Mike Stanton

Explainable Action Recognition using Skeletal Data

B.Sc. Computer Science

19th March 2021

Declaration of Originality

"I certify that the material contained in this dissertation is my own work and does not contain unreferenced or unacknowledged material. I also warrant that the above statement applies to the implementation of the project and all associated documentation. Regarding the electronically submitted work, I consent to this being stored electronically and copied for assessment purposes, including the School's use of plagiarism detection systems in order to check the integrity of assessed work.
I agree to my dissertation being placed in the public domain, with my name explicitly included as the author of the work."

Name: Mike Stanton

Date: 19/03/2021

Abstract

The introduction of large-scale datasets has allowed for recent developments in skeleton-based action recognition. Significant advancements in classification accuracy have been driven by a focus in Graph Convolutional Networks (GCN). State of the art (SOTA) models have become exceedingly over-parameterized, leading to very complicated and sophisticated solutions. This has obstructed development into the field of explainable action recognition. In this work, a lightweight and efficient method based on Convolutional Neural Network (CNN) is proposed, which will implement a two-stream input to extract features from both the skeleton and skeleton motion. In addition to this, the study outlines and designs an explanation module for the presented network.

Table of Contents

Chapter 1: Introduction

1.1 Problem Statements

Action recognition and detection prove to be a fundamental and challenging task in the study of computer vision. Its applications include advance intelligent surveillance systems, human-robotic augmentation, and human-computer interaction. The skeleton provides an excellent representation of human position and movements. Skeleton data is more robust against background noise and provides detailed information on high-level human action features. Compared with more commonly used data in the realm of the problem, i.e., RGB videos and depth map sequences, skeleton data is extremely small, enabling the design of lightweight and compact models.

This paper is conducted in two parts: firstly, addressing the problem of skeleton-based human action recognition. Then, secondly, designing a system that can provide explanations for the classification of actions.

The current most successful work in action recognition using skeleton data implements graph-based networks [Liu et al., 2020; Shi et al., 2020; Cheng et al., 2020], with many major advancements in the area having been developed in the past two years. In addition to this, multiple networks have been designed with two-stream input based upon earlier work into CNNs for action recognition [Simonyan and Zisserman, 2014], splitting the input into joints and bones [Cheng et al., 2020], as well as spatial-temporal splits [Liang et al., 2019]. The models mentioned above' commonality is the necessity for pretraining due to a large number of both trainable and non-trainable parameters within the network. The large size of these models also means a significant amount of processing power is required to utilise the network in acceptable time frames.

The decision process in a neural network is inherently a black box one. Due to this, recent focus has been directed at creating a system that can explain, with reason, its decision making in a classification process in order to verify its reliability as well as allowing researchers to identify problems in a model. A recent study found success in building a system that classified videos and provided explanations as to why a given sample is predicted to one class but not another [Kanehira et al., 2019]. Another study focussed on building a framework for explainable classification within video recognition tasks [Hiley et al., 2019]. These studies were able to demonstrate effective methodology for giving multimodal explanations while working with video input and suggest potential methodology to overcome common pitfalls in designing such a system for action recognition which is utilised in this study.

1.2 Project Aims

The main proposed contributions of this project can be summarised as follows:

- Design and implement a CNN model for learning global co-occurrences from skeleton data to accurately classify actions from the NTU RGB+D 120 dataset.
- Design and implement a two-stream CNN model, based upon the successes reported in previous studies [Simonyan and Zisserman, 2014], that increases classification accuracy compared to single-stream methods.
- The proposed model is significantly more lightweight and requires no pretraining without an unacceptable trade-off in performance.
- Collate existing research surrounding visual explanations of action recognition and design implementation for use in the proposed model with skeleton-based input.

