DeeperAction Workshop at ECCV 2022: 2nd place solution for Part-level Action Parsing on Kinetics-TPS Track

Technical Report: Unifying Comprehensive Knowledge into Part-level Action Parsing

Xiaojia Chen[1] Xuanhan Wang[1] Yan Dai[1] Jingkuan Song[12]

[1] Center for Future Media, University of Electronic Science and Technology of China, Chengdu, China

[2]Pengcheng Lab, Shenzhen, China

josonchan1998@163.com, xuanhan.wang@std.uestc.edu.com, yandai1019@gmail.com
jingkuan.song@gmail.com

Abstract

This technical report introduces our solution for Part-level Action Parsing on Kinetics-TPS Track in ECCV DeeperAction Workshop 2022. The proposed method is mainly based on Knowledge Embedded RCNN (KE-RCNN) [11] for part-level action parsing and CSN [5] for video action recognition. Specifically, KE-RCNN is an end-to-end framework and unifies comprehensive knowledge into part-level action parsing, which has three prediction heads for human detection, body part detection, and part state parsing. In the competition, our method achieved a score of 69% on the test set of Kinetics-TPS.

1. Introduction

The Part-level Action Parsing task aims at localizing the human instance and detecting their body part location and part states simultaneously in the frame level. In this report, we use KE-RCNN [11] for part-level action parsing and CSN [5] for video action recognition, which forms the basis for the submission to the Kinetics-TPS Challenge from **CFM-HAG** team.

As demonstrated in [10], the instance-aware body part detection and part state parsing is the bottleneck for this task. Inspired by [6], an intuitive approach is to directly adopt an RCNN-based framework to support part-level action parsing, which is derived from object detection models by applying two new branches for body part detection and part state parsing on part-level region features. However, local-wise part boxes with limited visual clues (i.e., part appearance only) will lead to unsatisfied parsing results, since many part-level states are not only decided by part-self but also relevant to others. To handle the above issue, we argue that not only visual information derived from local-wise part boxes but also relational knowledge representing rich clues of a part are needed.

Motivated by the above analysis, we propose a Knowledge Embedded regional convolution neural network (KE-RCNN), which is a simple yet effective RCNN-based framework for part-level action parsing. It follows an encoder-decoder design pattern that involves two novel components: (1) Implicit Knowledge-based Encoder (IK-En) and (2) Explicit Knowledge-based Decoder (EK-De). Specifically, the IK-En is designed to enhance part-level representation by encoding implicit knowledge about part-part relational contexts into part boxes, where it smartly decides which part-part relations are needed and what contexts to add. After that, the EK-De is proposed to identify the state from the part-level representation with the guidance of prior knowledge about part-state relations, which is derived from statistical priors.

Next, we present the detailed algorithm in the following section.

2. Method

The overall pipeline of the proposed method is shown in Fig. 1. Next, we orderly present the proposed method for human and body part detection, part state parsing, and video action recognition.

2.1. Human and Parts Detection

To detect persons and their body parts from video frames, we adopt an object detection approach based on Faster-RCNN [9], which starts with region proposal generation and then refines each proposal in the RCNN head for predicting persons' locations with their categories. Furthermore, we extend the Faster-RCNN by adding one RCNN branch for instance-level body part detection.

