Action Recognition Thesis - WIP

Nicolas Fleece
University of Ottawa
75 Laurier Ave. E, Ottawa, ON K1N 6N5
nflee092@uottawa.ca

Abstract

The ABSTRACT is to be in fully justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

Paragraphs: 1: General overview, action recognition 2: The approach of the paper 3: The rest of the paper

Human action recognition is a very difficult task and is a constantly evolving topic of research in the computer vision community. In the real world, human action recognition has many issues, not limited to things such as occlusion, cluttered and dynamic backgrounds, camera motion, and multiview variations [9]. Existing classical models [6] are vulnerable to these variables in the real world. The extraction of different features has tried to mitigate these background influences, such as rgb flow [13], however these can still be influenced by motion in the background. The emergence of lightweight, high quality pose estimation [5] has allowed for accurate skeleton data from nearly every video in the wild, and utilizing this skeleton data, the background factors can be mitigated.

This paper aims to mitigate these background factors, while providing a model that provides high-quality action recognition prediction at near state of the art performance. This involves constructing an intermediate representation using exclusively 2D pose data. With this representation, we aim to remove as much background influence as possible, thus our representation is independent of global position, or global rotation of the subject performing the action. The architecture of this representation was greatly inspired by [11], however their implementation suffers from largely relying on the global position of the person, which can reduce in the wild accuracy outside of curated datasets.

2. Related Work

Typical examples of action recognition [6] have used complex convolutional neural networks (CNNs) on the RGB frames, pulling features such as rgb flow [13] in order to enhance results. This method has proved to work well, however these models often require large amounts of GPU memory and are required to be run on high end hardware, which is a potential issue when applied to real-world scenarios where the necessary hardware may not be available. In addition, background noise is much more difficult to filter out and generalizing to different environments is difficult.

A subset of action recognition models utilize skeleton-based action recognition. These models aim to utilize the positions of joints/bones in the model in order to filter out background data, and allow the model to focus on the actual person rather than background data. Typically these involve simpler CNNs that allow for faster computation on lower quality hardware. Many different approaches are used to achieve this action recognition. [8], [16] utilize processed joint heatmap images and simple 2D-CNNs. [11] [17] use intermediate representations to construct custom image representations that can be easily processed by simple CNNs. RNNs and LSTMs [3] [15] [1] [10] and Transformers [2] [18] have been used as well to obtain good results with skeleton data.

Almost always, these skeleton based models utilize 2D pose data. There have been some examples of extracting 3D pose from depth cameras [4] as well as estimating 3D pose from 2D pose [7] and only the RGB frames [14]. These techniques have been used in the past for human action recognition [12], however 2D pose estimation is a much easier task and the models that have been used are generally much higher quality, and therefore is more reliable for the task of human action recognition.

3. Method

4. Experiment

5. Conclusion

6. Acknowledgement

References

[1] Federico Angelini, Zeyu Fu, Yang Long, Ling Shao, and Syed Mohsen Naqvi. 2d pose-based real-time human action recognition with occlusion-handling. *IEEE Transactions on Multimedia*, 22(6):1433–1446, 2020. 1

[2] Aizada Askar, Min-Ho Lee, Thien Huynh-The, and Nguyen Anh Tu. 2d skeleton-based action recognition using action-snippets and sequential deep learning. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2372–2377, 2022. 1

[3] Danilo Avola, Marco Cascio, Luigi Cinque, Gian Luca Foresti, Cristiano Massaroni, and Emanuele Rodolà. 2-d skeleton-based action recognition via two-branch stacked lstm-rnns. *IEEE Transactions on Multimedia*, 22(10):2481–2496, 2020. 1

[4] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 71–98, 2013. 1

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[7] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 1

[8] Vasileios Choutas, Philippe Weinzaepfel, Jerome Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. pages 7024–7033, 06 2018. 1

[9] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020. 1

[10] Xinghao Jiang, Ke Xu, and Tanfeng Sun. Action recognition scheme based on skeleton representation with ds-lstm network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2129–2140, 2020. 1

[11] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Simple yet efficient real-time pose-based action recognition. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 581–588, 2019. 1

[12] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018. 1

[13] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9945–9953, 2019. 1

[14] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2500–2509, 2017. 1

[15] Shenghua Wei, Yonghong Song, and Yuanlin Zhang. Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 91–95, 2017. 1

[16] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7914–7923, 2019. 1

[17] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM Multimedia Asia*, MMAsia '19, New York, NY, USA, 2020. Association for Computing Machinery. 1

[18] Yerassyl Zhalgasbayev and Nguyen Anh Tu. Two-branch stacked transformer for 2d skeleton-based action recognition. In *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–4. IEEE, 2023. 1