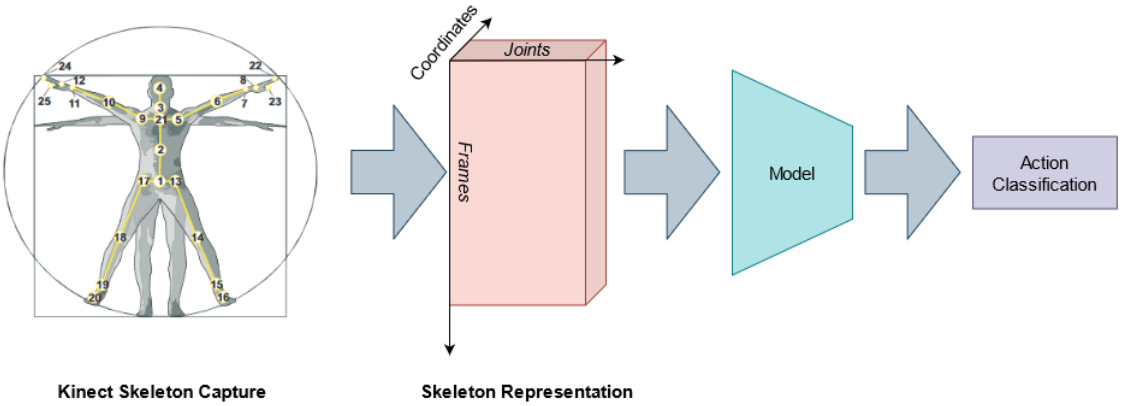

Figure 1: Proposed workflow for skeleton-based action recognition

Chapter 2: Background

2.1 Action Recognition

A study, vital to the completion of this work, gathered a large-scale dataset for 3D human activity to be used for analysis [Liu *et al.*, 2019]. In this study, they created the NTU RGB+D, and later NTU RGB+D 120, datasets. These datasets contain video sequences of 106 distinct subjects performing 120 different action classes. To do this, the study used the Microsoft Kinect [Kinect, 2014] to collect RGB videos, depth sequences, human skeleton data (3D coordinate locations of 25 specified joints), and infrared frames. For this work, the skeleton data is used to train the model to identify and classify actions.

Initially, long short-term memory (LSTM) networks were the natural choice for skeleton sequences due to the network being designed to model long-term temporal dependency problems. A standout model was produced, building upon previous works utilising recurrent neural networks, integrating a 3-layer LSTM framework [Zhang *et al.,* 2017]. However, over time more studies adopted convolution neural networks (CNN) to learn features. This led to an increase in performance in many subsequent studies over recent years, with many furthering advancements in model specifics. For example, [Du et al., 2016] proposed modifying the network's input to allow frame, joint, and coordinate dimensions of a skeleton sequence to be used as width, height, and channel of an image representation.

CNN models have a remarkable ability to extract high-level information from an input image and have been previously utilised to learn the desired spatial-temporal features from skeleton data [Du et al., 2016; Ke et al., 2017]. These CNN methods represent the skeleton sequence as an image, encoding skeleton joint coordinates in a 2D matrix which is fed into a CNN to recognise the underlying action as it would in image classification cases. In this case, the only features considered to learn co-occurrence are the neighbouring joints within a convolution kernel. This introduces the problem of effectively learning co-occurrences of features across all joints globally. To overcome this problem, the proposed model treats each joint of a skeleton as a channel so that the convolution layer can learn the co-occurrences from all joints globally.

[Simonyan and Zisserman, 2014] proposes a framework for a two-stream input CNN in which features are learned from both inputs explicitly and fused. The model proposed in this work accepts this framework to learn point and global level features of both the skeleton sequence and the temporal skeleton motion.

Finally, [Song *et al.,* 2019] acts as a target for classification accuracy while attempting to reduce the network's size and the number of trainable parameters within it. This study successfully implemented a multiple-stream CNN for action recognition on the NTU RGB+D dataset, making it an ideal comparison as a state of the art (SOTA) method.

2.2 Explanations

Explainable classification is a crucial component for understanding and interacting with complete action recognition systems. Explanations can be defined as the reasoning as to why a specific decision is consistent with visual evidence. [Hendricks et al., 2016] introduce two types of explanation systems: (1) Introspection: a system that explains how a model determines the final output due to high filter activation or presence of specific features. (2) Justification: a system which can produce a sentence in conjunction with visual evidence to highlight features relevant to the decision process. (e.g., "This is a zebra because it has black and white stripes.")

Another type of explanation system called counterfactual explanations [Kanehira et al., 2019] generate explanations in the form "X is classified to A not B because C and D exist in X.". Within that body of work, a successful explanation is classified as one that is interpretable by humans outside of the system's intricate knowledge, and the output should have fidelity to the explained target. A multimodal explanation module is integrated into a video classification CNN [Kanehira et al., 2019]. Presented in this work is an explanation module, designed for use in action recognition with skeleton sequence input, building upon the previously mentioned work's successes. A small number of studies have been recently conducted into more explainable video recognition networks and how these may be designed and implemented, providing valuable insights into creating a successful explanation module [Hiley et al., 2019; Nourani et al., 2020].