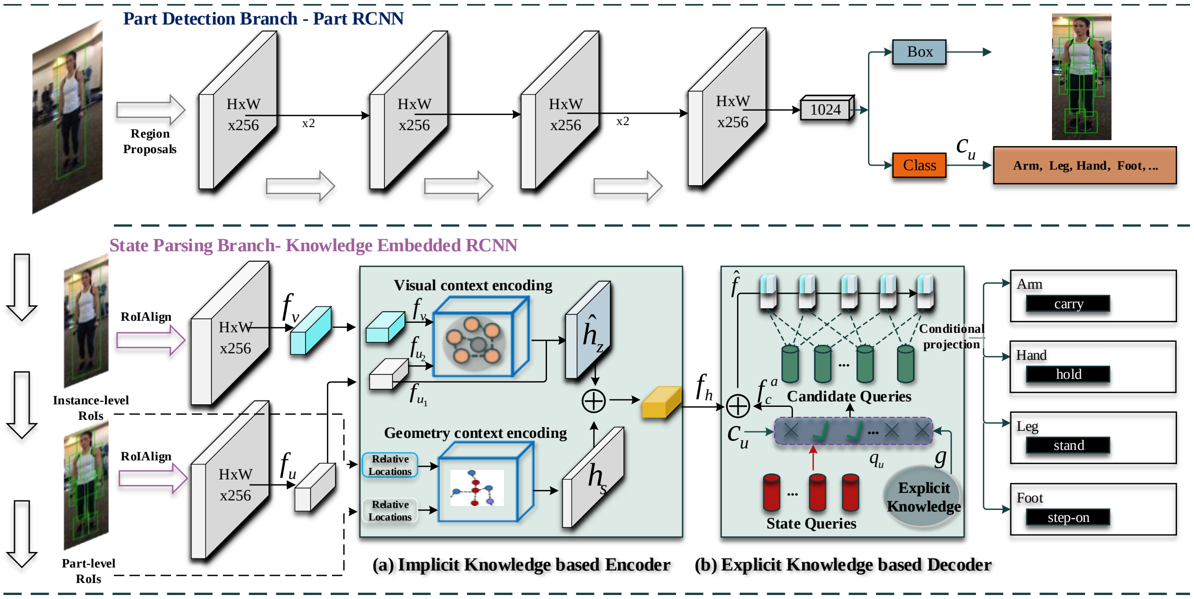

Figure 1. The overview of the proposed pipeline for part-level action parsing.

To implement our proposed detector, we take the Faster-RCNN with ResNet-50 as the basic detection model. In parallel with the person detection branch, a sub-network is built for body part detection, named Part-RCNN. Specifically, Part-RCNN is constructed by four consecutive convolutional layers with 256 channels followed by two linear layers for body part classification and regression. As a result, the proposed detector outputs a set of predictions, including the person bounding box $\{\mathcal{P}_{cls} \in \mathbb{R}^2, \mathcal{P}_{box} \in \mathbb{R}^8\}$ and the instance-aware body part bounding box $\{\mathcal{P}_{pcls} \in \mathbb{R}^N, \mathcal{P}_{pbox} \in \mathbb{R}^{4\times N}\}$, where N indicates the number of part categories.

Learning objectives: To enable the model to perform person detection and body part detection, we design our learning objectives as follows:

$$\begin{aligned}
\ell_{cls} &= Cross\_Entropy(\mathcal{P}_{cls}, \mathcal{P}^*_{cls}) \\
\ell_{box} &= SmoothL1(\mathcal{P}_{box}, \mathcal{P}^*_{box}) \\
\ell_{pcls} &= BCE(\mathcal{P}_{pcls}, \mathcal{P}^*_{pcls}) \\
\ell_{pbox} &= SmoothL1(\mathcal{P}_{pbox}, \mathcal{P}^*_{pbox}) \\
\ell_{det} &= \ell_{cls} + \ell_{box} + \ell_{pcls} + \ell_{pbox}
\end{aligned} \tag{1}$$

where $\mathcal{P}^*_{cls}$,$\mathcal{P}^*_{box}$,$\mathcal{P}^*_{pcls}$, and $\mathcal{P}^*_{pbox}$ is the corresponding ground truth. And BCE indicates the binary cross entropy loss.

2.2. Part State Parsing

In this section, we introduce our proposed method for part-level state parsing. As shown in Fig 1, our method decouples part state parsing by adding an independent branch, named KE-RCNN. To parse the state for each detected body part, our KE-RCNN first utilizes an Implicit Knowledge-based Encoder (IK-En) to enhance the part feature by incorporating part-part relational contexts. Then, under the guidance of explicit knowledge about part-state relations, candidate state queries that are relevant to the part are provided. Next, conditioning on candidate attribute queries, the enhanced part feature is further projected to state embeddings by applying an Explicit Knowledge-based Decoder (EK-De). Finally, a calculated similarity between generated part state embeddings and state classifier is used to recognize the state of the part.

