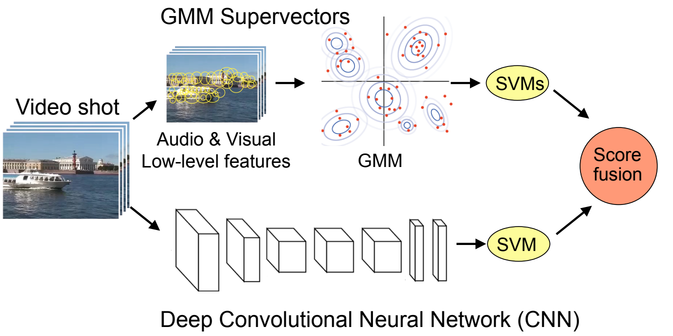
TokyoTech at TRECVID 2015

Semantic Indexing Using Deep CNN and GMM Supervectors

Nakamasa Inoue and Koichi Shinoda, Tokyo Institute of Technology

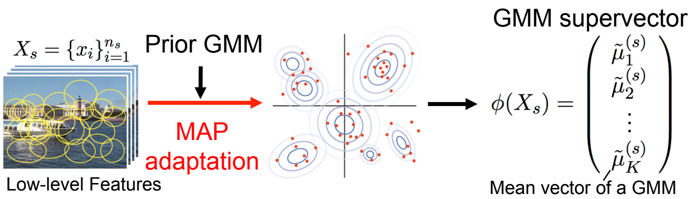
System Overview

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



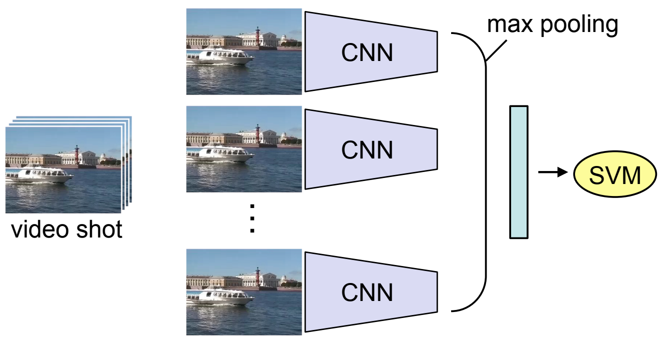
Gaussian-Mixuture-Model Supervectors

- 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.

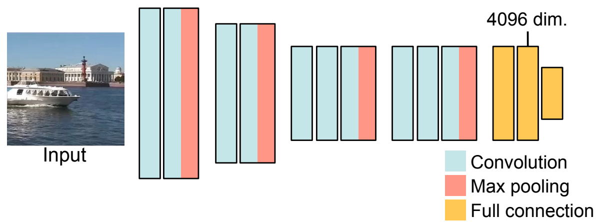


Convolutional Neural Network

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



- The convolutional network with 16 layers in [1] is used to extract 4096 dimensional features.
- Parameters of the CNN are trained on ImageNET 2012.

