UNIVERSITY OF OTTAWA

DOCTORAL THESIS

Efficient, Movement-Based Skeleton Action Recognition

Author:
Nicolas FLEECE

Supervisor:
Dr. Robert LAGANIÈRE

A thesis submitted in fulfillment of the requirements
for the degree of Master of Computer Science, Specialization in Applied AI

in the

VIVA Research Lab
School of Electrical Engineering and Computer Science

June 19, 2023

UNIVERSITY OF OTTAWA

Abstract

Faculty of Engineering

School of Electrical Engineering and Computer Science

Master of Computer Science, Specialization in Applied AI

Efficient, Movement-Based Skeleton Action Recognition

by Nicolas FLEECE

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

Acknowledgements

The acknowledgments and the people to thank go here, don’t forget to include your project advisor…

Contents

List of Figures

List of Tables

List of Abbreviations

LAH	**L**ist **A**bbreviations **H**ere
WSF	**W**hat (it) **S**tands **F**or

Chapter 1

Introduction

People interact with their environment in unique and nuanced ways, and throughout our lives, humans have learned to identify and categorize the different actions that we perform.

1.1 Action Recognition

For humans, the problem of Human Action Recognition is rather simple. We use past experiences throughout childhood and adult life to be able to pick out the various ways a person moves, and translate that into a familiar action that we have seen before. Combine that with objects that a person may be interacting with, and humans are remarkably good at discerning what actions other humans are involved in. However, as is with most things in the domain of computer vision, this ability does not translate well into the realm of artificial intelligence. The slightly different ways that people may perform these tasks add a layer of complexity that is difficult for a model to overcome.

1.2 Applications

Security

Health Care

Video Summarization

1.2.1 Ethical Issues

As with most applications of artificial intelligence, computer vision cannot be researched and discussed without taking into account the ethical issues that surround it. With AI being such a quickly evolving space, it is crucial that any researchers be aware of these issues. In this section, I will focus particularly on how it may affect action recognition, touching on other areas of AI in general to further illustrate my points.

Privacy

Accountability

Bias is one of the most common ethical issues in AI that can appear. In artificial intelligence, and computing in general, the principle of "garbage in, garbage out", is a common one to illustrate that if the inputs into a model are not of high quality, the outputs will not be of high quality. This can often be the case in datasets where say a group of people are not accurately represented, and while the model itself is not discriminatory, it will follow the data it has been given. Take the example previously discussed of airport security. Airport security has been scrutinized in the past for singling out individuals of particular races or who look a particular way. This may mean that if a model is being constructed that searches for people who may be flagged later in security, the majority of positive flags that were screened further would be of this group of people. The resulting dataset that is constructed would be biased against this group of people, therefore resulting in a model similarly biased. This type of issue has been shown in many different areas, another example being speech recognition models used by voice assistants not being able to recognize particular accents as they were not represented in the original dataset. These kinds of reasons are why it is crucial for researchers to be aware of and study their datasets when it comes to human data to avoid these biases and ensure that their data is well balanced.

Transparency

1.3 Challenges

Human action recognition is a very difficult task that comes with many issues, some of which continue to be major challenges moving forward with very complex modern models.

Background

Camera Movement

1.4 Problem Definition

1.5 Thesis Structure

Chapter 2

Convolutional Neural Networks

2.1 Structure

2.2 Kernel

2.2.1 3 Dimensional Convolutions

2.3 Classic Architectures

2.3.1 AlexNet

2.3.2 VGG-16

2.4 Modern Architectures

2.4.1 ResNet

2.4.2 Residual Attention Network

2.5 LSTMs

Chapter 3

Literature Review

3.1 Image Classification

3.2 Optical Flow

3.3 CNN Based Models

Naturally, the first approach to feeding video data into models is to process the raw RGB frames. The RGB frame data quite often

3.3.1 CNN + LSTM Models

The success of CNN's in the world of image classification makes the move to apply the same type of logic towards action recognition and the larger domain of video processing. Since a video broken down is just individual frames, the logic follows that we would be able to extract features from individual frames and combine these features to produce a classification outcome.

