Interaction Part Mining: A Mid-Level Approach for Fine-Grained Action Recognition

Yang Zhou[1], Bingbing Ni[2], Richang Hong[3], Meng Wang[3], and Qi Tian[1]
[1]University of Texas at San Antonio, US
[2]Advanced Digital Sciences Center, Singapore
[3]HeFei University of Technology, China

In recent years, fine-grained action recognition has raised extensive research [7, 9] due to its potential applications in assisted daily living, medical surveillance and smart home.

On one hand, fine-grained manipulation actions involve a large amount of interactions between human and objects. Therefore, how to model the interactions between human and objects plays a critical role in action representation and recognition. Many research works have devoted to modeling the contextual information between human and objects/scenes for action recognition. However, to model human and object contextual information, explicit detection of objects is often required by the above methods. Training these object detectors requires labour-extensive human annotation work. In fine-grained action recognition where many types of objects are manipulated in a single action, it is not feasible to 1) label the training object instances; and 2) detect those objects with decent detection accuracy.

On the other hand, the spatio-temporal features [6], especially the recent proposed dense trajectories [11] encoded with naive BoW [3] or Fisher vector representation [8] are commonly used for action recognition, but they can only capture the characteristics of the global motion in the entire video volume. The low-level feature extraction with global pooling might not be suitable for representing fine-grained actions because this global presentation easily attenuates the important localized interaction motion within the background movement. In contrast, for fine-grained motion, it is more important to highlight what kind of interaction motion is being performed in a local spatio-temporal sub-volume of the video. Therefore, we need a mid-level representation to capture the important local human-object interaction motion.

To address the above two issues, we propose a novel mid-level based pipeline for fine-grained action representation and classification, which is motivated by two recent successes in visual recognition, *i.e.*, object proposal technique [1] and mid-level discriminative visual element mining [2, 4, 5]. We show the general framework in Figure 1.

Firstly, we utilize the off-the-shelf object proposal BING [1] to generate a large number of object proposals (*i.e.*, candidate regions) on the segmented foreground motion pixels. Then we construct a spatio-temporal graph by matching the object proposal spatio-temproally based on both dense trajectory linkage and appearance similarity. An efficient graph segmentation algorithm is proposed to group the object regions which are temporally continuous and spatially compact into spatio-temporal sub-volumes. We name these sub-volumes as *interaction parts*. The extracted interaction parts have two advantages: 1) the extraction procedure is unsupervised which means object annotation is not required; and 2) the extracted interaction parts naturally contain mid-level information on what object and what motion is being performed.

Secondly, for each interaction part, we compute a Fisher vector representation by pooling the spatio-temporal motion features of the sub-volume. We learn a set of discriminative part detectors using an improved version of the image block mining approach [5], *i.e.*, to seed discriminative interaction parts by considering both appearance and motion features. Finally, we utilize the trained part detectors to score the interaction parts within each video, and summarize the part scores for each video sample using a generalization of max pooling technique, *i.e.*, Max-N pooling. We validate our mid-level based video representation on the MPII Cooking [10] and MSR Daily Activity 3D [12] datasets. The results show that our mid-level approach outperforms the state-of-the-art performance on both datasets. It is more effective than low-level features in capturing human-object interaction motion, and it is even better than the previous approach [13] which requires object detection.

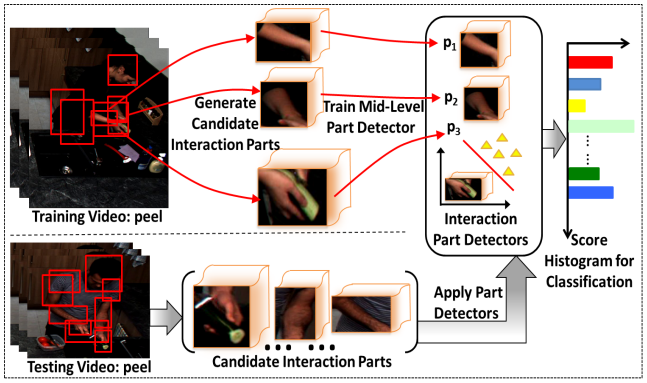

Figure 1: Our mid-level approach for fine-grained action recognition includes interaction part formation to generate candidate human-object interaction parts, and interaction part mining to train discriminative exemplar part detectors.

We conclude our contribution as follows. We propose a mid-level video representation for fine-grained action recognition. The method effectively captures the significant human-object interaction motion by object proposal based interaction part formation and discriminative interaction part mining. Most importantly, our method is free of explicit object detection, which gives us three major advantages: 1) it is more stable than the object detection approaches which are quite affected by the detection accuracy to different objects; 2) it saves extensive human labour for object annotation; and 3) it is quite applicable and feasible in real problems. Furthermore, we propose a novel Max-N pooling which improves performance compared to the previous naive max pooling approach.

[1] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[2] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[3] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[4] Arpit Jain, Abhinav Gupta, Mikel Rodriguez, and Larry S Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2013.

[5] Mayank Juneja, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[6] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[7] Jinna Lei, Xiaofeng Ren, and Dieter Fox. Fine-grained kitchen activity recognition using rgb-d. In *UbiComp*, 2012.

[8] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.

[9] Donald J Patterson, Dieter Fox, Henry Kautz, and Matthai Philipose. Fine-grained activity recognition by aggregating abstract object usage. In *ISWC*, 2005.

[10] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[11] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103 (1):60–79, 2013.

[12] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[13] Yang Zhou, Bingbing Ni, Shuicheng Yan, Pierre Moulin, and Qi Tian. Pipelining localized semantic features for fine-grained action recognition. In *ECCV*. 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.