Abstract

We submitted one run for THUMOS 13 action recognition task. The system used low level motion features and Fisher Vector coding. We trained 101 one-versus-rest action classifiers based on each type of low level features, and performed late fusion. Experimental results for 3 train/test partitions on UCF 101 are reported.

1. Introduction

We present our system’s action recognition performance on UCF101 dataset [3]. UCF101 is an action recognition data set of realistic action videos collected from YouTube. There are 13320 videos from 101 action categories.

We describe our system’s video level features in Section 2, classification framework in Section 3, and experimental results in Section 4.

2. Video Level Features

Raw features were extracted using the code provided by authors. For DT features, we set the sampling stride to 10, trajectory length to 15, the total feature dimension is 426. For MOSIFT features, we set the pixel movement threshold from 1 to 100, the feature dimension is 256.

To obtain video level features, we chose the Fisher Vector coding technique, and followed the procedure proposed in [4]:

1. Use PCA to reduce the dimension of raw feature space (426 for DT and 256 for MOSIFT) to 128.
2. Learn a Gaussian Mixture codebook with 64 clusters from randomly selected dimension reduced feature points.
3. Aggregate all feature points from the same video with Fisher Vectors. Use both first order terms and second order terms.

We also used a 2 level spatial pyramid to obtain a set of Fisher Vectors for each pyramid and concatenated them as a single video level feature.

3. Classification

LIBSVM [1] with Gaussian kernels was used to train action classifiers. We used a one-versus-rest strategy, there were 101 action classifiers trained by DT Fisher Vectors (DT_FV) and MOSIFT Fisher Vectors (MOSIFT_FV) respectively. For each train/test partition, we selected the SVM parameters based on 5-fold cross validation on the training partition.

To combine the results from DT_FV and MOSIFT_FV, we used late fusion by geometric mean

\[ s = \sqrt{s_1 \cdot s_2} \]  

where \( s_1 \) and \( s_2 \) are the action confidence scores to be fused.

For each video, action with the highest confidence score was assigned.

4. Experimental Results

Three train/test partitions are provided by THUMOS 13. We learned action classifiers for each training partition, and report the classification accuracy in Table 1. Performance with a single type of low level feature are listed in the table as DT_FV and MOSIFT_FV. The Fused column gives late fusion results. It can be seen that late fusion provides consistent improvement.

5. Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Depart-
ment of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. Computation for the work described in this paper was supported by the University of Southern California Center for High-Performance Computing and Communications (hpcc.usc.edu).

References