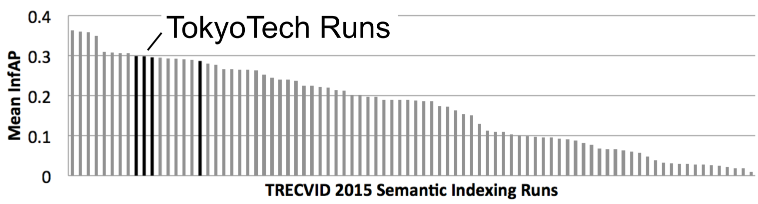


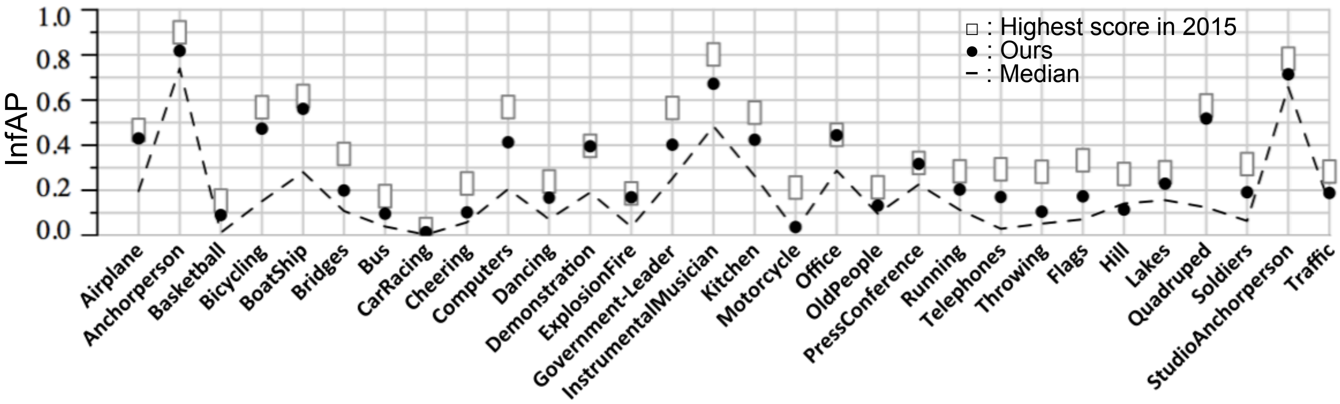
[1] K. Simonyan, and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition In Proc. of ICLR, 2015.

Results & Conclusion

- Our best result was **0.299** (Mean InfAP), which is ranked 3rd among participating teams.
- Future work: audio and motion analysis using deep neural networks.

Method	Mean InfAP
Deep CNN	0.274
GMM Supervector	0.226
Fusion	**0.299**





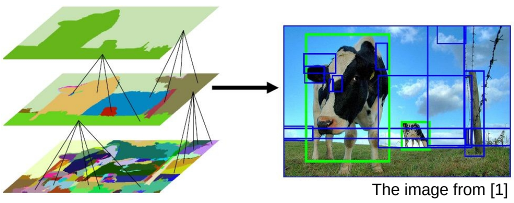
TokyoTech at TRECVID 2015

Localization with Spatio-Temporal Selective Search and SPPnet

Ryosuke Yamamoto, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology

Previous work: Selective Search[1] + Spatial Pyramid Pooling net[2]

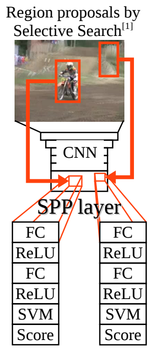
- Selective Search produces a large number of object region proposals from a still image
- An image is segmented hierarchically with several segmentation strategies including **useless ones**



The image from [1]

[1] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders, Selective search for object recognition. In IJCV, vol.104, pp.154-171, 2013

