Action Recognition with Shot Boundary Detection and Decoded iDT Features

Yu Wang and Jien Kato
{ywang, jien}@nagoya-u.jp

Graduate School of Information Science
Nagoya University, Japan

Abstract. We report our results on the THUMOS Challenge 2014 Action Recognition Task. Given an untrimmed video as input, our method recognizes its major action category in four main steps: (1) detect shot boundaries within the video; (2) decode pre-computed indices of iDTFs within each shot; (3) encode the recovered iDTFs into shot-wise fisher vectors and compute the shot-to-category classification scores; (4) summarize shot-wise classification scores into a video-wise classification score. We report results on four different strategies that are used for summarizing the video-wise classification scores.

Keywords: action recognition, shot boundary detection

1 Introduction

In THUMOS Challenge 2014, test videos are provided as untrimmed. In order to recognize the major action category in such kind of videos, a straight forward way is to conduct shot boundary detection to trim the videos before applying action classifiers. Based on this simple idea, we developed a system for this year's recognition task. The system encodes a input video into a number of shot-wise fisher vectors. Classification is then conducted on these fisher vectors using pre-trained action classifiers, with the resulted shot-wise shot-to-category classification scores being further summarized into a video-wise classification score (in four different ways) as the output. Beside the shot boundary detection, our another try is to compute fisher vectors from the pre-computed improved Dense Trajectory Features (iDTFs) [5] indices. As the raw features that provided by the organizers are indices of iDTFs, we decode it to the distorted full-length features first and then do the FV encoding. In the following section, we describe the implementation details of the system.

2 Implementation Details

Given an untrimmed video as input, our system recognizes its major action category in four main steps: (1) shot boundary detection; (2) decoding of the precomputed iDTFs indices; (3) encoding the recovered iDTFs into fisher vectors

and classify; (4) summarize the shot-wise classification scores into a video-wise classification score.

Shot boundary detection is a relative mature technology in video processing which has many promising methods in the literature [4]. In this work, we adapt the approach in [1], which utilizes both SURF and HSV histogram to measure the similarity between frames. Specifically, we utilized the binary release from the author's homepage. Because the original release has limitations on the lengths and the resolutions of input videos, as well as a little bug in the generated outputs, we wrote a Matlab script which consists of functions such like cutting-and-merging, resolution up-sampling, and late manipulation of raw outputs.

Once the shot boundaries have been detected, pre-computed iDTFs indices for each shot are collected. Since iDTFs are computed on trajectories across 16 frames, when collecting the iDTFs, we use a time window that starts in the detected starting boundary and ends 15 frames after the detected ending boundary, to make sure all trajectories that have overlap with the shot are included. For each trajectory, its indices of HOG, HOF and MBH are used to recover a distorted 396 dimensional feature vector by first looking up their corresponding items in the vocabulary and then concatenate these vocabulary items together.

The 396d feature vectors of the trajectories within each shot are then processed into a fisher vector through PCA and FV enccoding [3]. The PCs and GMMs are learned from $512,000$ such feature vectors from the training set. Specifically, the top 129 PCs were selected for projection in PCA (99% of the total variance can be preserved); the number of GMM was set to 256; L2 Normalization and Power Normalization were implemented. Such a setting results in a $66,048$ dimensional feature vector for each shot. The FVs of each shot are then classified using the 101 linear classifiers which are learned from the training data using LIBLINEAR library [2]. In this work, the results are produced without using the background data and the validation data.

After obtaining the shot-category classification scores, we summarize them to a video-wise classification score. The action category with the highest classification score are considered as the major category of the input video. We implemented four different strategies for summarizing the final classification score: (1) take the classification scores of the longest shot; (2) take the maximum score for each category (max pooling); (3) take the average score for each category (mean pooling); (4) take the weighted sum of all scores (scores of a shot is weighted by the shot's length) for each category. Beside these four strategies, in order to confirm how the decoded iDTFs work, we also implemented a baseline method which treat the input video as a single shot.

References

1. Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors (2014), 2014 IEEE International Conference on Acoustics, Speech and Signal Processing

2. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
3. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification (2010), 2010 European Conference on Computer Vision
4. Smeatona, A., Overb, P., Doherty, A.: Video shot boundary detection: Seven years of trecvid activity. Computer Vision and Image Understanding 114(14), 411–418 (2010)
5. Wang, H., Schmid, C.: Action recognition with improved trajectories (2013), 2013 International Conference on Computer Vision