In very classic models, this is a very simple process. The individual frames are passed through the CNN model, producing a feature map for each frame. These feature maps are then simply pooled and passed into dense layers which produces an output. While very simple and fast, this model completely ignores any temporal activity, meaning that the model cannot determine how a person moves throughout a video from one frame to another. This would make differentiating some reversible actions such as running forwards vs running backwards.

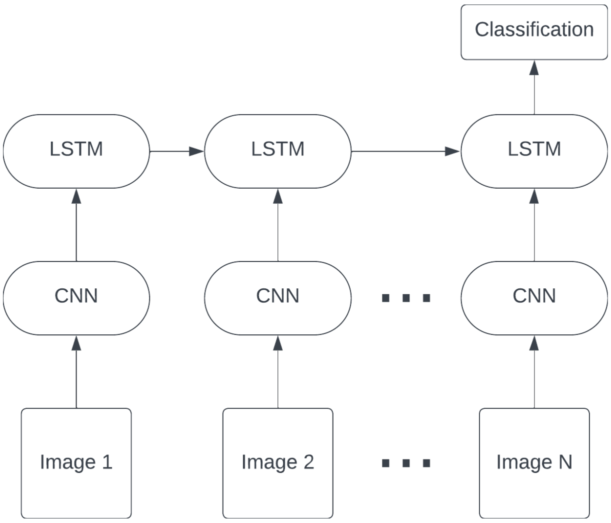

FIGURE 3.1: An example structure of a simple CNN-LSTM based model, each individual frame being individually fed into the CNN, and then passed to a LSTM.

Figure 3.1 shows the typical modern structure for this solution. After the features are extracted from each of the 2 dimensional CNN, they are passed through a LSTM. The goal of this LSTM module is to carry features from one frame to another.

The advantages of this model are that is is very lightweight and all of the individual parts are already well studied and efficient. This also means that the models are very lightweight, and relatively simple in comparison to more complex techniques.

The disadvantages of the model are also rooted in it's simplicity. The result of processing each image independently means that the interactions between frames is not very well represented. While the model is able to represent individual frame features very well, due to the fact that the feature maps are passed through the LSTM, classes that require specific movement from one frame to another are difficult to represent using this structure. Constructing these individual feature maps can also fall victim to background interference, meaning that a movement in the camera, or change in background could impact in a way that detracts from the main subject of the action more with this model than the other approaches discussed later in this chapter.

Long-term Recurrent Convolutional Networks [1], is a model constructed using this methodology. In the paper, they use the notation that each frame, x_i, is fed into the CNN in order to construct a fixed-length feature representation, $\phi_v(x_i)$. This is then passed into the recurrent sequence learning model. This is where the model

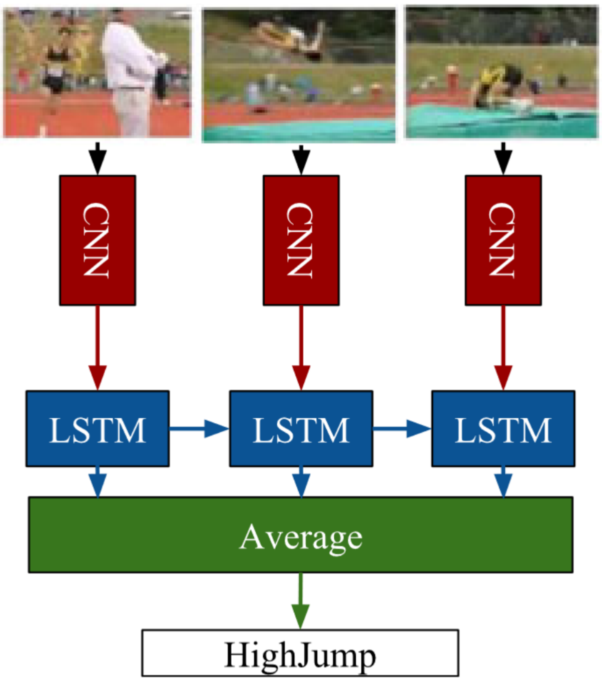

FIGURE 3.2: Action recognition structure for the LRCN model. [1]

differs from the previous example provided. In the LRCN model, the LSTM outputs at each frame are averaged to get the output class, rather than taking the last output. This removes any bias the model may have towards the later frames in long videos. In addition to RGB frames, this model additonally uses the optical flow feature, which easily adapts to this structure, replacing the RGB frames in figure 3.2. The LSTM structure is taken from [2], which is a structure devised from the original LSTM model, as we discussed in section 2.5. The CNN's, represented in the paper as ϕ, is described as a hybrid of the CaffeNet [3] (a variant of the AlexNet [4] model discussed in section 2.3.1) and the Zeiler and Fergus [5] models, which has been pre-trained on a large dataset.