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

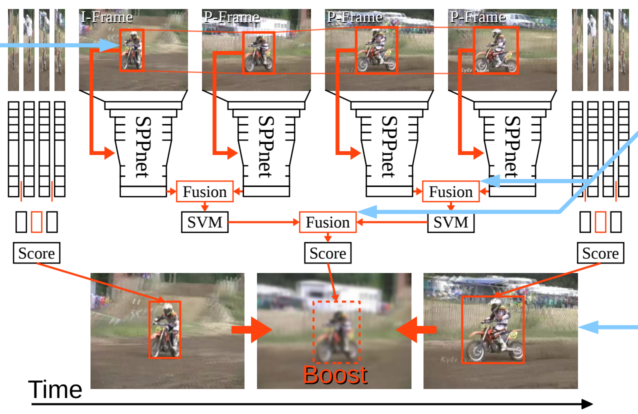


[2] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1904-1916, 2015

Our System

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

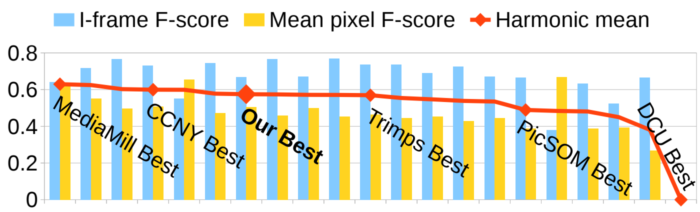
Novelty 3
Neighbor Frame Score Boosting

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

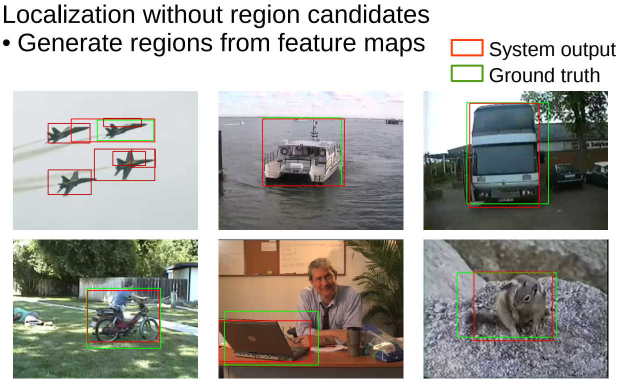
Results & Conclusion, Future Works

- 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

Method	Harm. Mean of F-scores	
	Val	Test
Selective Search + SPPnet	0.4481	0.5656
+ ST-Region Proposals, Multi-Frame Score Fusion	0.4518	0.5716
+ Neighbour-Frame Score Boost	**0.4569**	**0.5750**



- 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



TokyoTech at TRECVID Multimedia Event Detection 2015

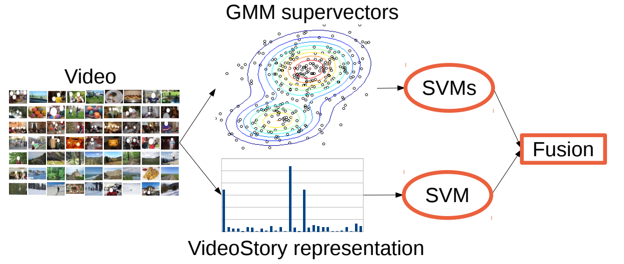
Combination of VideoStory and GMM supervectors

Tran Hai Dang, Nakamasa Inoue, and Koichi Shinoda

Tokyo Institute of Technology

System overview

We combine VideoStory representation with the GMM supervector system



Feature extraction

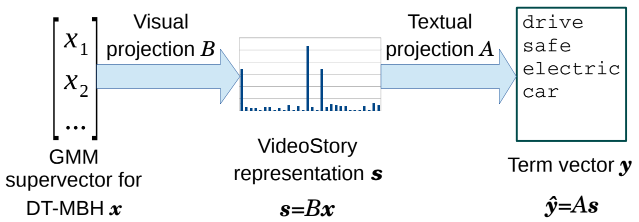
4 types of features are extracted:

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

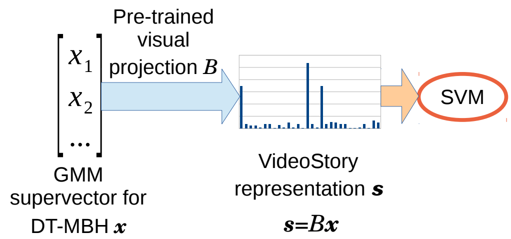
VideoStory[1]

Pre-training on the VideoStory46K dataset



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

VideoStory representations of TRECVID videos



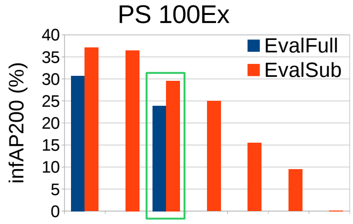
Only the visual projection is used to compute VideoStory representations of TRECVID videos

[1] Amirhossein Habibian, Thomas Mensink, Cees G. M. Snoek. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. ACM Multimedia, 2014

Results

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

Setting	infAP200(%)
Without VideoStory	13.88
With VideoStory	**13.98**



3rd among 7 teams

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