Semantic Indexing Using GMM Supervectors and Video-Clip Scores

Nakamasa Inoue, Kotaro Mori, and Koichi Shinoda,
Department of Computer Science,
Tokyo Institute of Technology
Outline

- System overview
- Baseline system
  - GMM spuervectors for 6 types of low-level features
- Spatial pyramid + Velocity pyramid*
- Re-scoring by video-clip scores
- Best result: Mean InfAP = 28.4%

System Overview

- Extend Bag-of-Words to a probabilistic framework

Velocity pyramid

Re-scoring
System Overview

- **STEP1: low-level feature extraction**

1) Har-SIFT
2) Hes-SIFT
3) Dense-HOG
4) Dense-LBP
5) Dense-SIFTH
6) MFCC
Low-Level Features (Visual)

1) Har-SIFT
   - Harris-affine detector [Mikolajczyk, 2004]
   - Multi-frame (every other frame)
2) Hes-SIFT
   - Hessian-affine detector
   - Multi-frame (every other frame)
3) Dense HOG
   - 32 dimensional HOG, 10,000 samples per frame
   - up to 100 frames per shot
4) Dense LBP
   - Local binary pattern, 10,000 samples per frame
   - up to 100 frames per shot
5) Dense SIFTH
   - SIFT + Hue histogram
   - 30,000 samples from a key-frame
Low-Level Features (Audio)

6) MFCC

- Mel-frequency cepstrum coefficients (MFCC)
- Audio features for speech recognition
- Targets: Speaking, Singing etc.
System Overview

- STEP2: GMM supervector extraction

Estimate GMM parameters
- Tree-structured GMM
- MAP adaptation

Extract GMM supervector
Spatial + Velocity pyramid
Gaussian Mixture Models (GMMs)

- Each shot is model by a GMM
  
  \[ X_F = \{ x_i \}_{i=1}^n \] : local features
  
  \[ \theta = \{ w_k, \mu_k, \Sigma_k \}_{k=1}^K \] : GMM parameters

- GMM parameters are estimated by using maximum a posteriori (MAP) adaptation

Universal background model (UBM): a prior GMM which is estimated by using all video data.
Gaussian Mixture Models (GMMs)

- MAP adaptation for mean vectors:
  \[
  \hat{\mu}_k = \frac{\tau \hat{\mu}_k^{(U)} + \sum_{i=1}^n c_{ik} x_i}{\tau + C_k}
  \]
  where
  \[
  c_{ik} = \frac{w_k N(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^K w_k N(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}, \quad C_k = \sum_{i=1}^{n_s} c_{ik}
  \]

  responsibility of component \(k\) for \(x_i\)

  Computational cost: high

GMM Supervector

- Combine normalized mean vectors.

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

where

$$\tilde{\mu}_k = \sqrt{\omega_k(U) \Sigma_k(U)} \frac{1}{2} \mu_k$$

normalized mean
Velocity Pyramid

- Extend spatial pyramid to motion
  - extract optical flow, quantize velocity vectors
  - concatenate GMM supervectors

Velocity Pyramid

$L = 0$

$L = 1$
- Vertical
- No-motion
- Horizontal

$L = 2$
- Down
- Up
- No-motion
- Right
- Left
System Overview

- STEP3: compute shot scores

Video clip → shot → Low-level features → GMM supervector → Shot score → Video-clip score → New shot score
Shot Scores

- Linear combination of SVM scores

\[ s = \sum_{F \in \mathcal{F}} \alpha_F f_F(X_F), \quad 0 \leq \alpha_F \leq 1, \quad \sum_{F} \alpha_F = 1 \]

where

\[ \mathcal{F} = \{ \text{SIFT-Har, SIFT-Hes, SIFTH-Dense, HOG-Dense, LBP-Dense, MFCC} \} \]

\( \alpha_F \) : optimized for each semantic concept (on IACC_1_B)

\[ S_1 \quad S_2 \quad S_3 \quad S_4 \quad S_5 \]
Video-Clip Score

- A semantic concept often reappears in a video clip
- Problem: occlusion, closed-up etc.
Video-Clip Score

