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

Zheming Yu, Lin Li, Jietian Guo
Hikvision Research Institute
{yuzheming, lilin37, Guojietian}@hikvision.com

Abstract

This short report introduces the implement details on Part-level Action Parsing and Action Recognition in ICCV DeeperAction Challenge Kinetics TPS Track.

In this challenge, a novel human-part-state recognition network is proposed based on the multi-head attention mechanism, in which various features from body parts are employed to enhance the robustness of model. Our experiment result achieves 63% scores .

1. Introduction

This task aims at locating the human location, body part location and their part states simultaneously in the frame level. The challenges include:

- Predict spatial position of the human body and body parts accurately and simultaneously;
- Define a new the human body part state recognition task;
- Whether the end-to-end framework is the only choice?

The pipeline of our work is showed in Fig. 1. A human detector is firstly performed on every frame, and then human instances in the frame will be fed to the part detection module, after that, the human body patches, human part detection results, and action recognition results will be input to the human part state recognition module.

For the human detector module, we use a common used object detection network, i.e., Cascade-rcnn[1], and the backbone is resnet50 pretrained on COCO dataset. The human body detection module is followed by human body parts detection module which also built based on the Cascade-rcnn network and the backbone is swin-transormer with object365 dataset pretrain weights.

As for action recognition module, the Swin-B [2] is utilized, which is initialized using parameters pre-trained on Something-Something dataset.

The prediction of state recognition of human body parts is a relatively difficult task, and its performance is hard to be improved. We find out that the state of human body parts and human body behavior are highly coupled. Even though the parts have the similar appearance, the part state labels are varying due to the various human behaviors.

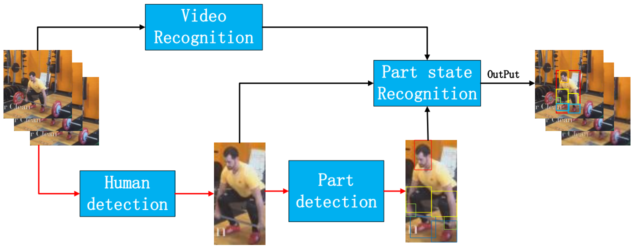

Figure 1. Illustration of the method

Since this dataset does not contain the tracking annotations of each human, the PastaNet[3] is used for our baseline.

Besides, tricks like model ensemble and multi-scale inference also contribute to the final performance.

Detailed algorithm is discussed in Section 2.

2. Approach

In this chapter, we will focus on introducing each module, i.e., human detection, human part detection, human part state recognition, action recognition, as shown in Fig. 1.

2.1. Human Detection Module

This module is proposed to predict the position of the human in each frame of the video. Although the video-based object detection algorithm can accept temporal context 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, so single-frame object detection network is used, i.e., cascade-rcnn[1]. We choose the Resnet50 as its backbone pretrained on COCO dataset. The same training strategy is employed with the original cascade-RCNN, and the mAP of the final human detector on the validation is 93%.

2.2. Even Human Parts Detection Module

The human parts detection module adopts a top-down strategy to decompose multi-person problems into single-person problems. After getting a number of bounding boxes for all instances in a frame, the human body part detector will output part positions of each person. We use

method	Backbone	Dataset size	mAP(%)
Retinanet[7]	Resnet-50	5k	25
GFL[8]	ResNeXt-101	5k	54
GFL	ResNeXt-101	50k	56.6
Cascade-rcnn[1]	Swin-L	50k	57.1

Table 1. The impact of different Object Detection model and dataset size.

Method	Backbone	Pesudo-label	mAP(%)
Cascade-rcnn	Swin-L	×	57.1
Cascade-rcnn	Swin-L	√	56.7

Table 2. The result of use pseudo-label

Augmentation	Action info	RoI Pooling	Multi-Head attention	Top1 (%)
√				72.4
√	√			73.6
√	√	√		74.2
√	√	√	√	74.9

Table 3. Ablation study of our human part state recognition network

Retinanet[7] as our baseline, and the final result is benefited from model architecture, dataset size, pseudo-label generation, and data augment.

2.2.1 Architecture

The performance of our baseline is poor. We tried other state-of-the-art object detection models, such as GFL[8], Cascade-rcnn[1]. The comparison results are shown in Table 1.

2.2.2 Pseudo-label

Though visualizing the data of human body parts, we found that the data contained a certain amount of noise, including part label missing and part location jitter. For the samples with missing labels in some parts, we use the form of generating pseudo-labels to complete the missing label information. Specifically, the GFL model is used to infer about the training dataset, and these part boxes are set positive samples with confidence greater than 0.4, while the prediction confidence between 0.1 and 0.4 will be ignored. The experiment results are shown in Table 2. After adding the pseudo-label information, we found that the effect of the validation is not be improved. Through analyzing the results, we found that parts of the pseudo-label information are inaccurate, which means noise will be introduced when label missing problem is alleviated.

