Towards Good Practices for Action Video Encoding

Jianxin Wu
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
wujx2001@nju.edu.cn

Abstract

In this notebook paper, we propose and evaluate different post-feature-extraction encoding processing for action videos in action recognition. Recently, high dimensional representation such as VLAD or FV has shown excellent accuracy in action recognition. This paper shows that a proper encoding built upon VLAD can get further accuracy boost comparable to the accuracy gains achieved by replacing bag-of-features with the VLAD representation.

1. Introduction

The past decade has witnessed increasing interests on action recognition in videos (e.g., [6, 13, 12, 8, 5, 16]). Many methods in this vast literature are based on the bag-of-features video encoding scheme. Without loss of generality, we can decompose a bag-of-features based action recognition pipeline into three steps:

- **Raw motion feature extraction.** We want to encode motion, arguably the most important cue for action recognition. Examples of popular raw motion feature include the space-time interest points (STIP) [6] and dense trajectory features (DTF) with motion boundary histograms (MBH) [16].
- **Action video encoding.** After raw motion features are extracted from a video, bag-of-features encodes this set into a single vector (e.g., [13, 7, 5, 16]).
- **Learning.** Various classifiers (e.g., support vector machines) learn action recognition models from the encoding of training videos.

Among these steps, motion features have enjoyed the most attentions in the past. The DTF features in [16] significantly outperformed previous state-of-the-art results on 9 benchmark datasets. As one example, recognition accuracy on HMDB51 dataset jumped from 26.9% (in [9]) to 48.3% (in [16]). Finding good motion features may continue to maintain its central role in action recognition research, e.g., the ω-flow features further increased accuracy on HMDB51 to 52.1% [2].

The VLAD encoding framework was used in [2] instead of the classic bag-of-features, a fact that contributed a lot to its accuracy gain. One very recent trend is that high-dimensional encoding frameworks such as Fisher Vector (FV) [10] or VLAD [3] are gradually gaining popularity in video representation—mainly due to their excellent recognition accuracies. For example, VLAD exhibited significant improvements on two difficult datasets (Hollywood2 and HMDB51) over bag-of-features [2]. In [15], using FV or VLAD can both improve event classification accuracy by a large margin. Similar results are reported in [11] for FV with small number of code words.

In this paper, through empirical evaluations, we show that proper video encoding can further boost action recognition accuracy, and the gains are comparable to how FV or VLAD improves bag-of-features.

2. Bimodal encoding

Suppose $x$ is a variable representing one dimension in VLAD (after power and $\ell_2$ normalization), we denote $\pi(>0)$ and $\varepsilon(<0)$ as the maximum and minimum value of this dimension observed in the training set, respectively. The encoding transformations are defined as:

- **Even-Scale.** $x \leftarrow \frac{x - \pi}{\varepsilon - \pi}$. The new $x$ has varying minimum and maximum values in different dimensions, but the range (difference between minimum and maximum) in each dimension is always 1. Note that 0 remain unchanged.
- **Side-Info.** $x \leftarrow \frac{x - \varepsilon}{\pi - \varepsilon}$. Note that 0 will be transformed to $\frac{-\varepsilon}{\pi - \varepsilon}$, thus it is a simple way to incorporate the side information (i.e., which code word is missing in which video). The new range is $[0 1]$ in all dimensions. Thus, the scale is normalized in this transformation.
- **Bimodal.** $x$ is transformed as follows:
  
  $x \leftarrow x / \pi$ if $x \geq 0$

  and

  $x \leftarrow x / |x|$ if $x < 0$.  

This encoding handles the bimodality explicitly and fix scales in both sides of 0. The new range is \([-1, 1]\). Note that 0 remain unchanged.

Evaluation results for different encoding transformations are presented in Table 1. In this table, we used the dense trajectory features (DTF) [16] and the VLAD representation. In HMDB51, spatiotemporal pyramid (split into 3 temporal regions) are used. The codebook size is 256 in VLAD. In classification, we used the LIBLINEAR software package [1] with its default parameters, except that the option “-B 1” is used to replace the default option value “-B -1”. For HMDB51, we followed the original evaluation protocol to use 3 train / test splits and report the average accuracy [5]. For Youtube, we followed the original protocol to perform 25-fold cross validation [8]. For UCF101 [14], we followed the THUMOS 2013 challenge protocol [4].

### 2.1. Comparison of encodings and comparison with state-of-the-art

Youtube is an easy dataset (with around 88% accuracy), and all improvement techniques (root descriptor, residue normalization, spatio-temporal pyramid and the proposed encodings) cannot improve performance significantly. On the other two datasets, however, there is 1–2% improvements.

The first row in the second block of Table 1 shows bag-of-features results. HMDB51 and Youtube results are from [16]. For UCF101, we used \(K = 4000\) in bag-of-features and the PmSVM classifier [18] with the histogram intersection kernel (\(p = -16, C = 0.1\) in PmSVM). Comparing with bag-of-features, in UCF101 and HMDB51, 1.74% and 2.81% are gained by using VLAD to replace bag-of-features, respectively; while 1.06% and 1.90% are further contributed by the bimodal encoding, close to what VLAD buys us in action recognition. Further considering its ease of implementation and light computational overhead, the proposed bimodal encoding is a good practice for action video encoding.

### 3. THUMOS challenge

In the THUMOS 2013 challenge, we further improve the performance by replacing the DTF motion features with the improved trajectory features in [17]. With this improvement, our accuracy on the UCF101 dataset is 83.54%.

### References


