Semantic Indexing Using Deep CNNs and GMM Supervectors

Nakamasa Inoue and Koichi Shinoda
Tokyo Institute of Technology

Zhang Xuefeng and Kazuya Ueki
Waseda University
Outline

- Part 1: Our system at TRECVID 2014
  - Deep CNNs + GMM supervectors
  - n-gram models for re-scoring
  Best result: Mean InfAP = 0.281

- Part 2: Motion features & Future work
System Overview

- Deep CNN + GMM Supervectors
Deep CNN

- A 4096 dimensional feature vector at the sixth layer is extracted
- A pre-trained model on ImageNET 2012 [1]

GMM Supervectors

- Extend BoW to a probabilistic framework
1) Extract 6 types of visual/audio features: Har-SIFT, Hes-SIFT, Dense HOG, Dense LBP, Dense SIFTH, and MFCC
2) Estimate GMM parameters for each shot
3) Combine normalized mean vectors

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

GMM supervector
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 \( F \) is a feature type, \( \alpha_F \) is a weight.

Shot 1  Shot 2  Shot 3  Shot 4  Shot 5

\[ s_1 \quad s_2 \quad s_3 \quad s_4 \quad s_5 \]
n-Gram Models

- n-consecutive video shots are dependent
- Bigram (n=2)

Re-scoring by

\[ p(w_i = +1 | s_{i-1}, s_i) \]

Label (+1 or -1)

A Full-Gram Model

- n-consecutive video shots are dependent
- Full-gram
  - we simply add the maximum shot score in a video clip

\[ s'_i = (1 - p)s_i + ps_{\text{max}} \quad p = r \left\langle \frac{\#(\text{positive shots in a video clip})}{\#(\text{shots in a video clip})} \right\rangle \]
# Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Method</th>
<th>Mean InfAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TokyoTech-Waseda_4</td>
<td>baseline: GMM Supervectors + Full-gram re-scoring</td>
<td>0.260</td>
</tr>
<tr>
<td>TokyoTech-Waseda_3</td>
<td>+ sampling</td>
<td>0.262</td>
</tr>
<tr>
<td>TokyoTech-Waseda_2</td>
<td>+ Deep CNN</td>
<td>0.280</td>
</tr>
<tr>
<td>TokyoTech-Waseda_1</td>
<td>+ Deep CNN (optimized weight)</td>
<td><strong>0.281</strong></td>
</tr>
</tbody>
</table>

![Graph showing Mean InfAP for TRECVID 2014 Semantic Indexing Runs](image)
InfAP by Semantic Concepts
# Evaluation of n-Gram Models

- Mean AP on SIN 2012

<table>
<thead>
<tr>
<th>Method</th>
<th>MeanAP SIN 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.306</td>
</tr>
<tr>
<td>Bi-gram(n=2)</td>
<td>0.312</td>
</tr>
<tr>
<td>Tri-gram(n=3)</td>
<td>0.312</td>
</tr>
<tr>
<td>Full-gram</td>
<td><strong>0.321</strong></td>
</tr>
</tbody>
</table>
Conclusion (Part 1)

- Deep CNN + GMM Supervector
- n-gram models for re-scoring
- Experimental Results
  - Mean InfAP: 0.281
- Future work
  - Improving audio analysis
  - Introducing motion features for object tracking with deep CNNs
Motion features

- Our baseline system did not include any motion information
  - 5 visual (Har-SIFT, Hes-SIFT, Dense HOG, Dense LBP, and Dense SIFTH) + 1 audio features

- Tried to introduce Dense trajectories into our system
  - Probably effective for some actions / movements.
    - ex.) “Running”, “Swimming”, “Throwing” and etc.

- But unfortunately, we could not finish before the submission deadline.
Dense trajectories

- 4 types of features were extracted from each shot
  - **Trajectory** (a sequence of displacement vectors)
  - **HOG** (Histogram of Oriented Gradient)
  - **HOF** (Histogram of Optical Flow)
  - **MBH** (Motion Boundary Histogram)
Dense trajectories

- Setting
  - Use every other frames
  - Trajectory length L=15
    → More than 30 frames are needed to extract features,
      but about 40% of shots have less than 30 frames...
  - Volume is subdivided into a spatio-temporal grid of size 2 x 2 x 3
  - Orientations are quantized into 8 (or 9) bins.

L = 15 [frames]

2 x 2

5 [frames]

- HOG: 96 dim → 32 dim
- HOF: 108 dim → 32 dim
- MBH: 108x2 dim → 64 dim

PCA
Dense trajectories

- Trajectory → GMM Supervectors → SVMs → Scores
- Video shot → HOG on trajectories → SVMs → Scores
- HOF on trajectories → SVMs → Scores
- MBH on trajectories → SVMs → Scores
Performance of dense trajectories

Mean AP on SIN 2010

<table>
<thead>
<tr>
<th>Method</th>
<th>MeanAP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (6 features)</td>
<td>14.07</td>
</tr>
<tr>
<td>Trajectory</td>
<td>1.28</td>
</tr>
<tr>
<td>HOG on trajectories</td>
<td>8.30</td>
</tr>
<tr>
<td>HOF on trajectories</td>
<td>4.79</td>
</tr>
<tr>
<td>MBH on trajectories</td>
<td>7.14</td>
</tr>
</tbody>
</table>
Complementarity

Mean AP (%) on SIN 2010

Dense HOG + HOG on trajectories

10.90 + 8.30 = 9.82

Late fusion

- We have not tried the fusion weight optimization, but Dense HOG and HOG on trajectories is not so complementary.
Complementarity

- HOF and MBH are different from other features.
- Finally, we could slightly improve mean AP by combining MBH with our baseline method.

Mean AP (%) on SIN 2010

\[
\begin{array}{ccc}
\text{6 features} & + & \text{MBH} \\
14.07 & 7.14 & 14.29
\end{array}
\]

Late fusion

(*) no fusion weight optimization
Future work

- Adapt velocity pyramid to dense SIFT/HOG/LBP …

- Motion features with deep CNN