2.2.1 Implicit Knowledge Encoder

In the following, we discuss the two major components (i.e., visual and geometry context encoding) of our Implicit Knowledge based Encoder (IK-En).

Visual context encoding: Given the detected person bounding boxes and their parts' bounding boxes, we firstly extracted their features by RoIAlign [6], which are denoted as f_v and f_u, respectively. Following [4], we start by evenly splitting part features f_u into two subsets, respectively denoted as f_{u1} and f_{u2}. f_{u1} is used to represent part visual information and f_{u2} is used to encode visual contexts by interacting with f_v. Specifically, we treat f_{u2} as the query, and let f_v be the key and value, then perform cross attention to obtain the part context feature h_z. Finally, we linearly fuse the part visual feature f_{u1} and the part context feature h_z to attain the enhanced part representation $\hat{h_z}$.

Geometry context encoding: In addition to visual contextual relations, we represent geometry context of the part through Eq. 2:

$$h_s = W_g(\tfrac{x_u - x_v}{w_u}, \tfrac{y_u - y_v}{h_u}, log(\tfrac{w_v}{w_u}), log(\tfrac{h_v}{h_u}))^T, \qquad (2)$$

where $\langle x_u, y_u, w_u, h_u \rangle$ are coordinates and scales extracted from part region and $\langle x_v, y_v, w_v, h_v \rangle$ are counterpart from person region. $W_g \in \mathbb{R}^{D \times 4}$ is a linear matrix that maps the relative geometry context into a high dimensional vector h_s.

After that, a part representation f_h with implicit knowledge (i.e., visual relation and geometry relation) is obtained by simply fusing h_s and $\widehat{h_z}$.

2.2.2 Explicit Knowledge Decoder

In this section, we introduce how to parse the state of the part by our Explicit Knowledge Embedded Decoder (EK-De). We start by initializing the state queries $q_u \in \mathbb{R}^{D \times C}$ and use explicit knowledge to filter state queries, where C is the number of state categories and D is the number of channels. The filter process can denote as Eq. 3:

$$\begin{aligned} c_u^* &= {c_u}^T g, \\ q_u^{\prime} &= q_u \circ c_u^*, \\ \hat{q}_u &= \theta(q_u^{\prime} | c_u^*), \end{aligned} \qquad (3)$$

where $c_u \in \mathbb{R}^{N \times 1}$ indicates the probability of each part class, $c_u^* \in \mathbb{R}^{1 \times C}$ denotes a weighing vector that decides which state is the candidate. And $g \in \mathbb{R}^{N \times C}$ is the explicit knowledge by calculating a frequent statistics matrix from the occurrence among all part-state pairs. $\theta(\cdot|\cdot)$ denotes a filtering function that outputs $\hat{C}$ candidate state queries. $\hat{q}_u \in \mathbb{R}^{\hat{C} \times D}$ conditioning on c_u^*, where each state query is selected as a candidate if its' corresponding score in c_u^* is higher than a predefined threshold value (*e.g.*, 0). After that, we perform the multi-head cross attention between filtered state queries $\hat{q}_u$ and part representation f_h to obtain the decoded state embeddings $f \in \mathbb{R}^{\hat{C} \times D}$, where f_h is the key and value in cross attention. Finally, the state of the part is parsed through a similarity matrix calculated between f and the state classifier $\hat{W}_s \in \mathbb{R}^{\hat{C} \times D}$, as formalized in Eq 4.

$$\mathcal{O} \;\; = \mathcal{P}(\sum_{i=1}^{D} \hat{W}_s^i \circ f^i), \qquad (4)$$

where $\mathcal{P}(\cdot)$ is the softmax nonlinear function. $\mathcal{O} \in \mathbb{R}^{\hat{C}}$ is the state categorical distribution.

2.3. Video-level Action Recognition

Following [3], we use the CSN [5] model pre-trained on IG-65M dataset and then finetune on the Kinetics-TPS training set. As a result, we achieve around 97% top-1 accuracy on the testing set. Furthermore, we finetune the Video Swin Transformer [8] model pre-trained on the something-something dataset and then perform a model ensemble between the CSN model and the Video Swin Transformer model. Finally, we achieve 98% top-1 accuracy on the testing set.

