Deeper Action

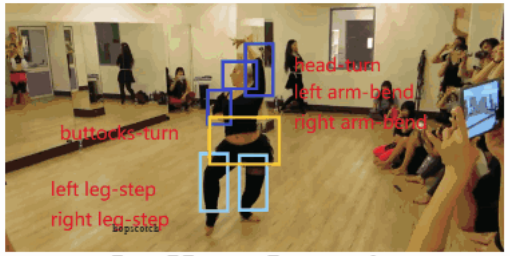
ICCV 2021 Workshop

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

Zheming Yu[1] Lin Li[2] Jietian Guo[3]

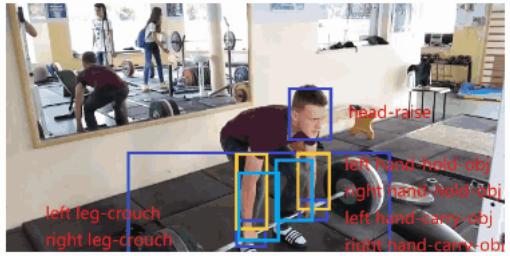
Hikvision Research Institute

- Challenge introduction



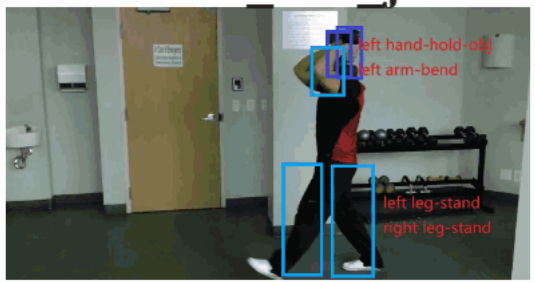
belly_dancing

clean_and_jerk

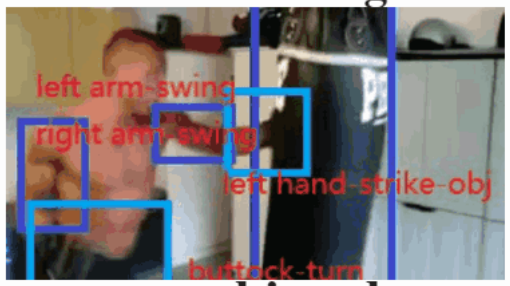


deadlifting

hopscotch

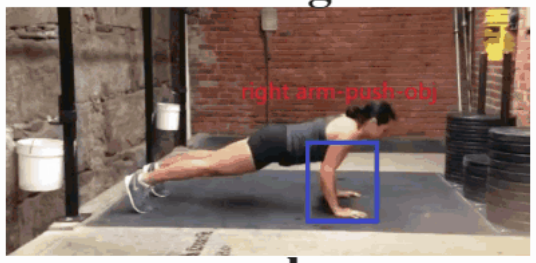


lunge

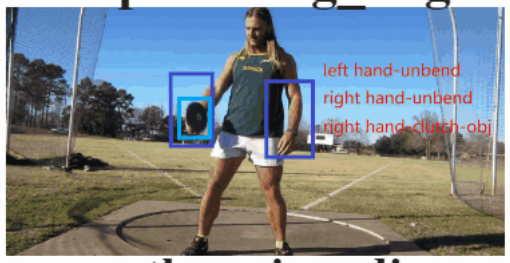


punching_bag

pull_ups



push_up



throwing_discus

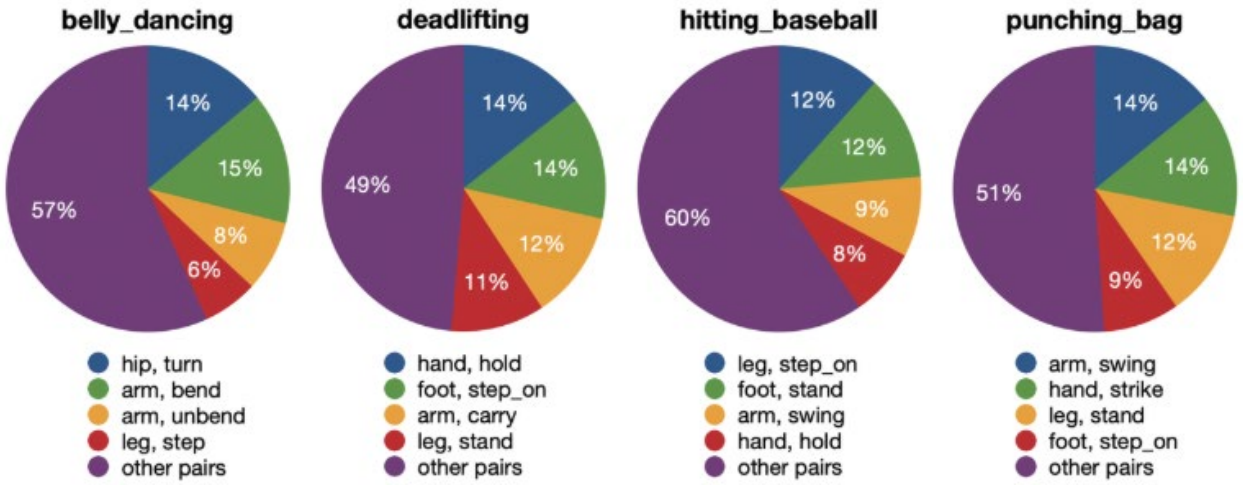
needs to predict:

- human location
- body part location
- part state in the frame level
- human action in the video level

