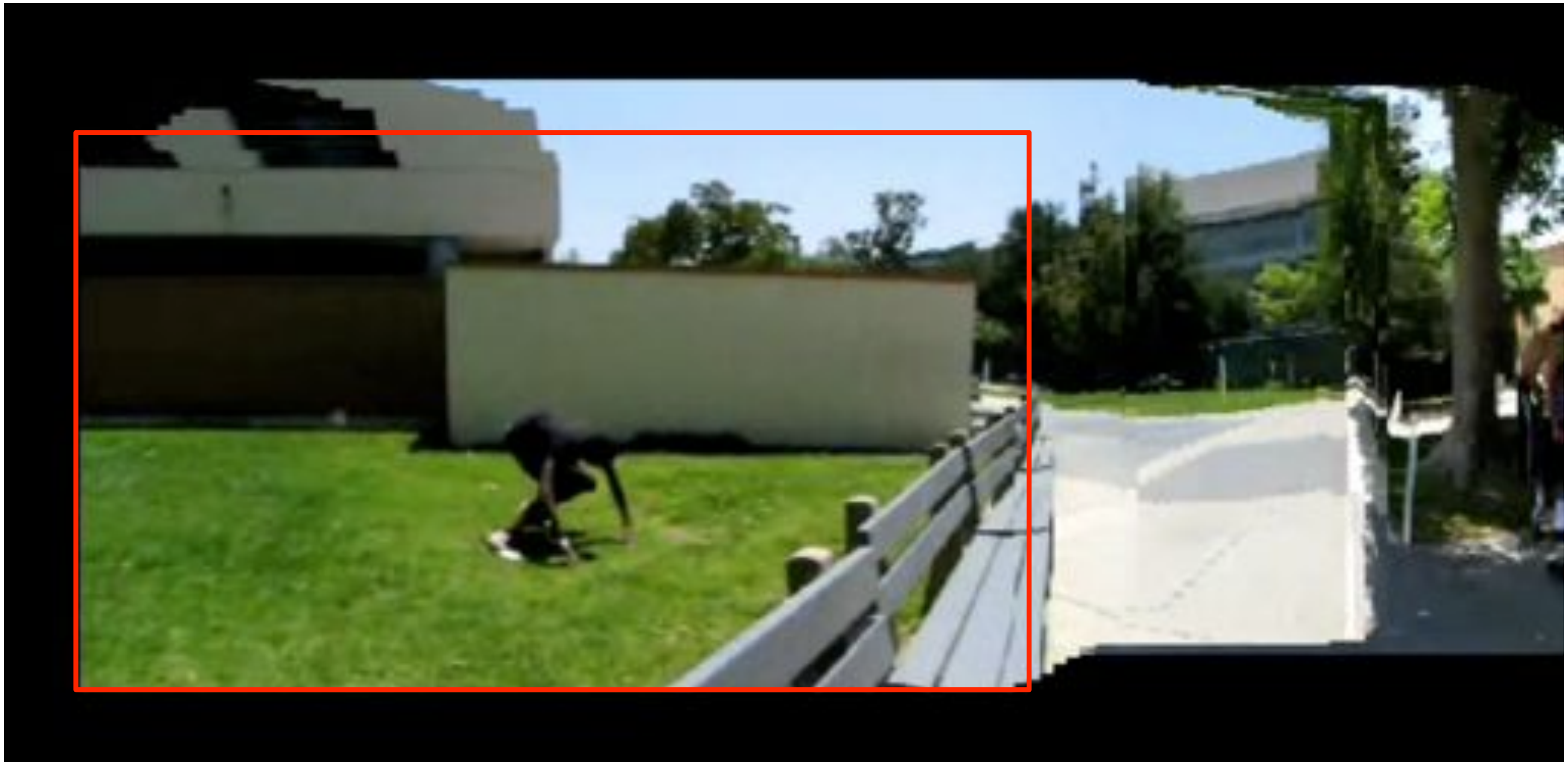
- Audio features

8*) MFCC

*: 5 features used in our MED 11 method

Camera-Motion Cancellation

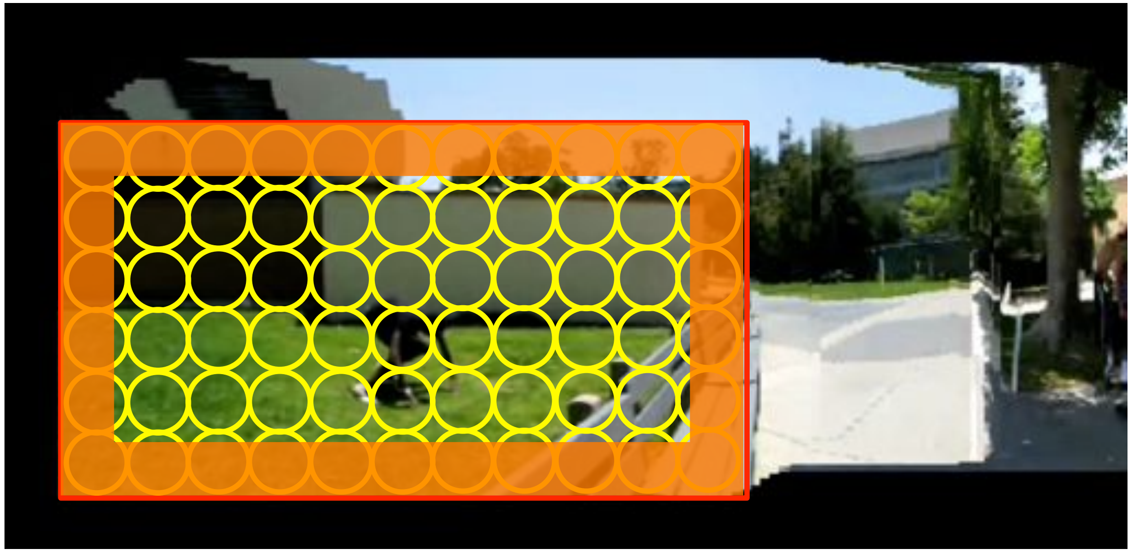
- Separate camera motion and object motion



Example (Video)

CC-DSTIP

- Camera-motion-cancelled dense (CC-D) STIP
  1. **Estimate the camera motion** by using optical flows in the peripheral region.
  2. **Remove the camera motion** by shifting a frame to the same direction as the optical flows.
  3. **Extract dense STIP** features



STIP+CC-DSTIP

- Experimental results on MED 11

Feature	Mean MNDC
STIP	0.677
DSTIP	0.706
CC-DSTIP	0.694
STIP+CC-DSTIP	0.635

- STIP: original STIP*
- DSTIP: dense STIP