- Video-clip score: the maximum shot score in a clip
- Re-scoring:

$$s'_i = (1 - p)s_i + ps_{\text{max}}$$

$$p = r \left( \frac{\text{#(positive shots in a video clip)}}{\text{#(shots in a video clip)}} \right)$$

Video-clip score

Shot scores:
- $s_1$
- $s_2$
- $s_3$
- $s_4_{\text{max}}$
- $s_5$

Re-scoring:
- $s'_1$
- $s'_2$
- $s'_3$
- $s'_4$
- $s'_5$
Experimental Condition

- **TokyoTech_Canon_4**
  - 6 types of GMM supervectors
  - Video-clip score (r=1.0)
- **TokyoTech_Canon_3**
  - + Spatial and velocity pyramid for HOG
- **TokyoTech_Canon_2**
  - set r=0.9 for video-clip scores
- **TokyoTech_Canon_1**
  - set r=0.8 for video-clip scores
## Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Method</th>
<th>Mean InfAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TokyoTech_Canon_4</td>
<td>6 types of GMM sv + video-clip scores</td>
<td>0.280</td>
</tr>
<tr>
<td>TokyoTech_Canon_3</td>
<td>+ Spatial and velocity pyramid</td>
<td>0.283</td>
</tr>
<tr>
<td>TokyoTech_Canon_2</td>
<td>set r = 0.9</td>
<td>0.284</td>
</tr>
<tr>
<td>TokyoTech_Canon_1</td>
<td>set r = 0.8</td>
<td>0.284</td>
</tr>
</tbody>
</table>
InfAP by Semantic Concepts
Evaluation of Velocity Pyramid

- Mean NDC on the MED task (HOG features)

<table>
<thead>
<tr>
<th></th>
<th>MED 10</th>
<th>MED 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pyramid</td>
<td>0.661</td>
<td>0.688</td>
</tr>
<tr>
<td>Spatial pyramid (SP)</td>
<td>0.635</td>
<td>0.617</td>
</tr>
<tr>
<td>Velocity pyramid (VP)</td>
<td>0.617</td>
<td>0.620</td>
</tr>
<tr>
<td>SP+VP</td>
<td>0.607</td>
<td>0.600</td>
</tr>
</tbody>
</table>

- Mean AP on the SIN task

<table>
<thead>
<tr>
<th></th>
<th>SIN 12 (HOG)</th>
<th>SIN 12 (Fusion)</th>
<th>SIN 13 (Fusion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pyramid</td>
<td>0.236</td>
<td>0.321</td>
<td>0.280</td>
</tr>
<tr>
<td>SV+VP</td>
<td>0.245</td>
<td>0.323</td>
<td>0.283</td>
</tr>
</tbody>
</table>

* Fusion: fusion of 6 types of visual and audio features, but SV+VP is applied to only HOG
Evaluation of Video-clip Scores

- Mean AP on SIN 2012

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Video-Clip Score</th>
<th>Video-Clip Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Har-SIFT</td>
<td>0.183</td>
<td>0.208</td>
</tr>
<tr>
<td>Hes-SIFT</td>
<td>0.179</td>
<td>0.207</td>
</tr>
<tr>
<td>Dense-SIFTH</td>
<td>0.202</td>
<td>0.224</td>
</tr>
<tr>
<td>Dense-HOG</td>
<td>0.236</td>
<td>0.259</td>
</tr>
<tr>
<td>Dense-LBP</td>
<td>0.235</td>
<td>0.260</td>
</tr>
<tr>
<td>MFCC</td>
<td>0.079</td>
<td>0.086</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.306</td>
<td>0.321</td>
</tr>
<tr>
<td>Fusion (r=0.9)</td>
<td>0.306</td>
<td>0.324</td>
</tr>
</tbody>
</table>
Conclusion

- 6 types of audio and visual GMM supervectors
  + Velocity pyramid
  + Re-scoring by video-clip scores

- Experimental Results
  - Mean InfAP: 0.284

- Future work
  Improve audio analysis
  Audio-visual localization