3. Experiments

3.1. Experimental Settings

We train our models on the training set of the Kinetics-TPS dataset. There are 3809 annotated videos in the training set. In the test phase, all models are tested on the official server[1]. Our models are implemented based on mmdetection[2] on an Ubuntu server with eight Tesla V100 graphic cards. We adopt FPN as the backbone model and use Adam solver to train for 12 epochs. The learning rate is 1e-4 and decreases by 10 at the 8-th epoch.

3.2. Main Result

Our experimental results are summarized in Tab 1. We firstly simply extend Faster-RCNN by adding two branches for body part detection and part state parsing. And we finetune TimeSformer [1] model as our video classification result, which achieves 85% top-1 accuracy on the testing set. When replacing the baseline with the KE-RCNN model, we find comprehensive knowledge is critical to part-level action parsing, where it improves the baseline model by 4.7%. After that, we replace the ResNet-50 backbone with the Swin-B [7] backbone, which improves the score by 4.5%. Furthermore, we use the ensembled video classification model as described in subsection 2.3, which achieves a higher score of 66.6%. Finally, with the help of testing time augmentation (TTA) and parsing model ensemble, we attain the final score of 69.1%.

3.3. Ablation Study

In this section, we investigate the effect of IK-En and EK-De. Note that we randomly pick 30% of the training set as the minival set for the ablation study, resulting in 2686 videos for training and 1123 videos for validation. And we use the same video classification result across all the models. The experimental results are reported in Tab. 2. From the results, we have the following findings:1) The KE-RCNN with IK-En shows better performance than that of KE-RCNN with EK-De. 2) Jointly applying IK-En and EK-De brings the best results, suggesting that each component is complementary to each other for state parsing. We refer readers to the paper of KE-RCNN [11] for more details.

4. Conclusion

In this report, we present our method for part-level action parsing. We unify comprehensive knowledge into part-level

[1] `https://codalab.lisn.upsaclay.fr/competitions/4392`

Table 1. Ablation results of different submissions on the Kinetics-TPS testing set.

Baseline	KE-RCNN	Better backbone	Video model ensemble	TTA	Parsing model ensemble	Acc_p
✓						49.5%
✓	✓					54.2%
✓	✓	✓				58.7%
✓	✓	✓	✓			66.6%
✓	✓	✓	✓	✓		67.2%
✓	✓	✓	✓	✓	✓	69.1%

Table 2. Component ablation studies on Kinetics-TPS. Investigating the effect of proposed modules.

Parsing Branch	IK-En	EK-De	Acc_p
Standard RCNN	-	-	49.1
KE-RCNN	✓		52.2
KE-RCNN		✓	51.7
KE-RCNN	✓	✓	**53.5**

action parsing by implicit and explicit knowledge. By building Implicit Knowledge based Encoder (IK-En), we enhance part representations by incorporating visual contexts as well as geometry contexts. Then Explicit Knowledge-based Decoder (EK-De) is proposed to identify the state of a part by human prior knowledge. Extensive experiments on the Kinetics-TPS dataset demonstrate the effectiveness of the proposed method, and it obtain 69.1% Acc^p score in the 2022 Kinetics-TPS Challenge.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Xiaodong Chen, Xinchen Liu, Kun Liu, Wu Liu, and Tao Mei. A baseline framework for part-level action parsing and action recognition. *arXiv preprint arXiv:2110.03368*, 2021.

[4] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[5] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[10] Xuanhan Wang, Xiaojia Chen, Lianli Gao, Lechao Chen, and Jingkuan Song. Technical report: Disentangled action parsing networks for accurate part-level action parsing. *arXiv preprint arXiv:2111.03225*, 2021.

[11] Xuanhan Wang, Jingkuan Song, Xiaojia Chen, Lechao Cheng, Lianli Gao, and Heng Tao Shen. Ke-rcnn: unifying knowledge based reasoning into part-level attribute parsing. *arXiv preprint arXiv:2206.10146*, 2022.