- CC-DSTIP: camera-motion-canceled dense SITP

* Space-time interest points by Harris 3D detector

162-dimensional features (HOG+HOF) are computed in STIP.

Appearance Features (Sparse)

- SIFT with Harris-Affine detector (**SIFT-Har**)
  - 128-dimensional features robust for illumination and scale change.
  - Harris-Affine detector : used for corner detection
- SIFT with Hessian-Affine detector (**SIFT-Hes**)
  - Hessian-Affine detector : used for blob detection
- SURF features (**SURF**)
  - 64-dimensional feature extracted using the sum of 2D Haar wavelet responses.

<u>They are extracted from 1 frame in every 2 seconds.</u>

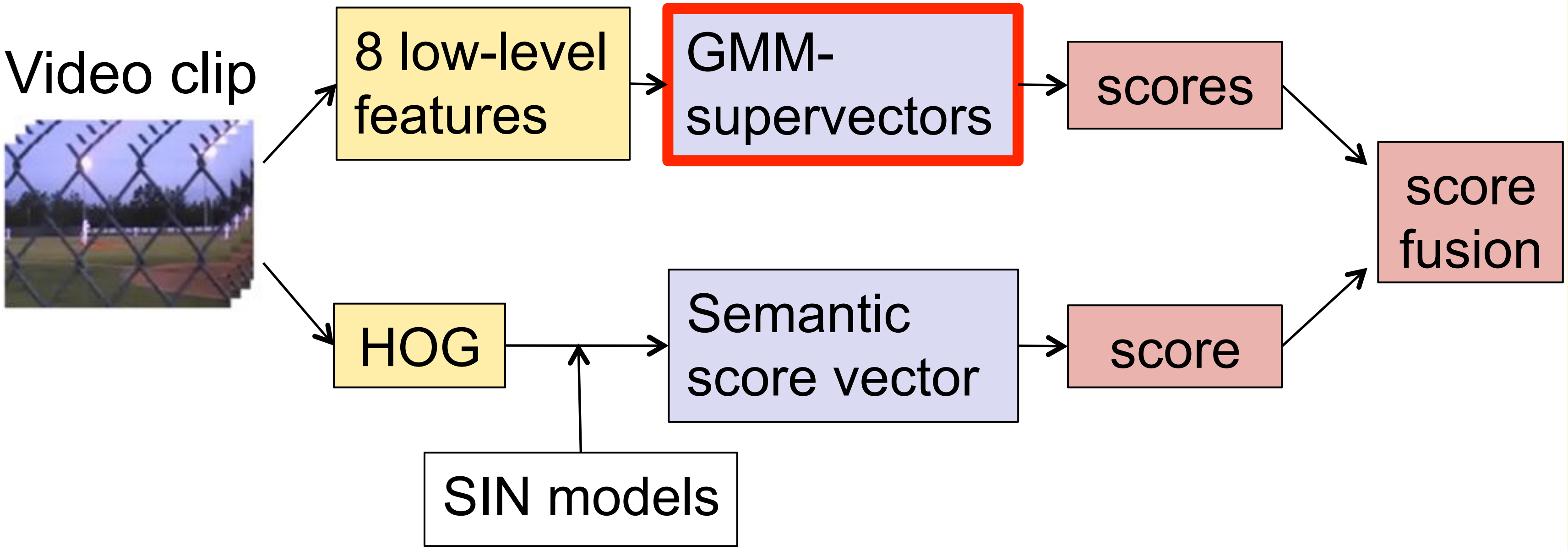
Appearance Features (Dense)