Beyond Short Snippets: Deep Networks for Video Classification [6], is another approach to this structure, which explores a more complex deep-LSTM based module, as well as more classical feature pooling. Similarly to the previously discussed model [1], the model utilizes a combination of two popular CNN models, AlexNet [4] and GoogLeNet [7]. The paper did explore many more classical feature pooling architectures, and were proven to have good results, however these techniques were outperformed by the LSTM model. The paper utilized a deep LSTM architecture for the feature aggregation step, shown in figure 3.4, which further adds to it's

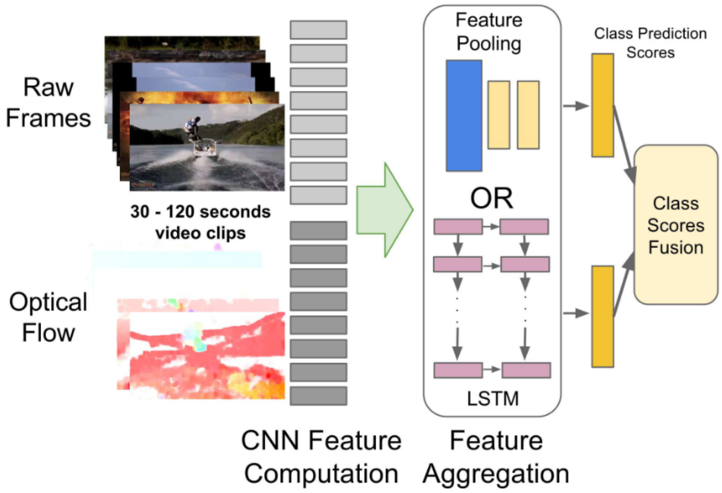

FIGURE 3.3: Overview of the Beyond Short Snippets: Deep Networks for Video Classification model [6]

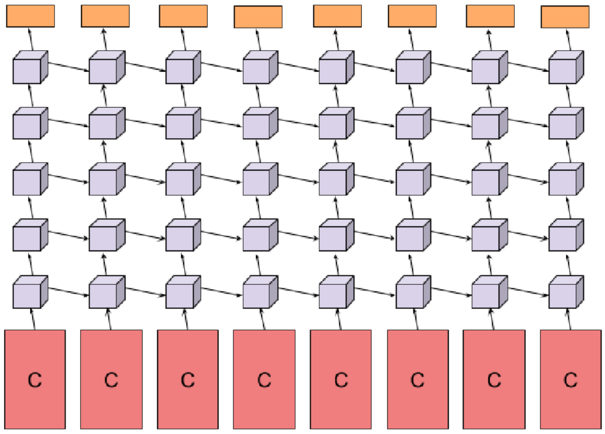

FIGURE 3.4: Deep LSTM architecture utilized by [6] in the feature aggregation step as shown in figure 3.3.

complexity, moving it above the CNN-LSTM architectures described previously.

3.3.2 3D CNN Models

The natural next step for action recognition is to move to 3D CNN models.

3.4 Modern Methods

3.5 Datasets

3.6 Skeleton-Based Action Recognition

3.6.1 Pose Detection

3.6.2 Intermediate Representations

Appendix A

Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.

Bibliography

[1] J. Donahue, L. A. Hendricks, M. Rohrbach, *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017. DOI: 10.1109/TPAMI.2016.2599174.

[2] W. Zaremba and I. Sutskever, *Learning to execute*, 2015. arXiv: 1410.4615 [cs.NE].

[3] Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural networks see the world — a survey of convolutional neural network visualization methods," *Mathematical Foundations of Computing*, vol. 1, pp. 149–180, Jan. 2018. DOI: 10.3934/mfc.2018008.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[5] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, 2013. arXiv: 1311.2901 [cs.CV].

[6] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702. DOI: 10.1109/CVPR.2015.7299101.

[7] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.