Improving Bag-of-Features Action Recognition with Non-Local Cues

Muhammad Muneeb Ullah
Muhammad.Muneeb.Ullah@inria.fr
Sobhan Naderi Parizi
Sobhan.Naderi_Parizi@inria.fr
Ivan Laptev
Ivan.Laptev@inria.fr

INRIA - Willow Project
Laboratoire d'Informatique
École Normale Supérieure
CNRS/ENS/INRIA (UMR 8548)

Local space-time features have recently shown promising results within Bag-of-Features (BoF) approach to action recognition in video. Pure local features and descriptors, however, provide only limited discriminative power implying ambiguity among features and sub-optimal recognition performance. In this work, we propose to disambiguate local space-time features and to improve action recognition by integrating additional non-local cues with BoF representation. For this purpose, we decompose video into region classes and augment local features with corresponding region-class labels. For example, regions of a parking lot and side walks in Figure 1 are likely to correlate with specific actions such as opening a trunk and running. Propagating region labels to the local feature level in this example is therefore expected to increase discriminative power of local features with respect to particular actions.

We build upon BoF framework and represent videos with Harris3D features [2] and associated HOG/HOF descriptors. Feature descriptors are vector-quantized using either k-means visual dictionary, or a supervised quantization method based on ERC-Forests [3]. Our baseline BoF video representation corresponds to l_1-normalized histograms of visual words. To enrich BoF representation, we propose to decompose video into a set of regions r associated with labels l, $l \in \{L^1, \ldots, L^M\}$ such that regions with the same labels share common properties. We then accumulate a separate BoF histogram h^i from all features within L^i-labeled regions. Video descriptor (a *channel*) is constructed by concatenating BoF histograms for all region labels, i.e. $x = [h^1, \ldots, h^M]$ as illustrated in Figure 2. For action classification we use SVM with RBF-χ^2 kernel, and use product of kernels to combine multiple channels.

We test our approach using readily-available segmentation methods and explore alternative segmentation strategies to (i) improve discrimination of different action classes and (ii) to reduce effects of errors of each segmentation approach. Below we briefly summarize five types of video segmentation used in this work (see also Figure 3 for the illustration).

1. Spatio-temporal grids (STGrid24): We divide a video into a set of 24 pre-defined spatio-temporal grids.

2. Foreground/background motion segmentation (Motion8): We segment video into foreground and background regions using motion segmentation. Separate histograms for 2 types of regions and 4 values of segmentation threshold generate 8 channels.

3. Action detection (Action12): We train Felzenszwalb's detector [1] on action images collected from the Web and segment video into action and non-action regions according to detected bounding boxes and six values of detection threshold. We generate 12 channels per action for 6 threshold values and 2 types of grids on action regions.

4. Person detection (Person12): We use Calvin upper-body detector and segment video into person and non-person regions. Following the procedure for Action12 channels above, we obtain 12 channels for 6 threshold values and 2 types of grids on person regions.

5. Object detection (Objects12): We use Pascal VOC08 pre-trained Felzenszwalb's object detectors [1] and segment video into object and non-object regions for four object classes: car, chair, table, and sofa. We generate 12 channels per object class using 6 threshold values and 2 types of grids on object regions as for Action12 and Person12 channels above.

We report results of action classification on Hollywood-2 dataset using mean Average Precision (mAP). Table 1 compares baseline results for the two alternative quantization methods. Table 2 presents results for individual channels as well as for their combination using ERC-Forest quantization. All new channels improve the baseline performance when combined with STGrid24 channel. Moreover, the combination of all channels further improves performance significantly up to mAP 0.553.

In conclusion, the proposed method improves action classification significantly and can benefit further from additional segmentations.

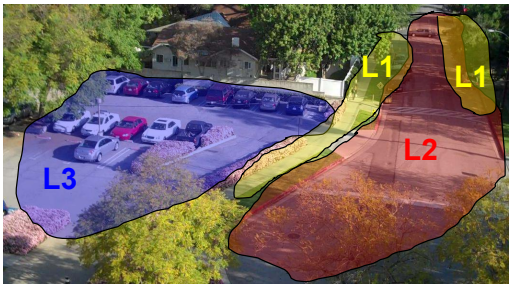

Figure 1: Regions in video such as road, side walk and parking lot frequently co-occur with specific actions (e.g. driving, running, opening a trunk) and may provide informative priors for action recognition.

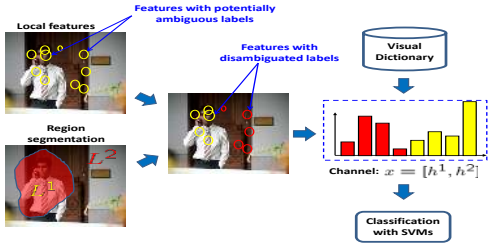

Figure 2: An illustration of our approach to disambiguate local descriptors with the help of semantic video segmentation.

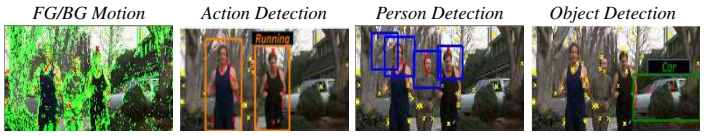

Figure 3: Illustration of proposed semantic region extraction and feature separation in video.

Channels	Performance (mean AP)
BoF with K-Means	0.479
BoF with ERC-Forest	**0.486**
STGrid24 with K-Means	0.504
STGrid24 with ERC-Forest	**0.518**

Table 1: Hollywood-2 classification obtained with the baseline channels.

Channels	Performance (mean AP)
Motion8	0.504
Person12	0.493
Objects12	0.499
Action12-class (specific)	**0.528**
STGrid24 + Motion8	0.532
STGrid24 + Person12	0.532
STGrid24 + Objects12	0.530
STGrid24 + Action12-class (specific)	**0.557**
STGrid24 + Motion8 + Action12-class + Person12 + Objects12	**0.553**

Table 2: Overall performance of individual channels and their different combinations.

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[2] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2/3):107–123, 2005.

[3] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1632–1646, 2008.