- HOG features with dense sampling (**HOG**)
  - Histograms of oriented gradients extracted densely in a image.
  - 7,200 features are sampled in 1 frame image in every 2 seconds
- RGB-SIFT features with dense sampling (**RGB-SIFT**)
  - 384-dimensional color features with dense sampling
  - Sampled from every 6 pixels, and 1 frame in every 6 seconds

Audio Features

- MFCC features (**MFCC**)
  - Audio features often used in speech recognition
  - In addition to MFCC, ΔMFCC + ΔΔMFCC + Δpower + ΔΔpower are also used. → Total dimensions are 38.

System Overview

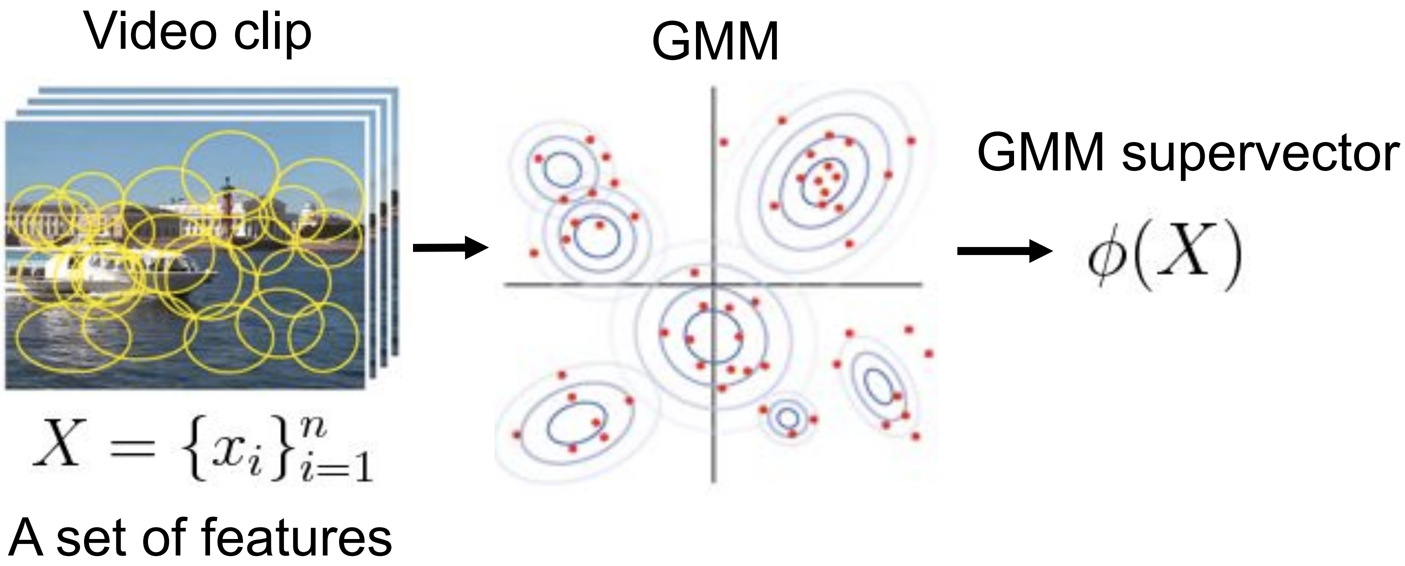


Gaussian mixture model (GMM)

