Hand Description in Egocentric Vision for Future Action Recognition

-

194.077 - Applied Deep Learning

Wiktor Mucha

October 2022

1 Introduction

Short description Egocentric vision is a computer vision domain where the world is seen from a perspective of a person wearing a camera on her/his body. The image collected in this manner is similar to the view visible to the camera wearer. This perspective is interesting for understanding human actions and information from camera user surroundings which are valuable for this person's behaviour understanding. The majority of actions performed by humans involve interaction with objects by using the hands. The aim of this project is to to create a deep learning algorithm to detect the key points of the hands for skeleton description.

Motivation State-of-the-art action recognition models in videos for egocentric data involve high computation caused by a need for feature extraction in every frame in a sequence and later full sequence understanding. I am working on my PhD in the field of egocentric vision and came to the conclusion to create a model for action recognition in which the learning process is accelerated by employing in a sequence model the hand positions and the position of interacting objects instead of extracting features in a traditional way.

Methodology In the literature, there are existing examples of similar work in this field. Dousty et al. [1] models hands to detect tenodesis grasp. First, hands are detected using YOLOv2[2] including differentiating between left and right. The next step is pose estimation using OpenPose[3]. The algorithm extracts feature maps using VGG-19 backbone. Further, a 2-branch multistage Convolutional Neural Network (CNN) simultaneously predicts the locations of body parts outputting 2D maps with position confidence.

Cho et al.[4] performs action recognition basing on hand key-points position. First, to estimate these positions they use an algorithm named DETR[5] with a

modified head to predict key-points. DETR model combines CNN with a transformer. CNN plays as a backbone learning 2D representation of an input image that is passed to the encoder with positional embeddings. Learned positional embeddings are the decoder's input and finally, the feed-forward neural network classifies them as an object or not and outputs bounding boxes.

These two examples show that this problem is based on object detection where hand key-points are predicted. These key-points can be hands joints and fingers to create a 2D skeleton representation.

2 Description of datasets

This section contains datasets for the described task of hand key-point finding with a short description for each one. Unfortunately, for this moment I could not find this for the egocentric domain. To test this in the egocentric environment it is very likely that creating annotations will be necessary.

2.1 FreiHAND Dataset [1]

A dataset for hand pose and shape estimation from a single colour image, which can serve both as a training and benchmarking dataset for deep learning algorithms. Contains 130240 training and 3960 evaluation samples with RGB images and hand segmentation masks.

2.2 COCO-WholeBody-Hand [2]

This dataset is a version of COCO dataset for pose estimation. It includes key-points for the whole body, but they are divided between categories and the left and right hands create a separate category that can be distinguished from them.

2.3 CMU Panoptic HandDB [3]

This dataset included RGB images with keypoint annotations for 1912 training samples and 846 testing ones. Additionally, there are synthetic data with 14261 annotations in total.

2.4 Rendered Handpose Dataset [4]

The dataset contains of 41258 training and 2728 testing samples. There are RGB images (320x320 pixels), depth maps (320x320 pixels), and segmentation

[1] FreiHAND Dataset `https://lmb.informatik.uni-freiburg.de/resources/datasets/FreihandDataset.en.html`

[2] COCO-WholeBody-Hand Dataset `https://github.com/jin-s13/COCO-WholeBody/`

[3] CMU Panoptic HandDB `http://domedb.perception.cs.cmu.edu/handdb.html`

[4] Rendered Handpose Dataset `https://lmb.informatik.uni-freiburg.de/resources/datasets/RenderedHandposeDataset.en.html`

masks. each hand is described with 21 key-points.

3 Work structure and schedule

Estimated schedule of work:

- Downloading datasets, and familiarizing myself with them. Creating data loaders - 2 days
- Network planning
- Network building - 3 days
- Training, fine-tuning, and improvements - 7 days
- Creating annotations for egocentric datasets - 2 days
- Application to present the work - 1 day
- Creating final presentation - 1 day

References

[1] M. Dousty and J. Zariffa, "Tenodesis grasp detection in egocentric video," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1463–1470, 2020.

[2] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[4] H. Cho and S. Baek, "Transformer-based action recognition in hand-object interacting scenarios," *arXiv preprint arXiv:2210.11387*, 2022.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.