2.2.3 Data Augment

During the training of the human body part detection, we use random Crop and random rotation technical to augment in the training phrase.

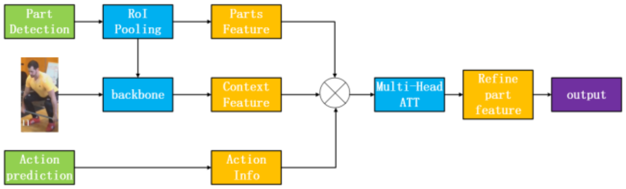

Figure 2. Part state recognition network structure

2.3. Human Part State Recognition

Since the state of human body parts is highly correlated with current human behavior, we also need to consider the following problems:

- How to get robust features for human body parts?
- How to fuse global context information and local features among human body parts?
- How to get accurate information about current human behavior?

Regarding the above issues, we propose the following solutions: (1). use the pre-trained Transformer network as the backbone, and perform RoI Pooling to obtain part features; (2). connate human body feature on each part specific to obtain global context; (3). use Multi-Head Attention to fuse information from body parts; (4). add action category information to the network. The whole structure is showed in Fig. 2.

2.3.1 Backbone

The Transformer architecture has recently shined in image classification and object detection tasks, and its strong feature extraction capabilities can effectively benefit downstream tasks. We use swin-large-224[4] pretrain on Imagenet-21K as the backbone, input a single frame of human body image, and perform RoI pooling operation on its middle layer features to obtain the part features, and the output of the last stage which contain high-level semantic information is used as the current human body feature.

2.3.2 Multi-Head Attention Mechanism

Since Multi-head attention machoism endows the model the ability to jointly attend information from different representation subspaces at different positions[5]. We borrow the idea of Multi-Head Attention in BERT[6], input each body part into the attention network as a patch.

2.3.3 Action Category Information

We find the state of the human body part are highly coupled with human behavior, the human body parts with the same appearance have different states due to different behavior categories. Therefore, how to add action category information into the network becomes particularly essential. We propose to embed the behavior category of the current video into a high-dimensional space, and is fed to the network to help the network output is more accurate.

2.3.4 Data Augmentation

We have selected some data augmentation operations including RandCrop, Part box random jitter and other color space transformations, like random brightness, contrast, etc.

2.3.5 Result

Table 3 show our ablation studies. The impact factors include human part state recognition network, Multi-Head Attention, Action Information. We can see that these factors are contributed to the final results.

2.4. Action Recognition

In Action Recognition task, TimeSformer[9], Video-Swin-Transformer[2] have achieved excellent performance. On the other hand, ActionClip[10] uses multi-modal strategy to effectively improve the performance. We use the Swin-B model from Video Swin Transformer. Since the use of Kinetics pre-train is not allowed, we train the pre-train on something-something dataset by ourselves, which load ImageNet-22K pre-train. Tricks like TTA, model ensemble also contribute to the final performance. Finally, we achieve 93% top-1 accuracy.

3. Experiments

3.1. Implementations Details

Dataset. We only use the competition training set for experiments.

Training. For action recognition, we conduct experiments on a Cluster with A100. For Human part state recognition , due to the large imbalance of human body parts data, resampling strategy is used to balance the various states of each part. The input is scaled to 256x256 and then randomly cropped to 224. Besides, we use AdamW for optimizer and set batchsize to 512 with weight decay 1e-8.

4. Conclusions

In this technical report, we introduced the scheme details used in this challenge. We decoupled the tasks and tried the current excellent algorithm model on each module. Besides, for human body part state recognition, we add action category information and use the Multi-Head Attention mechanism to polish the part feature. Finally, our team achieve 63% score in this challenge and win the first place.

References

[1] Cai Z , Vasconcelos N . Cascade R-CNN: Delving into High Quality Object Detection[J]. 2017.

[2] Liu Z, Ning J, Cao Y, et al. Video swin transformer[J]. arXiv preprint arXiv .2103. 13230,2021.

[3] Li Y L , Xu L , Liu X , et al. PaStaNet: Toward Human Activity Knowledge Engine[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[4] Liu Z , Lin Y , Cao Y , et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J]. 2021.

[5] Attention Is All Your Need.

[6] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[7] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[8] Li X, Wang W, Hu X, et al. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11632-11641.

[9] Bertasius G, Wang H, Torresani L. Is Space-Time Attention All You Need for Video Understanding?[J]. arXiv preprint arXiv:2102.05095, 2021.

[10] Wang M, Xing J, Liu Y. ActionCLIP: A New Paradigm for Video Action Recognition[J]. arXiv preprint arXiv:2109.08472, 2021.