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

CNN's have been widely used for many years on the problem of human action recognition. They have shown to be capable of detecting actions performed by people in simple and complex settings. This quite often involves the use of 3D-Convolutions, which operate on multiple frames of video at a time. Complex models that often use several GPUs are constantly being developed and tested on many different datasets.

1.1. Pose-Based Action Recognition

Pose involves extracting the skeleton of the person and using this data over multiple frames of a video to classify an action. Pose is a common addition used in action recognition as it relates most to how humans view actions and the movement of different bones.

1.1.1 Intermediate Representations

The approach of the majority of this paper involves creating intermediate representations for pose data over multiple frames. This typically has the aim of creating some kind of image that represents either the motion of the persons bones and/or joints through the image at different points in the video. These images can then be used either by the model independently or added to traditional two-stream architectures.

The advantage of these types of representations is that the model can be a small neural network (often CNN's) that can be trained end-to-end very quickly and with little memory. This quite often allows for real-time evaluation and in some cases mobile-capable models.

1.2. Person-Based Action Recognition

2. Related Work

Typical examples of action recognition REFERENCES HERE have used complex convolutional neural networks (CNNs) on the RGB frames, pulling features such as flow 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. [5], [11] utilize processed joint heatmap images and simple 2D-CNNs. [7] [12] use intermediate representations to construct custom image representations that can be easily processed by simple CNNs. RNNs and LSTMs have been used as well to great effect utilizing skeleton data [2] [10] [1] [6].

Almost always, these skeleton based models utilize 2D pose data. There have been some examples of extracting 3D pose from depth cameras [3] as well as estimating 3D pose from 2D pose [4] and only the RGB frames [9]. These techniques have been used in the past for human action recognition [8], 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] 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

[3] 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

[4] 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

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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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