A Discriminative CNN Video Representation for Event Detection

Zhongwen Xu[†], Yi Yang[†], Alexander G. Hauptmann[§]
[†]QCIS, University of Technology, Sydney. [§]SCS, Carnegie Mellon University.

Dense Trajectories and its enhanced version *improved Dense Trajectories* (IDT) [1, 6] have dominated complex event detection in recent years. Despite good performance, heavy computation costs greatly restrict the usage of the improved Dense Trajectories on a large scale. In the TRECVID MED 2014, NIST introduced a very large video collection, containing 200,000 videos of 8,000 hours in duration. Even after the spatial re-sizing and temporal down-sampling processing, it still takes 500 cores one week to extract the features [1]. It becomes important to propose an efficient representation for complex event detection with only affordable computational resources, while at the same time attempting to achieve better performance.

One instinctive idea would be to utilize the deep learning approach, especially Convolutional Neural Networks (CNNs), given their overwhelming accuracy in image analysis and fast processing speed. However, it has been reported that the event detection performance of CNN based video representation is worse than the improved Dense Trajectories in MED 2013 [1, 4], as shown in Table 1.

	MEDTest 13	MEDTest 14
IDT [1, 6]	**34.0**	**27.6**
CNN in Lan et al. [4]	29.0	N.A.
CNN_{avg}	32.7	24.8

Table 1: Performance comparison (mAP). CNN_{avg} are our results from the average pooling representation of frame level CNN descriptors.

The contributions of this paper are threefold. First, this is the first work to leverage the encoding techniques to generate video representation based on CNN descriptors. Second, we propose to use a set of latent concept descriptors as frame descriptors, which further diversifies the output with aggregation on multiple spatial locations at deeper stage of the network. The approach forwards video frames for only one round along the deep CNNs for descriptor extraction. With these two contributions, the proposed video CNN representation achieves more than 30% relative improvement over the state-of-the-art video representation on the large scale MED dataset, and this can be conducted on a single machine in two days with 4 GPU cards installed. In addition, we propose to use Product Quantization [2] based on CNN video representation to speed up the execution (event search) time.

The Fisher vector [5] and Vector of Locally Aggregated Descriptors (VLAD) [3] have been shown to have great advantages over Bag-of-Words (BoWs) in local descriptor encoding methods. The Fisher vector and VLAD have been proposed for image classification and image retrieval to encode image local descriptors such as dense SIFT and Histogram of Oriented Gradients (HOG). Attempts have also been made to apply Fisher vector and VLAD on local motion descriptors such as Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) to capture the motion information in videos. To our knowledge, *this is the first work on the video pooling of CNN descriptors and we broaden the encoding methods from local descriptors to CNN descriptors in video analysis*.

In our experiments, we utilize the largest event detection datasets, namely TRECVID MEDTest 13 and TRECVID MEDTest 14. Our proposed representation beats two other strong baselines in *15 out of 20 events* in MEDTest 13 and *14 out of 20 events* in MEDTest 14, respectively. Table 2, Figure 1 and Table 3 show the main results.

	fc_6	fc_6_relu	fc_7	fc_7_relu
Average pooling	19.8	24.8	18.8	23.8
Fisher vector	28.3	28.4	27.4	29.1
VLAD	**33.1**	32.6	**33.2**	31.5

Table 2: Performance comparison (mAP) on **MEDTest 14 100Ex**

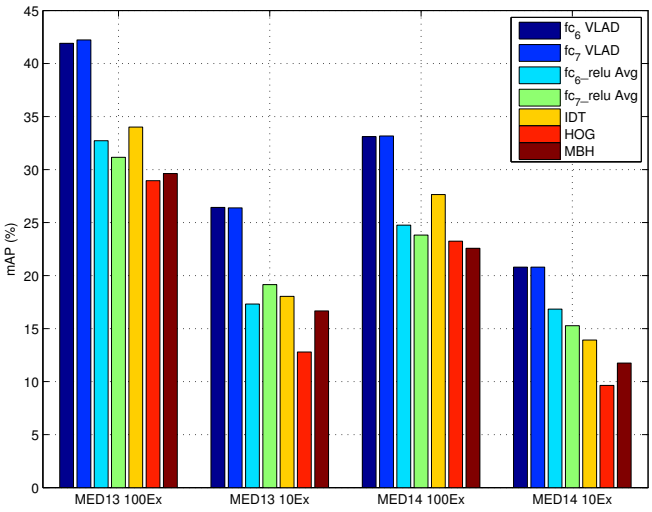

Figure 1: Performance comparisons on MEDTest 13 and MEDTest 14

	Ours	IDT	Relative Improv
MEDTest 13 100Ex	**44.6**	34.0	31.2%
MEDTest 14 100Ex	**36.8**	27.6	33.3%

Table 3: Performance comparison of IDT and our proposed representation.

We compare the MEDTest 13 results with the top performers in the TRECVID MED 2013 competition. Natarajan et al. report mAP 38.5% on 100Ex, 17.9% on 10Ex from their *whole visual system* of combining all their low-level visual features. Lan et al. report 39.3% mAP on 100Ex of their *whole system including non-visual features* while they conducted 10Ex on their internal dataset. Our results achieve 44.6% mAP on 100Ex and 29.8% mAP on 10Ex, which significantly outperforms the top performers in the competition who combine more than 10 kinds of features with sophisticated schemes. To show that our representation is complementary to features from other modalities, we perform average late fusion of our proposed representation with IDT and MFCC, and generate a lightweight system with static, motion and acoustic features, which achieves **48.6% mAP on 100Ex, and 32.2% mAP on 10Ex**.

[1] Robin Aly, Relja Arandjelovic, Ken Chatfield, Matthijs Douze, Basura Fernando, Zaid Harchaoui, Kevin McGuinness, Noel E O'Connor, Dan Oneata, Omkar M Parkhi, et al. The AXES submissions at TrecVid 2013. 2013.

[2] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *TPAMI*, 33(1):117–128, 2011.

[3] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, 2012.

[4] Zhen-Zhong Lan, Lu Jiang, Shoou-I Yu, et al. CMU-Informedia at TRECVID 2013 Multimedia Event Detection. In *TRECVID 2013 Workshop*, 2013.

[5] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[6] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.