- Each video clip is represented by a GMM

- Estimate GMM parameters

- GMM supervector: concatenation of the parameters



GMM Parameter Estimation

- Maximum a posteriori (MAP) adaptation

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

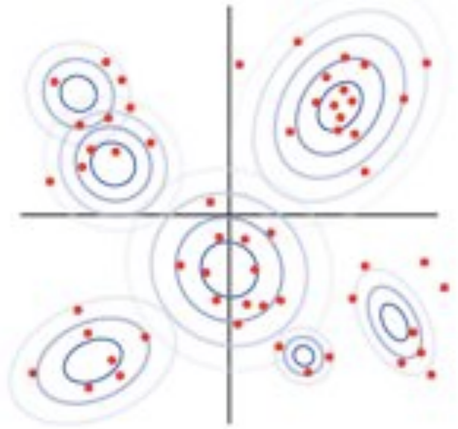
where

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

$X = \{x_i\}_{i=1}^{n}$

*UBM

MAP
adaptation



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

GMM Supervector

- Concatenate mean vectors of a GMM

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

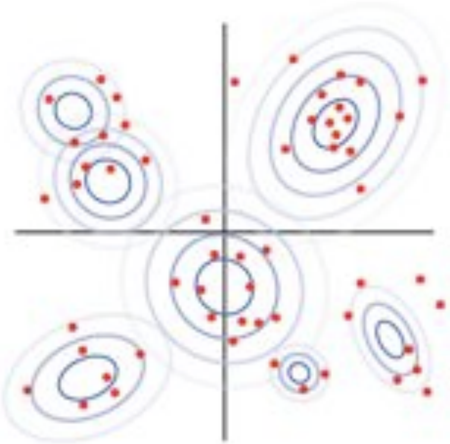
where

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

$X = \{x_i\}_{i=1}^{n}$

UBM

MAP adaptation

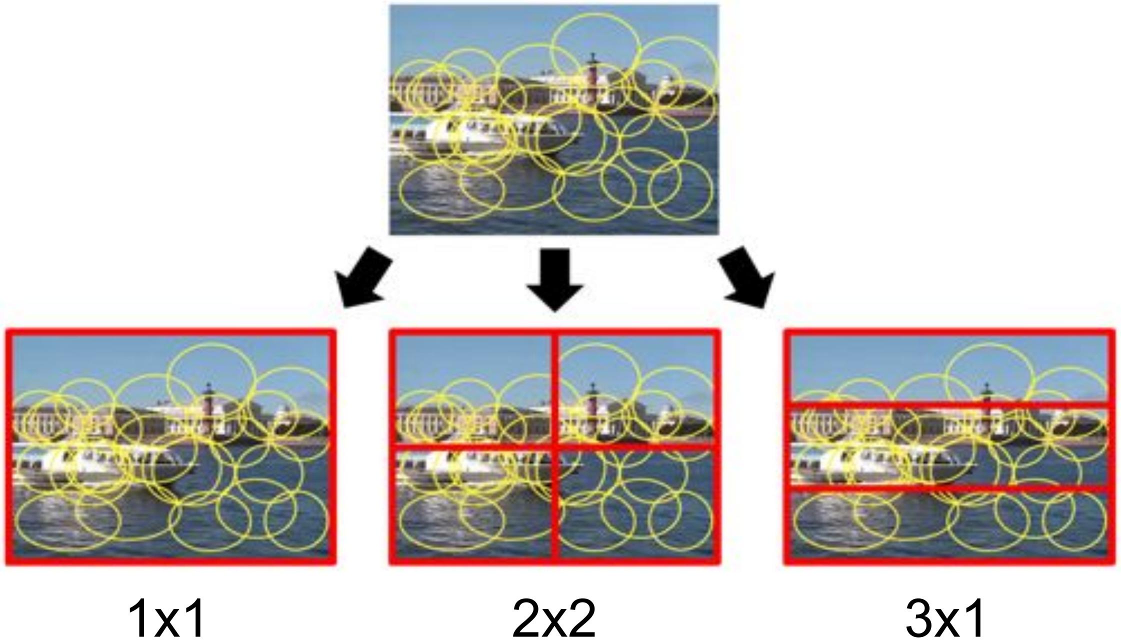


$\phi(X)$

GMM supervector

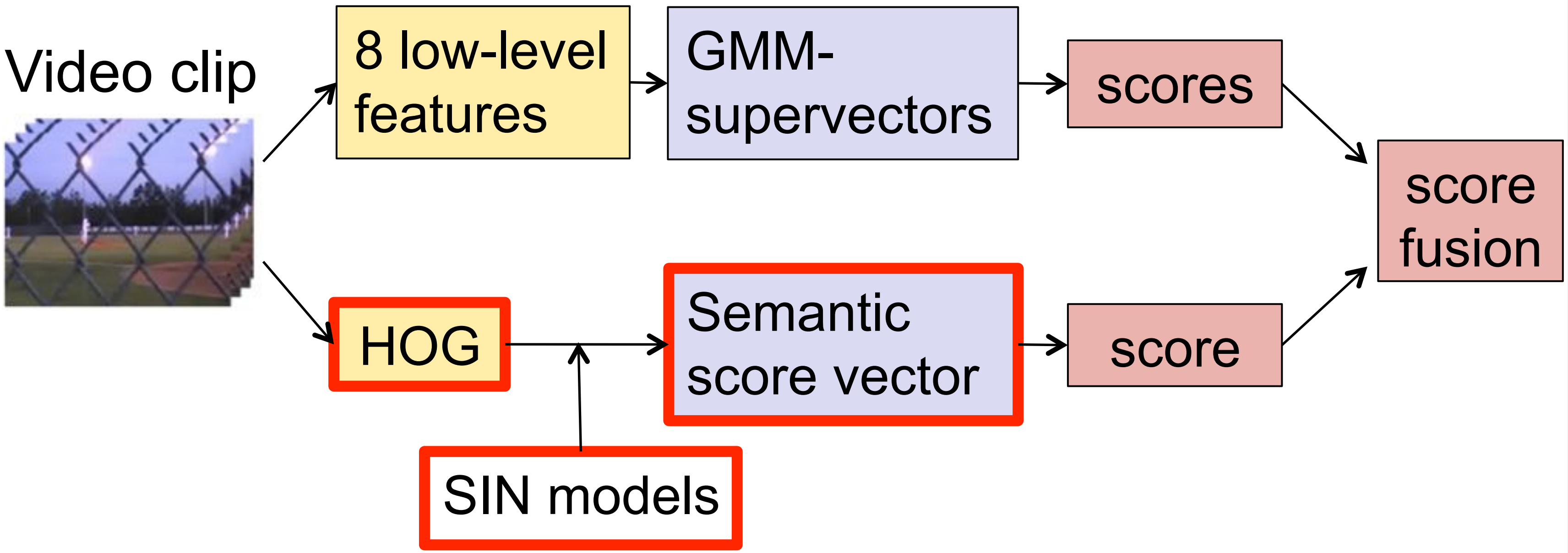
Spatial Pyramids

- Use spatial information of low-level features
  1. Extract GMM supervectors for each 8 regions
  2. Concatenate 8 GMM supervectors into a vector.



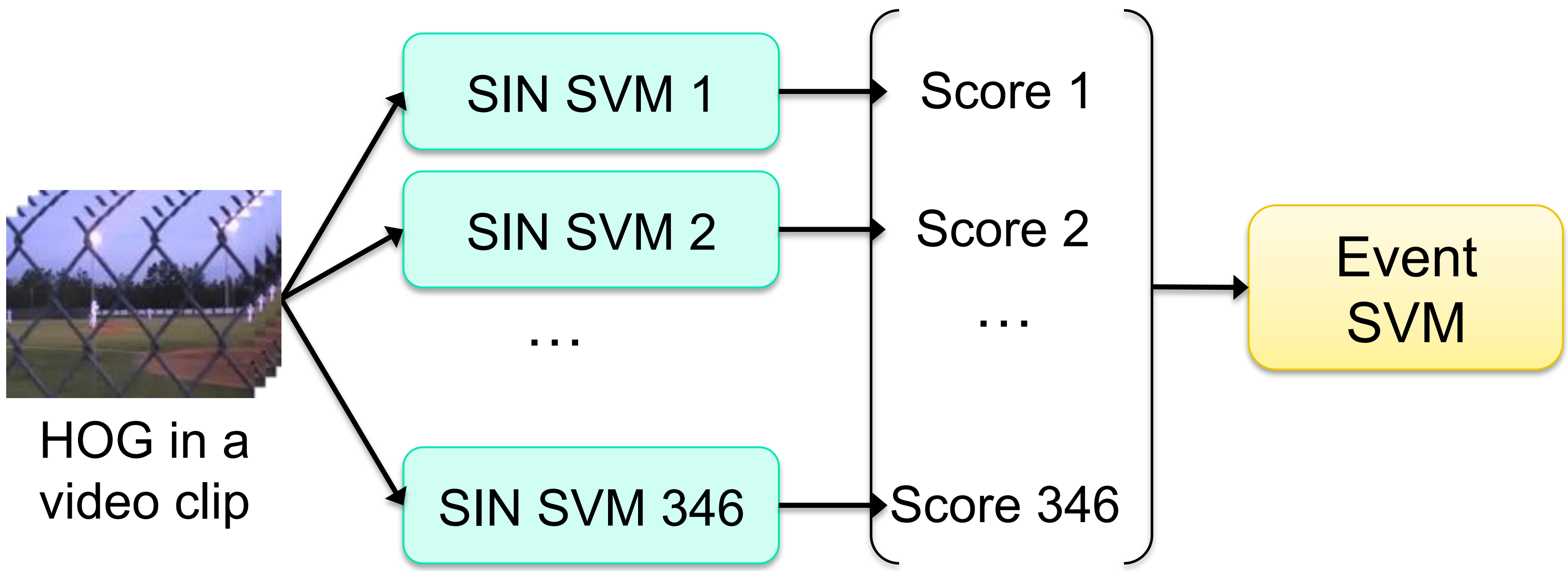
- For SIFT-Har, SIFT-Hes, HOG, SURF, and RGB-SIFT

