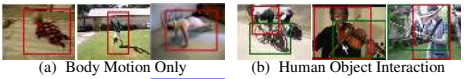
Two-Stream SR-CNNs for Action Recognition in Videos

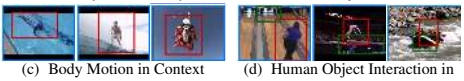
Yifan Wang[1]
yifan.wang@student.ethz.ch
Jie Song[1]
jsong@inf.ethz.ch
Limin Wang[2]
07wanglimin@gmail.com
Luc Van Gool[2]
vangool@vision.ee.ethz.ch
Otmar Hilliges[1]
otmar.hilliges@inf.ethz.ch

[1] Advanced Interactive Technologies Lab, ETH Zurich, Switzerland
[2] Computer Vision Lab, ETH Zurich, Switzerland

Action recognition belongs to the most challenging tasks in computer vision. An action is usually defined by multiple elements, called "cues", e.g. person, object and scene. Accordingly, common actions can be divided into four types as shown in Figure 1.



(a) Body Motion Only (b) Human Object Interaction



(c) Body Motion in Context (d) Human Object Interaction in Context

Figure 1: Action types with different composition of semantic cues, e.g. human body (red box), interacting objects (green box), and global context (blue box)

Despite the high overall classification accuracy [4], the conventional two-stream CNN approach [2] performs poorly on human-centric categories (shadowed plot in Figure 3). This discovery indicates overfitting from uninformative variances possibly from "scene". Hence we propose a semantically aware CNN-based framework for action recognition in video, which uses the locational information of various semantic cues as an explicit attention guidance during training and testing.

1 Approach

First of all, we propose a generic and efficient method to extract action relevant persons and objects from video sequences using the output of an object detector, e.g. Faster R-CNN [1]. This method recovers detection errors and removes irrelevant "bystanders" devoid of ground truth. The obtained bounding boxes are incorporated into the conventional two-stream CNNs network via a RoiPooling layer as shown in Figure 2. Each semantic cue constructs an individual channel, which is combined by a fusion layer to produce the final prediction.

2 Experiment and Result

We conduct a series of experiments on UCF101 dataset [3] and determine the best performing model, namely SR-CNNs with sum-fused person and scene

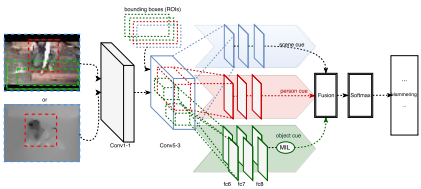


Figure 2: Architecture of two-stream SR-CNNs

channels, denoted as *S+P*. Our empirical study demonstrates that (1) our approach outperforms the original two-stream CNNs in terms of global accuracy (Table 1) (2) the robustness against ambiguous variances in scene (Figure 3) is improved (3) semantic channels exhibit complementary properties and improves spatial and temporal streams differently.

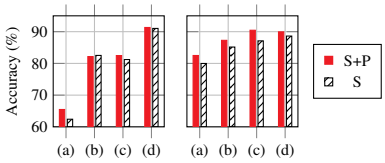


Figure 3: Performance comparison on UCF101 split1 in different action types defined in Figure 1. (left: spatial stream; right: temporal stream)

Models	Spatial	Temporal	Two Stream
S	77.93	86.79	91.15
S+P	**78.32**	**88.29**	**92.60**

Table 1: Comparison to conventional two-stream CNNs on UCF101 averaged over 3 splits

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[2] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[3] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[4] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.