USC at THUMOS 2014

Chen Sun and Ram Nevatia

University of Southern California, Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089, USA

Abstract. We submitted one run for THUMOS 2014 action recognition task. The system used improved dense trajectory features and fisher vector coding. Since testing videos are temporally untrimmed, we applied a sliding window of 100 frames for both training and testing videos. We utilized the background videos by iteratively training SVM classifiers and selecting hard negative samples. Video-level scores were generated by maximum pooling.

Keywords: fisher vector, hard negative mining

1 Introduction

We present our system's action recognition performance on UCF101 dataset [3]. UCF101 is an action recognition data set of action videos collected from YouTube. There are 13,320 temporally segmented videos from 101 action categories, the typical length of the videos is less than 10 seconds. This temporal segmentation of testing videos might not reflect the real world as most of the web videos are temporally untrimmed.

To address this issue, this year's challenge is evaluated on around 1,500 temporally untrimmed videos, where the actions of interest might happen anywhere within the videos. We decided to use temporal sliding windows, and utilized the additional background videos by hard negative mining.

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

2 Clip Level Features

We extracted improved dense trajectories features (iDTF) [5]. 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 [2, 6], and followed the procedure proposed in [4]. One difference however, is that we encoded the four modalities of iDTF (i.e. HOG, HOF, MBHx and MBHy) separately, source code for Fisher Vector generation is available online[1].

We used sliding windows of 100 frames to extract clip-level features of training, background and testing set.

[1] https://github.com/chensun11/dtfv

3 Classification

We used LIBLINEAR [1] to train 1 vs rest action classifiers. Each feature modality was trained separately, and combined later by taking the average of confidence scores.

A large collection of long background videos are available. To make use of it, we applied a multi-stage training strategy by iteratively mining negative samples. For each iteration, we kept all the positive training instances for that class, and applied the trained linear classifier from previous iteration to all the negative samples. We then selected negative samples with the highest positive confidence scores. The first set of negative samples was selected by random.

During inference, we applied the SVM classifiers to all the testing video clips. Video level confidence scores were computed by taking the average of top 2 clip-level scores.

4 Experiment Setup

We used the training and background videos for training, and verified the performance on validation set. To speed up the whole process, we set the number of cluster centers of Fisher Vectors to 64, and projected each iDTF modality to half of their original dimensions with PCA. The length and step size of sliding windows were 100 frames and 50 frames respectively.

We set the SVM parameters empirically, where cost was set to 1 and bias was set to 10. This was found to yield state-of-the-art performance on the THUMOS 13 setting. 5,000 negative samples were mined for training at each iteration, up to 5 iterations.

References

1. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research (2008)
2. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
3. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. CRCV-TR-12-01
4. Sun, C., Nevatia, R.: Large-scale web video event classification by use of fisher vectors. In: WACV (2013)
5. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
6. Wang, X., Wang, L., Qiao, Y.: A comparative study of encoding, pooling and normalization methods for action recognition. In: ACCV (2012)