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

<table>
<thead>
<tr>
<th>Method</th>
<th>MANDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours in MED 11</td>
<td>0.550</td>
</tr>
<tr>
<td>+ 3 feature types</td>
<td>0.530</td>
</tr>
<tr>
<td>+ semantic score</td>
<td>0.533</td>
</tr>
</tbody>
</table>
System Overview

Video clip → 8 low-level features → GMM-supervectors → scores

HOG → Semantic score vector → score

SIN models

score fusion
System Overview

Video clip → 8 low-level features → GMM-supervectors → scores

SIN models → HOG → Semantic score vector → score

score fusion
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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean MNDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIP</td>
<td>0.677</td>
</tr>
<tr>
<td>DSTIP</td>
<td>0.706</td>
</tr>
<tr>
<td>CC-DSTIP</td>
<td>0.694</td>
</tr>
<tr>
<td>STIP+CC-DSTIP</td>
<td>0.635</td>
</tr>
</tbody>
</table>

- 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.

They are extracted from 1 frame in every 2 seconds.
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

Video clip → 8 low-level features → GMM-supervectors → scores

HOG → Semantic score vector → score

SIN models

score fusion
Gaussian mixture model (GMM)

- Each video clip is represented by a GMM
- Estimate GMM parameters
- GMM supervector: concatenation of the parameters

\[ X = \{x_i\}_{i=1}^n \]

A set of features
GMM Parameter Estimation

- Maximum a posteriori (MAP) adaptation

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

Where

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

*UBM

MAP adaptation

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

- Concatenate mean vectors of a GMM

\[
\phi(X) = \left( \begin{array}{c}
\tilde{\mu}_1 \\
\tilde{\mu}_2 \\
\vdots \\
\tilde{\mu}_K 
\end{array} \right)
\]

where

\[
\tilde{\mu}_k = \sqrt{\frac{w_k^{(U)}}{(\Sigma_k^{(U)})^{-1/2}}} \mu_k
\]

Normalized   Mean

UBM

MAP adaptation

\[X = \{x_i\}_{i=1}^n\]

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

Video clip

8 low-level features → GMM-supervectors → scores

HOG

Semantic score vector → score

SIN models

score fusion
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

HOG in a video clip

SIN SVM 1 → Score 1
SIN SVM 2 → Score 2
SIN SVM 346 → Score 346

Event SVM
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

Video clip

8 low-level features

GMM-supervectors

scores

score fusion

HOG

Semantic score vector

score

SIN models
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

\[ s(X) = \sum_{F} \alpha_F f_F(X) \]

where

- \( f_F \) : detection score for feature type \( F \)
- \( \alpha_F \) : Fusion weight for feature type \( F \)
Results
Pre-Specified Task

<table>
<thead>
<tr>
<th>Run ID</th>
<th>System ID</th>
<th>Features</th>
<th>Mean ANDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 1</td>
<td>p-GSSVM7PyramidCcScv-r1</td>
<td>Run 2 + Semantic</td>
<td>0.533</td>
</tr>
<tr>
<td>Run 2</td>
<td>c-GSSVM7PyramidCc-r2</td>
<td>Run 3 + CC-DSTIP</td>
<td>0.530</td>
</tr>
<tr>
<td>Run 3</td>
<td>c-GSSVM7Pyramid-r3</td>
<td>Run 4 + RGBSIFT, SURF + spatial pyramids</td>
<td>0.534</td>
</tr>
<tr>
<td>Run 4</td>
<td>c-GSSVM5-r4</td>
<td>5 types in MED11</td>
<td>0.550</td>
</tr>
</tbody>
</table>

- 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)

Run 1: Run 2 + Semantic scores
Run 2: Run 3 + CC-DSTIP
Run 3: Run 4 + SURF + RGB-SIFT + Spatial pyramids
Run 4: 5 features used in 2011
Ad-Hoc Task

<table>
<thead>
<tr>
<th>Run ID</th>
<th>System ID</th>
<th>Features</th>
<th>Mean ANDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 5</td>
<td>p-GSSVM7PyramidCcScv-r5_1</td>
<td>The same 9 types as Run 1</td>
<td>1.7490</td>
</tr>
<tr>
<td>Run 6</td>
<td>c-GSSVM5-r6_1</td>
<td>5 types in MED11</td>
<td>2.5351</td>
</tr>
</tbody>
</table>

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

▶ These unexpected results are due to a bug of our script.
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