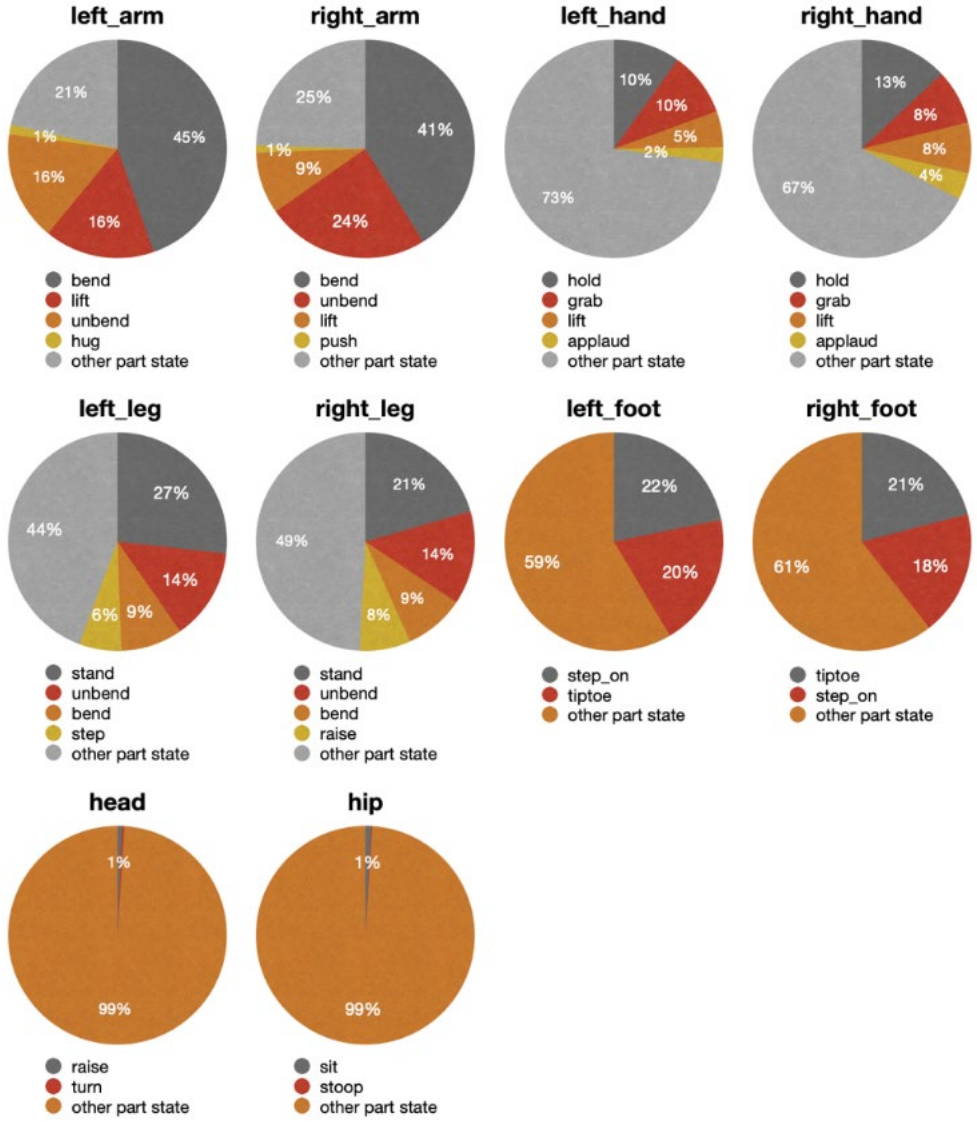
- Dataset Statistics

Kinetics-TPS contains 4741 videos:

1. Bounding boxes of human instances: 1.6 M
2. Bounding boxes of body parts: 7.9 M
3. Part state tags of each annotated part: 7.9 M
4. Bounding boxes and tags of objects: 0.5 M
5. ‘body part, part state’ pairs of four exemplar classes in the dataset

