USC & THU at THUMOS 2015

Chuang Gan
Tsinghua University
IIIS, Beijing, China
ganchuang1990@gmail.com

Chen Sun Rama Kovvuri Ram Nevatia
University of Southern California
IRIS, Los Angeles, USA
{chensun,nkovvuri,nevatia}@usc.edu

Abstract

This notebook paper describes our approach for the action classification task of the THUMOS Challenge 2015. Our system combines motion and appearance features. For motion features, we adapt the Fisher vector representation with improved dense trajectories. For appearance feature, we compute feature activations from deep convolutional neural networks. We then train SVM classifiers and kernel ridge regression classifiers for each action class. During testing, these classifiers are applied to whole videos as well as temporal sliding windows with different durations. Finally, we combine the classification scores of the whole video and top ranked windows to generate video-level classification scores. Experimental results show that the proposed multi-duration fusion strategy improves the classification results significantly.

1. Introduction

This paper describes our framework in the action classification task of THUMOS Challenge 2015. The goal of the THUMOS Challenge is to recognize a large number of human action classes from open source videos in a realistic setting. In particular the test data consists of temporally untrimmed videos videos, where the actions of interest might happen anywhere within the videos, and multiple instances can be present in each video. For full details on the definition of the challenge, task, and datasets, we refer to [3].

In the following sections, we describe our system's features representation in Section 2, classification framework in Section 3 and experimental details in Section 4.

2. Feature Extraction

2.1. Motion Features

We extracted improved dense trajectories features (iDT-F) [7]. It employs a camera compensation step before feature extraction. We used the default settings as provided by authors.

To obtain video level features, we chose the Fisher Vector coding technique [4, 8], and followed the procedure proposed in [6]. One difference however, is that we encoded the four modalities of iDTF (i.e. HOG, HOF, MBHx and MBHy) separately.

We first extracted the video-level iDTF features for both training and testing sets, then used sliding windows with durations of 20, 60, 100 and 160 frames over videos to extract clip-level iDTF features of testing set.

2.2. Appearance Features

We employed deep net features. We used the output of the first and second fully connected layers of 16 weight layers in the VGG ILSVRC 2014 classification task winning solutions [5]. These features were extracted for 5 frames per second. For the final video representation, we applied average pooling and cross-frame max pooling [2] to each of these output vectors over the frames.

3. Classification

For motion features, we used LIBLINEAR [1] to train 1 vs rest action classifiers. Each feature modality was trained independently, and was combined using late fusion by taking the geometric mean of the confidence scores. For appearance features, we used kernel ridge regression as described in [9] to train 1 vs rest action classifiers.

During testing , we applied the classifiers to both video-level features and clip-level features. To obtain video-level confidence scores, we took the average of top 5 clip-level scores for motion features and top 30 clip-level scores for CNN features.

Then we weighted fuse the scores of different features and use the best combinations obtained on Validation set for the Test set.

4. Experiment Setup

For all runs, we only used the videos in training set for training, and verified the performance on validation set.

Note that the performance of the proposed multi-duration fusion strategy can improve the action classification performance from 66.2% to 71.4% in term of mean Average Precision (mAP) on the validation set.

For iDTF features, we set the number of cluster centers of Fisher vectors to 256, and projected each iDTF modality to half of their original dimensions with PCA.

We fixed the SVM cost to 100 and bias to 10. For kernel ridge regression, we fixed the λ to 1.

5. Acknowledgement

Chuang Gan was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003.

References

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[2] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015.

[3] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015.

[4] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[6] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In *WACV*, 2013.

[7] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[8] X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *ACCV*, 2012.

[9] S.-I. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, et al. Informedia@ TRECVID 2014 MED and MER.