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.
  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.

GMM Supervectors

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

Results & Conclusion

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