# TokyoTech-Waseda at TRECVID 2014

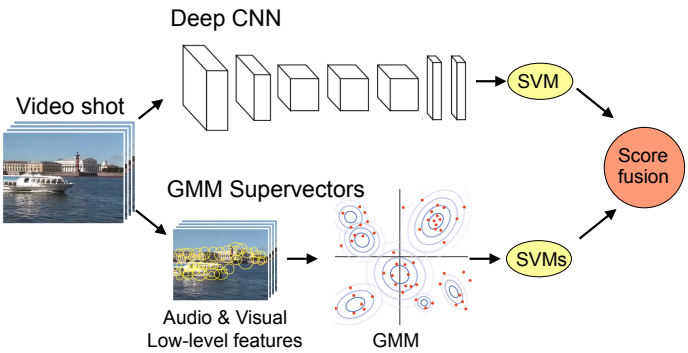
# Semantic Indexing Using Deep CNN and GMM Supervectors


Nakamasa Inoue and Koichi Shinoda
Tokyo Institute of Technology

Zhang Xuefeng and Kazuya Ueki
Waseda University

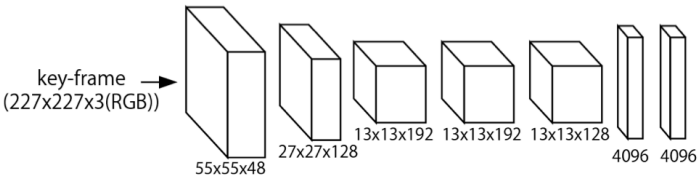

## System Overview

➢ We propose a fast and high-performance semantic indexing system.




## Deep Convolutional Neural Network (CNN)

➢ A 4096 dimensional feature vector at the sixth layer is extracted from the key-frame image of a video shot.
➢ Parameters of the CNN are trained on ImageNET Challenge 2012.




## Local Feature Extraction

➢ 6 types of audio and visual features are extracted from video data.

Multi-frame

**1) Har-SIFT**
SIFT features with Harris-affine detector

**2) Hes-SIFT**
SIFT features with Hessian-affine detector

**3) Dense-HOG**
HOG features with dense sampling

**4) Dense-LBP**
Local binary pattern (LBP) features with dense sampling

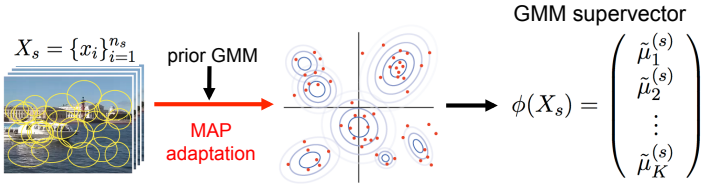
**5) Dense-SIFTH**
SIFT+Hue histogram with dense sampling (key-frame)

**6) MFCC**
Audio features first proposed for speech recognition. Targets: Speaking, Singing, etc.

Audio

## GMM Supervectors

➢ Each video shot is modeled by a Gaussian-mixture-model (GMM) supervecor.
➢ Maximum a posteriori (MAP) adaptation is used to estimate GMM parameters.




## Video-Clip Scores

➢ A semantic concept often reappears in shots in the same video clip.
➢ Share a video-clip score with other shots to detect the reappearance.
➢ Video-clip score: the maximum value of shot scores among all the shots in a video clip. The scaling parameter *r* is set to 0.8.

Video-clip score: $s_{\max} = \max_i s_i$ ← Shot score

New shot score: $s'_i = (1-p)s_i + p s_{\max}$

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

Reappearance (Boat_Ship)

## Results & Conclusion

➢ Our best result was **0.281** (Mean InfAP), which is ranked 3rd among participating teams.
➢ Future work: object tracking and localization using deep CNNs.