System Overview



Semantic Score Vector

- Use semantic concept models in SIN task
  - A semantic score vector consists of the SVM scores for the 346 concepts in SIN task
  - Use it as input to an SVM for each event



Test SIN Models on MED

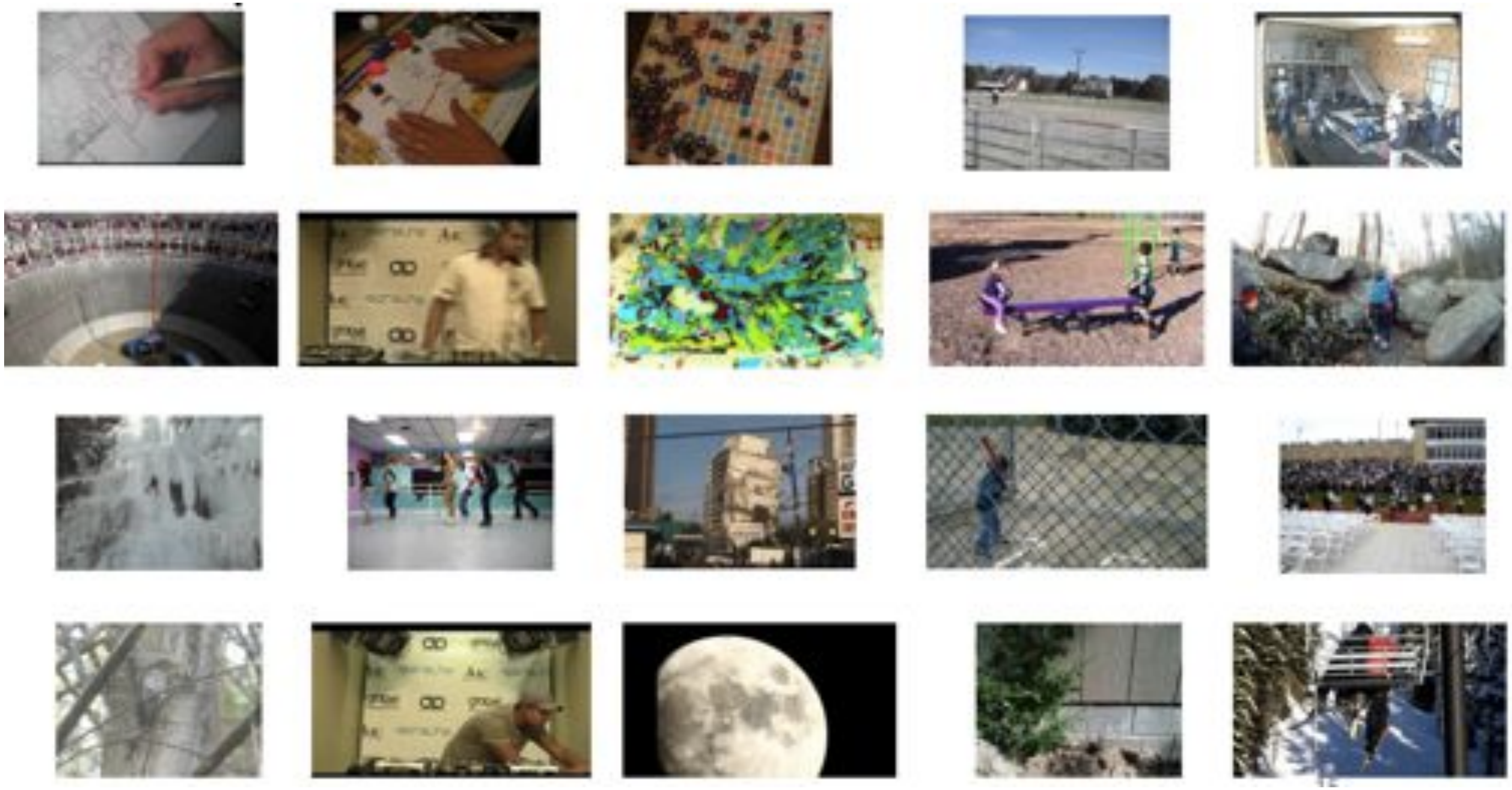
- **Car (Top 20)**

Test SIN Models on MED

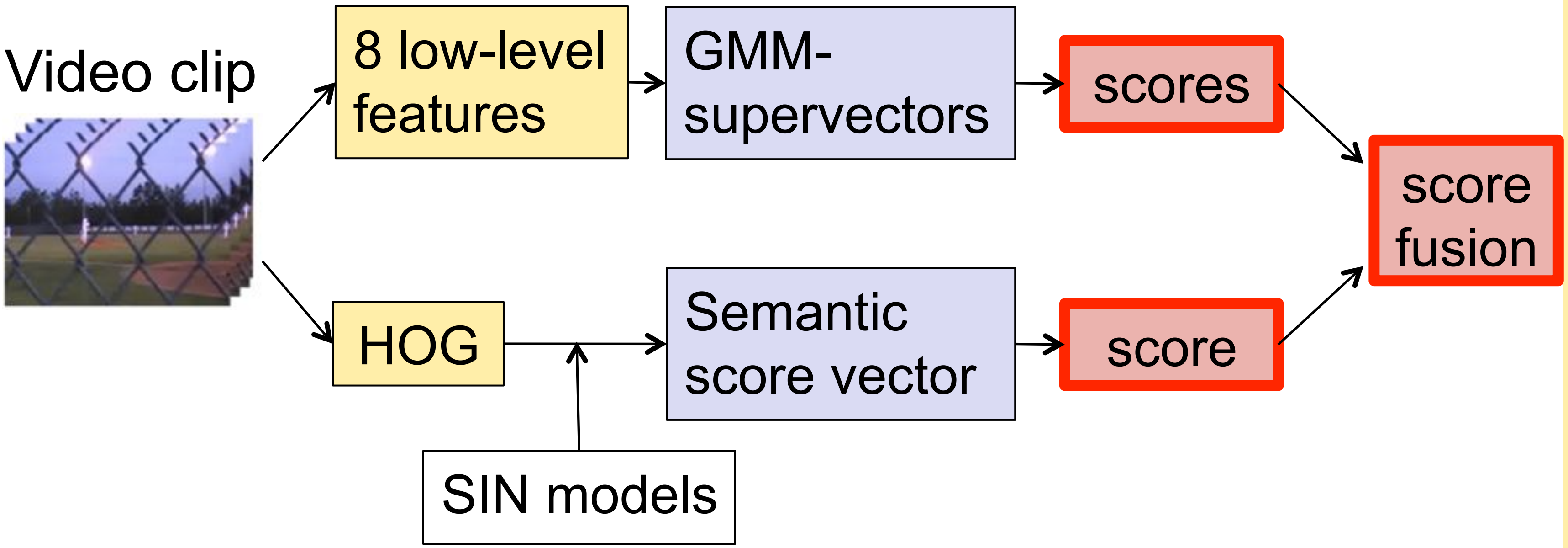
- Dogs (Top 20)

Test SIN Models on MED

- **Map (Top 20)**



System Overview



Fusion of SVM Scores

- One-vs-all SVM
  - for each event and for each feature type with RBF-kernels.

$$k(X_i, X_j) = \exp(-\gamma \|\phi(X_i), \phi(X_j)\|_2^2)$$

- Detection score

$$s(X) = \sum_{\mathrm{F}} \alpha_{\mathrm{F}} f_{\mathrm{F}}(X)$$

where

f_{F} : detection score for feature type F
α_{F} : Fusion weight for feature type F

