NII at TRECVID Multimedia Event Detection 2014

Sang Phan, Duy-Dinh Le, Shin'ichi Satoh

Abstract

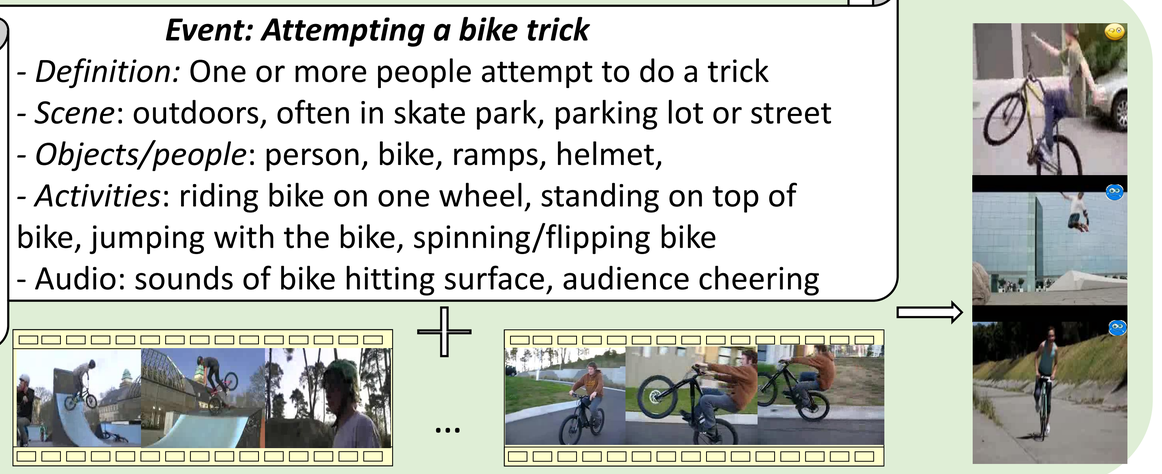
We report our Multimedia Event Detection system with following parts: (1) pre-processing, (2) feature extraction, (3) feature representation and (4) model learning. We use both audio and visual features with Fisher vector encoding.

In the evaluation, we compare the technical improvements of using motion and image features. We also investigate how to use related training videos for learning event model properly.

Multimedia Event Detection

Event: Attempting a bike trick

- *Definition:* One or more people attempt to do a trick
- *Scene*: outdoors, often in skate park, parking lot or street
- *Objects/people*: person, bike, ramps, helmet,
- *Activities*: riding bike on one wheel, standing on top of bike, jumping with the bike, spinning/flipping bike
- Audio: sounds of bike hitting surface, audience cheering



Approach

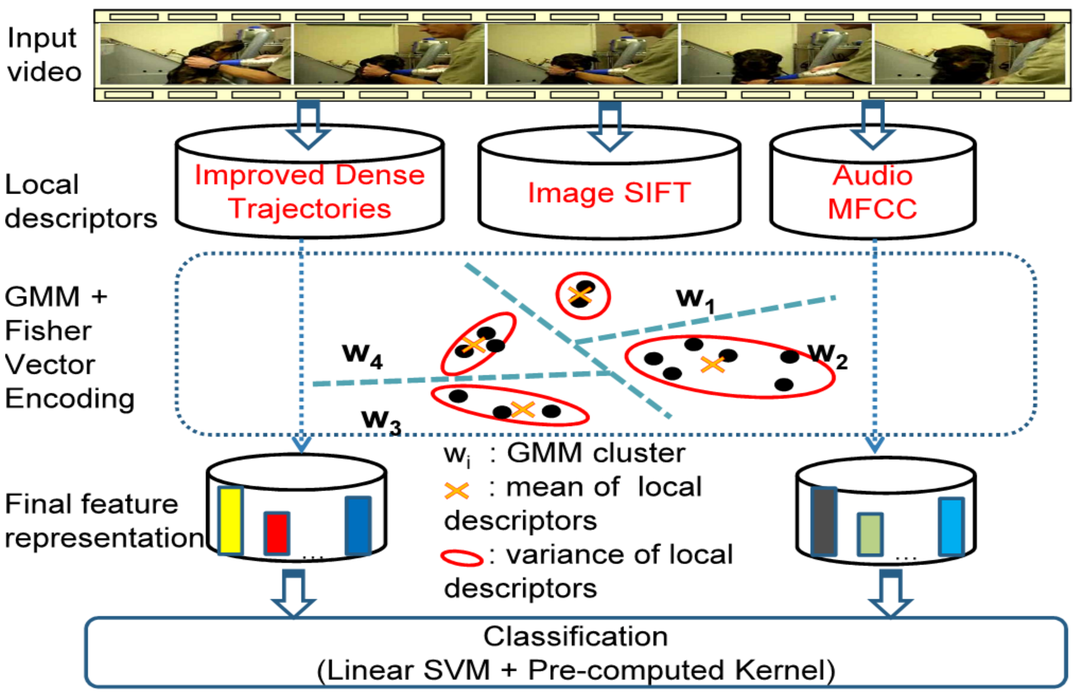


Figure 1. Overview of our MED framework.