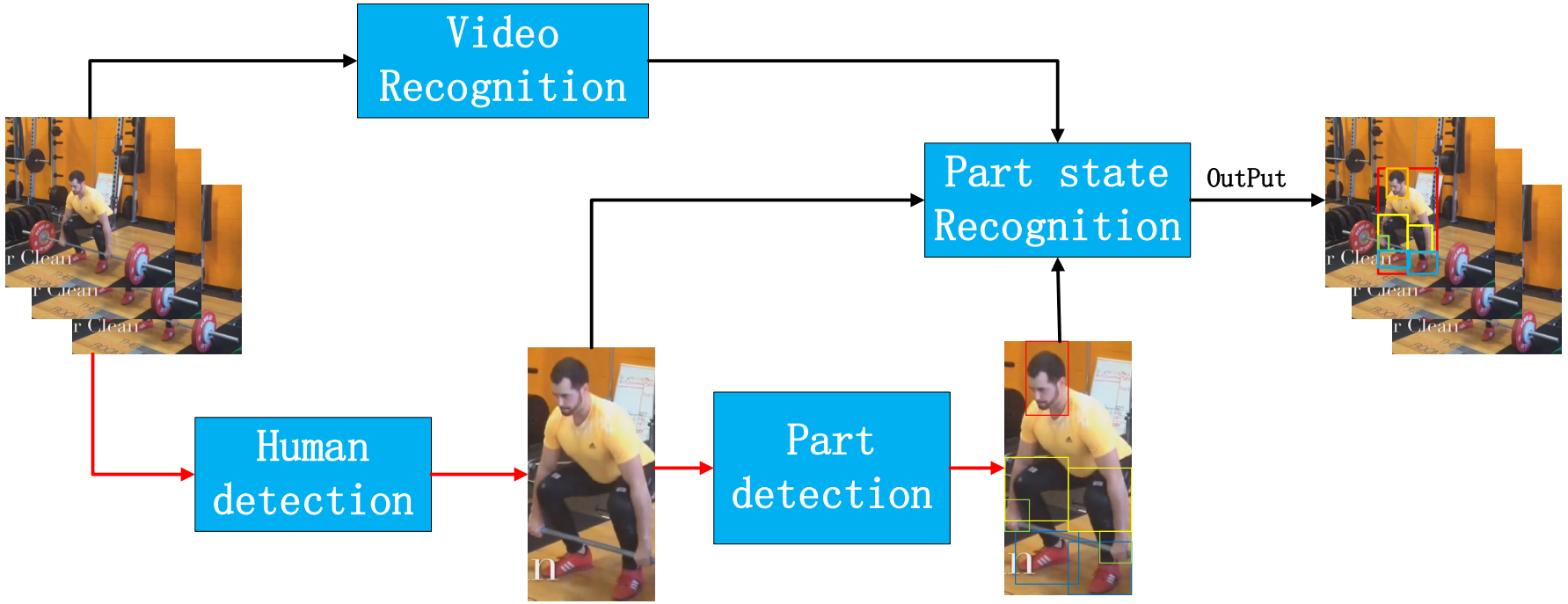


6. Top-5 ‘part state’ tags of each body part in Kinetics-TPS



- Our Method

This Challenge aims at locating the human location, body part location and their part states simultaneously in the frame level. There are four modules to predict the results.