Results

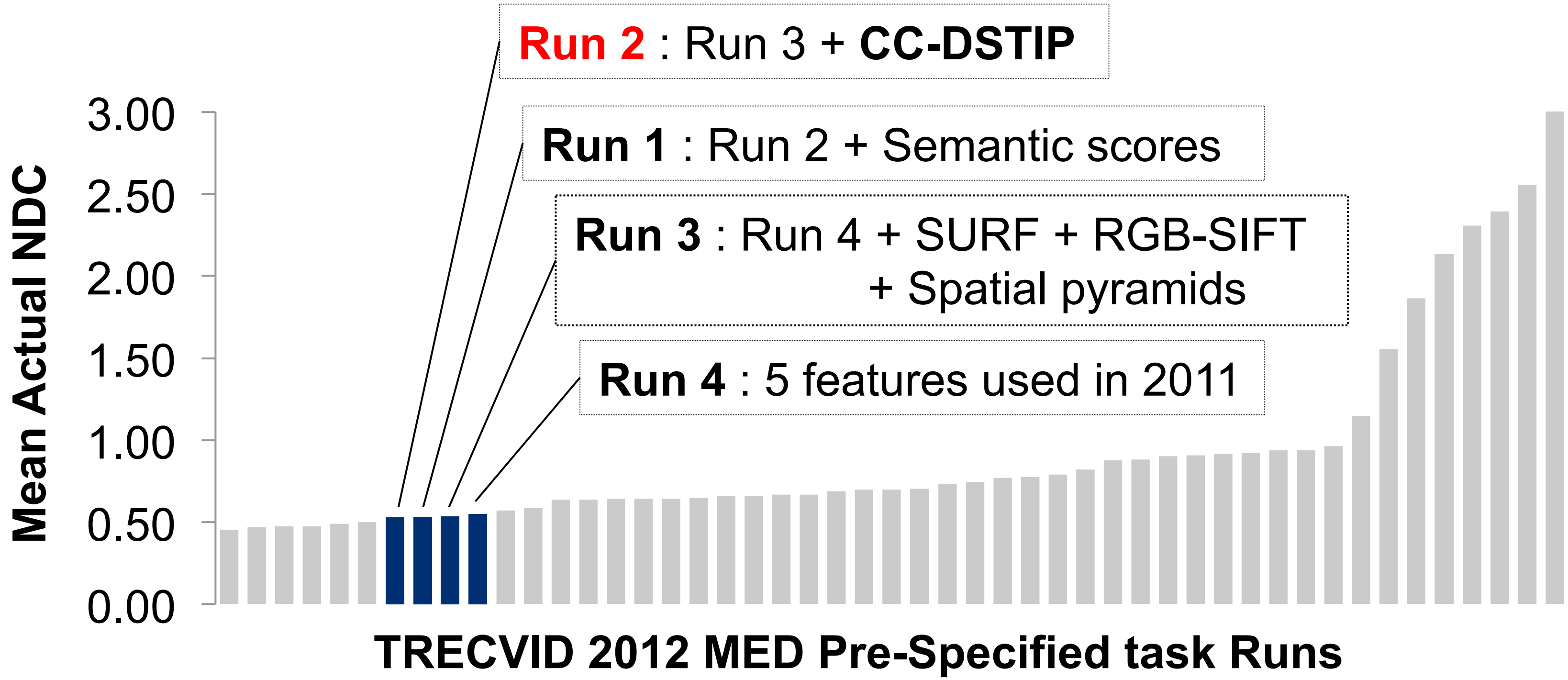
Pre-Specified Task

Run ID	System ID	Features	Mean ANDC
Run 1	p-GSSVM7PyramidCcScv-r1	Run 2 + Sematic	0.533
Run 2	c-GSSVM7PyramidCc-r2	Run 3 + CC-DSTIP	**0.530**
Run 3	c-GSSVM7Pyramid-r3	Run 4 + RGBSIFT, SURF + spatial pyramids	0.534
Run 4	c-GSSVM5-r4	5 types in MED11	0.550

- Detection thresholds and the fusion weights are optimized by using 2-fold cross validation.

Performance Comparison

- Ranked 7th /49 runs and 3rd /17 teams
  (among the "EKFull" runs)



Ad-Hoc Task

Run ID	System ID	Features	Mean ANDC
Run 5	p-GSSVM7PyramidCcScv-r5_1	The same 9 types as Run 1	1.7490
Run 6	c-GSSVM5-r6_1	5 types in MED11	2.5351

- As the detection thresholds, we used the average of those of Pre-Specified events.
- The fusion weights were determined by the same way.

➢ <u>These unexpected results are due to a bug of our script.</u>

Conclusion

- Camera motion cancellation for STIP
  - Provided **complementary information** to other features and was **more effective than feature without cancellation**.
- GMM supervectors with 8 low-level features
  - Our best mean Actual NDC was **0.5296** ranked **3rd among the 17 teams** in MED12 Pre-Specified task.
- Future works
  - more on using the SIN models for the MED task
  - improve the fusion method of multiple features