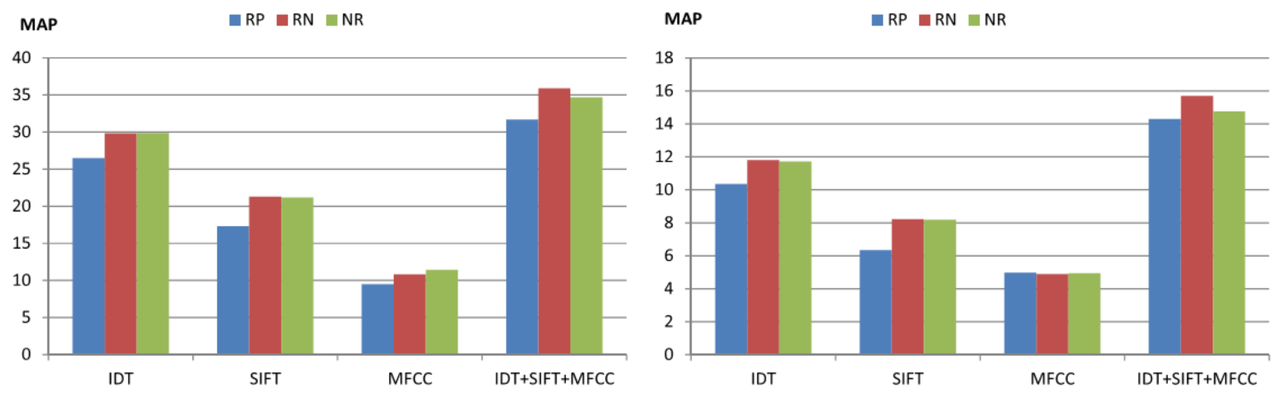
- Still image features: Hessian detector + RootSIFT, no spatial pooling.
- Motion features: Improved Dense Trajectories, Histogram of Optical Flow (HOG) + Histogram of Oriented Gradient (HOG) + Motion Boundary Histogram (MBH).
- Audio features: MFCC with length of 25 ms for audio segments and a step size of 10 ms. The 13d MFCCs + first and second derivatives are used.

Table 1: Performance comparison of different motion feature configurations.

MED13 System	MED14 System	
Dense Trajectories (MBH)	Improved Dense Trajectories (MBH)	Improved Dense Trajectories (HOGHOF + MBH)
28.33	35.07	40.77

Table 2: Performance comparison of different image feature configurations.

MED13 System	MED14 System	
SIFT	SIFT (New aggregation)	SIFT (New aggregation + RootSIFT)
23.41	24.24	27.02



(a) On EK100 setting (b) On EK10 setting

Figure 2. Comparison of different ways to use related videos: Related as Positive (RP), Related as Negative (RN) and No Related (NR).

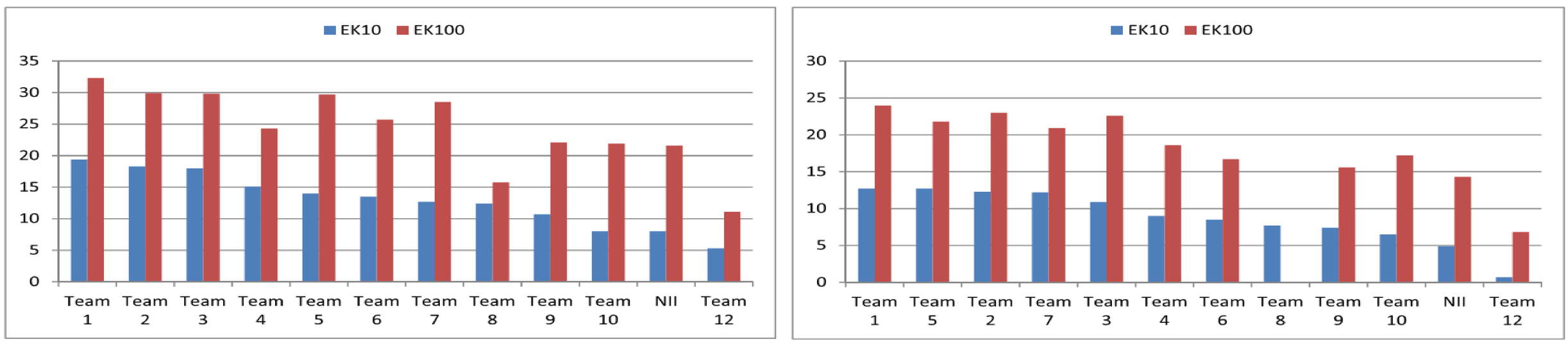
Submitted System & Results

Submitted system:

- Features: we use the same motion feature configuration as in MED13. For image and audio features, we use the new configuration.
- We fixed our system to use related videos as negative training samples for both EK10 and EK100 settings.

Results:

- We achieved rank 11th out of 12 teams in the EK10 setting and rank 10th in the EK100 setting.
- Our system is significantly worse in the EK10 setting. Concept detection is more useful when the number of training video are limited.



(a) Pre-Specified systems (b) Ad-Hoc Systems

Figure 3. Comparison of our MED system with others on the full evaluation set for both PS and AH tasks.

NII
Contact : Sang Phan／ National Institute of Informatics, Japan
TEL : 03-4142-2527 FAX : 03-4142-2120 Email : plsang@nii.ac.jp