TokyoTech at TRECVID 2015

Semantic Indexing Using Deep CNN and GMM Supervectors

Nakamasa Inoue and Koichi Shinoda, Tokyo Institute of Technology

System Overview
- A hybrid system of Gaussian-mixture-model (GMM) supervectors and deep convolutional neural networks.

Convolutional Neural Network
- Features are extracted from multiple frames in each video shot by using convolutional neural networks.

Gaussian-Mixture-Model Supervectors
- Each video shot is modeled by a GMM. Maximum a posteriori adaptation is used to estimate parameters.
- 6 types of low-level features: Harris SIFT, Hessian SIFT, Dense SIFT, Dense HOG, Dense LBP, and MFCC.

Results & Conclusion
- Our best result was 0.299 (Mean InfAP), which is ranked 3rd among participating teams.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean InfAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep CNN</td>
<td>0.274</td>
</tr>
<tr>
<td>GMM Supervisor</td>
<td>0.226</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.299</td>
</tr>
</tbody>
</table>

TokyoTech at TRECVID 2015

Localization with Spatio-Temporal Selective Search and SPPNet
Ryosuke Yamamoto, Nakamasu Inoue, Koichi Shinoda
Tokyo Institute of Technology


- Selective Search produces a large number of object region proposals from a still image
- An image is segmented hierarchically with several segmentation strategies including useless ones


- An efficient method to extract CNN scores from a large number of object regions of an image
- Achieved a state-of-the-art result in object localization
- The Selective Search results are used as region proposal


Novelty 1
Spatio-Temporal Region Proposals

- Selective Search[1] with temporal dimensional extended region proposals
- This will produce a large number of temporally continuous region proposals
- As Selective Search, Results contain a lot of useless proposals

Novelty 2
Multi-Frame Score Fusion

- To avoid noise or object deformation, fuse feature maps among several frames
- Exclude useless proposals

Novelty 3
Neighbor Frame Score Boosting

- If system fails to localize in some frames, Neighbor Frame Score Boosting will recover using neighbors

Our System

Results & Conclusion, Future Works

- Multi-Frame Score Fusion and Neighbor-Frame Score Boosting improved the score
- We archived 3rd place among all teams with harmonic mean of F-scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Harm. Mean of F-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective Search + SPPnet</td>
<td>0.4481</td>
</tr>
<tr>
<td>+ ST-Region Proposals, Multi-Frame Score Fusion</td>
<td>0.4518</td>
</tr>
<tr>
<td>+ Neighbour-Frame Score Boost</td>
<td>0.4569</td>
</tr>
</tbody>
</table>

- Future work: The detection results strongly depend on quality of ST-Region Proposals
- Improve ST-Region Proposals quality
- Localization without region candidates
- Generate regions from feature maps

![Image of results and conclusions]
TokyoTech at TRECVID Multimedia Event Detection 2015
Combination of VideoStory and GMM supervectors
Tran Hai Dang, Nakama Inoue, and Koichi Shinoda
Tokyo Institute of Technology

System overview
We combine VideoStory representation with the GMM supervector system

Feature extraction
4 types of features are extracted:
1. GMM supervectors for dense HOG (DHOG)
2. — for RGB-SIFT (SIFT)
3. — for dense trajectory from HOG, HOF, and MBH (DT)
4. VideoStory representations (New !)
Maximum a posteriori (MAP) adaptation and Universal Background Model (UBM) are used to make GMM supervectors

VideoStory[1]
Pre-training on the VideoStory46K dataset

VideoStory representations of TRECVID videos

Visual projection and textual projection are trained jointly using videos and their titles from the VideoStory46K dataset

Only the visual projection is used to compute VideoStory representations of TRECVID videos

Results
Comparison of our different settings in the condition of PS 10Ex EvalSub

<table>
<thead>
<tr>
<th>Setting</th>
<th>infAP200(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without VideoStory</td>
<td>13.88</td>
</tr>
<tr>
<td>With VideoStory</td>
<td>13.98</td>
</tr>
</tbody>
</table>

Conclusions
- VideoStory shows effectiveness in events such as "Rock climbing", "Fixing musical instruments", "Parking a vehicle", "Tuning musical instruments"
- It is needed to increase the amount of training data for improving the performance