# Kinetics-TPS Track on Part-level Action Parsing and Action Recognition Technical Report

Hailun Zhang[1], Zhi Li[2], and Qijun Zhao[3]

[1] Sichuan University, Chengdu, China tamakoko@stu.scu.edu.cn
[2] Sichuan University, Chengdu, China lizhi@stu.scu.edu.cn
[3] Sichuan University, Chengdu, China qjzhao@scu.edu.cn

**Abstract.** This short report introduces the implementation details on Part-level Action Parsing and Action Recognition in ECCV DeeperAction Challenge Kinetics TPS Track. We designed a human-part-state recognition network based on multiple attention blocks, in which features from body parts can be extracted and employed for action recognition. In the competition, we achieved 25.19% mAP on the test set of Kinetics-TPS.

**Keywords:** Action Recognition, Video Recognition.

## 1 Introduction

Understanding action from images is crucial for building an intelligent system. Recent works [8][6] about action recognition mainly treat it as high-level classification problem, i.e. from pixels to action concept directly based on instance-level semantics. In the Kinetics-TPS track, we aim at understanding the human actions based on human part states.

The pipeline of our work is showed in Fig. 1. A human part and object detector is firstly performed on every frame, and then the part and object feature in the frame will be extracted and fed to the part relevance predictor, after that, we extract the language priors and bridge the gap between part states and action semantics, and then sequential model will be used to classify actions.

Detailed algorithm is discussed in Section 2.

## 2 Approach

The overall framework of our method is shown in Fig.1. We will orderly introduce the proposed models for human part detection, part parsing and action recognition.

### 2.1 Human part and object detection

To detect human part and object from each frame of the video, we adopt the state-of-the-art object detection approach based on Cascaded-RCNN [1]. Although the video-based object detection algorithm can accept temporal context

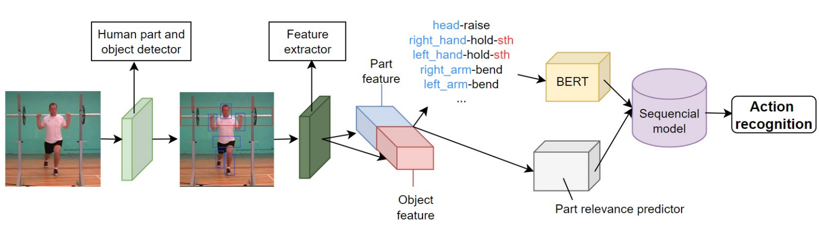


**Fig. 1.** The overall framework of our method.

information to effectively alleviate the position frame deviation caused by motion blur, small targets, etc., there is not a large impact on the accuracy of our task. Then we use Faster-RCNN [4] to extract the body part and object features in each frame of the video. We choose the VGG-16 as its backbone pretrained on COCO[7] dataset. The same training strategy is employed with the original Faster-RCNN.

### 2.2 Part Parsing

Since the state of human body parts is highly correlated with current human activity, we consider fuse the body part and object features to parse part state.

We input part feature and object feature to a part relevance predictor. Part relevance represents how important a body part is to the action. For example, feet usually have weak correlations with "drink with cup". And in "drink tea", only hands and head are essential. The part relevance predictor consists of FC layers and Sigmoids, which infers the relevance of each part and the corresponding objects. With part relevance predictor, we use part relevance labels as supervision and construct cross-entropy loss to obtain relevance score for each part. Because a part can have multiple states, e.g. head performs "eat" and "watch" simultaneously. Hence we use multiple Sigmoids to do this multi-label classification.

### 2.3 Action Recognition

Our goal is to bridge the gap between part state and activity semantics. Language priors are useful in visual concept understanding[5]. Thus the combination of visual and language knowledge is a good choice for establishing this mapping. To further enhance the representation ability, we utilize the uncased BERT-Base pre-trained model [2] as the language representation extractor. Bert [2] is a language understanding model that considers the context of words and uses a deep bidirectional transformer to extract contextual representations.

In specific, for the i-th body part with n part state, we convert each part state to a $f_{Bert} \in R^{2304}$ (concatenating three 768 sized vectors of part, verb, object). Second, we multiply $f_{Bert}$ with predicted part state probabilities $P_{part-state}$,

$i.e. f_{part-state} = f_{Bert} \times P_{part-state}$, where $P_{part-state} = Sigmoid(S_{part-state})$, $S_{part-state}$ denotes the part state score of the i-th part.

We pool and resize the $f^{L(i)}_{part-state}$, and concatenate it with its corresponding visual part state feature $f^{V(i)}_{part-state}$. Then we obtain the part state representation $f^{(i)}_{part-state}$ for each body part. It is the part-level activity representation and with it, we use a Hierarchical Activity Graph (HAG) [3] to model the activities. Then we can extract the graph state to recognize the action. We adopt sequential model implementation of HAG, which uses LSTM to take the part feature gradually, and uses the output of the last time step to classify actions.

## 3 Experiments

**Dataset.** We only use the competition training set for experiments. Kinetics-TPS contains 3,809 training videos (4.96GB in size) and 932 test videos (1.26GB in size).

**Training.** For action recognition, we conduct experiments on a service with a single RTX3090. The input is scaled to 256x256 and then randomly cropped to 224. Following the guidelines of the challenge, we set HUMAN IOU THRESH as 0.5 and set PART IOU THRESH as 0.3 in frame-level action prediction.

## 4 Conclusions

We designed a human-part-state recognition network based on multiple attention blocks, in which features from body parts can be extracted and employed for action recognition. In the competition, we achieved 25.19% mAP on the test set of Kinetics-TPS.

## References


1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Li, Y.L., Xu, L., Liu, X., Huang, X., Xu, Y., Wang, S., Fang, H.S., Ma, Z., Chen, M., Lu, C.: Pastanet: Toward human activity knowledge engine. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 382–391 (2020)
4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
5. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
6. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR 2011. pp. 3177–3184. IEEE (2011)
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
8. Yao, B., Fei-Fei, L.: Action recognition with exemplar based 2.5 d graph matching. In: European conference on computer vision. pp. 173–186. Springer (2012)