The framework of the proposed method

- Module results

➢ Human detection

Method	Backbone	Pre-train	mAP(%)
Cascade-RCNN	Resnet-50	COCO	93

➢ Action recognition

Method	Pre-train	ACC(%)
Video Swin Transformer	SSv2	93

➢ Human Body Parts Detection

- We adopts a top-down strategy to decompose multi-person problems into single-person problems;

➢ Best result is cascade-RCNN with Swin-L backbone and trained on larger dataset.

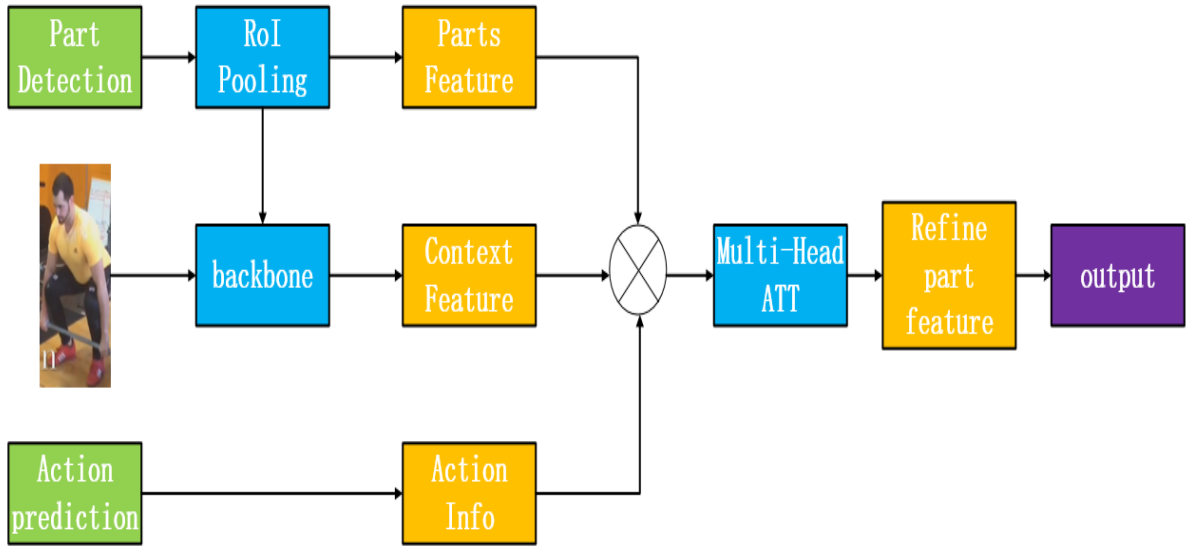
method	Backbone	Dataset size	mAP(%)
Retinanet	Resnet-50	5k	25
GFL	ResNeXt-101	5k	54
GFL	ResNeXt-101	50k	56.6
Cascade-rcnn	Swin-L	50k	57.1

➢ The risk of pseudo-label

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

- Module results

➢ Human Part State Recognition



Human part state recognition structure

- Use the pre-trained Transformer network as the backbone, and perform ROL Pooling to obtain part features;
- Connate human body feature on each part specific to obtain global context;
- Use Multi-Head Attention to fuse information from body parts;
- Add action category information to the network.

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

- Prospects

  - Incorporating middle level information to improve the action recognition performance.
  - Incorporating stable and accurate temporal information to improve the performance of human body part localization and human body part state prediction.
  - Incorporating pre-trained contrastive Language-Video Model, like ActionCLIP, and open concepts.

Thanks