Chapter 3: Methods

3.1 Co-occurrence Feature Learning

CNNs have quickly become one of the most successful neural network models for applications in image classification, video classification, and object detection. Compared with sequential structures such as RNNs, CNNs can simultaneously encode both spatial and temporal information. Convolutional operations can be decomposed into two parts: local feature aggregation and global feature aggregation across channels. This can be represented as a 3D tensor, denote T as dimension1 x dimension2 x dimension 3. The channels dimension will always be aggregated globally; therefore, any feature may become channels through transposition.

Previously, the more successful CNN-based action recognition methods had been expected to have the joint coordinates as channels aggregated globally. This would mean that co-occurrence features between joints are only learned locally within the convolution layer. Problematically this would not allow for capturing long-range joint interactions involved in some actions within the dataset. More recent studies have shown that aggregating these features globally will result in greater performance in action recognition. This can be achieved by having the joint dimension in the channels for the CNNs input.

3.2 Skeleton Motion

Besides the individual representation of joint positions, it is of equal importance to consider the temporal movements of joints when training a network to classify movements into action categories. This can be learned implicitly within the CNN; however, it has been shown to achieve higher accuracy rates when a representation of skeleton motion is introduced and is explicitly fed to the network [Simonyan and Zisserman, 2014].

To represent the skeleton of a person in frame f, we formulate it as:

$$S^f = \{ J_1^f , J_2^f , \dots , J_{25}^f \}$$

where $J = (x, y, z)$ represents one set of 3D joint coordinates.

The skeleton motion is defined as the temporal difference of each joint between two consecutive frames:

$$M^f = S^{f+1} - S^f = \{ J_1^{f+1} - J_1^f, J_2^{f+1} - J_2^f, \dots , J_{25}^{f+1} - J_{25}^f \}$$

Raw skeleton coordinates S and the skeleton motion M are independently fed into the network using a two-stream paradigm. The network concatenates their feature maps later in the model to utilize the resulting information from two outputs.

3.3 CNN Network

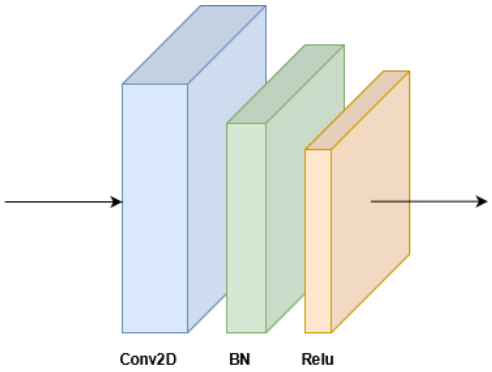

Figure 2: Illustration of a basic block. Conv2D represents the 2D spatial convolution over the input, followed by a Batch Normalisation layer and a ReLU activation function to introduce non-linearity.

Two different versions of the CNN model have been produced as part of this project due to physical computing limitations. The difference between the two models is the number of input frames fed into the network from each sample video action. One model receives a select sample of 32 frames, with the other model receiving all possible frames from each video, normalised to the maximum possible length of 300 frames. Unfortunately, due to hardware limitations and the lack of access to the hardware available outside of COVID times, results discussed within this project come from the first model. With the addition of theoretical discussion based upon what can be inferred surrounding the complete input model.

Figure 3 shows the network architecture of the framework. A skeleton sequence X can be represented with a F x J x D tensor, where F is the number of frames in the sequence, J is the number of joints in the skeleton (25 for NTURGB+D dataset), and D is the coordinate dimension. The temporal skeleton motion outlined above is also of the same shape as X. Both the skeleton and skeleton motion are fed into the network directly as part of a two-stream input. The two network branches share the same architecture. However, their parameters are learned independently, with their feature maps being fused by concatenation along the channels dimension after the third basic block.

Given the skeleton sequence and motion inputs, the features are learned hierarchically. Firstly, point-level features are encoded with 1 x 1 and 3 x 1 convolution layers. The kernel size along the joint dimension are kept as 1 to force the network to learn point-level representation from the 3D coordinates for each joint independently. After the point-level feature learning, we permute the feature map with parameters (1, 3, 2) so that the joint dimension is moved to channels of the tensor. Then, in the next stage, all subsequent convolution layers extract global co-occurrence features from all skeleton joints described in Section 3.1. Finally, the feature map passes through two fully connected layers for final classification.

The network contains only 36,000 total parameters making it significantly smaller than any current SOTA networks [Yan et al., 2018; Song et al., 2019; Shi et al., 2019]. As a result of the model's minimal size, it is easy to train the network from scratch without the need for pretraining.

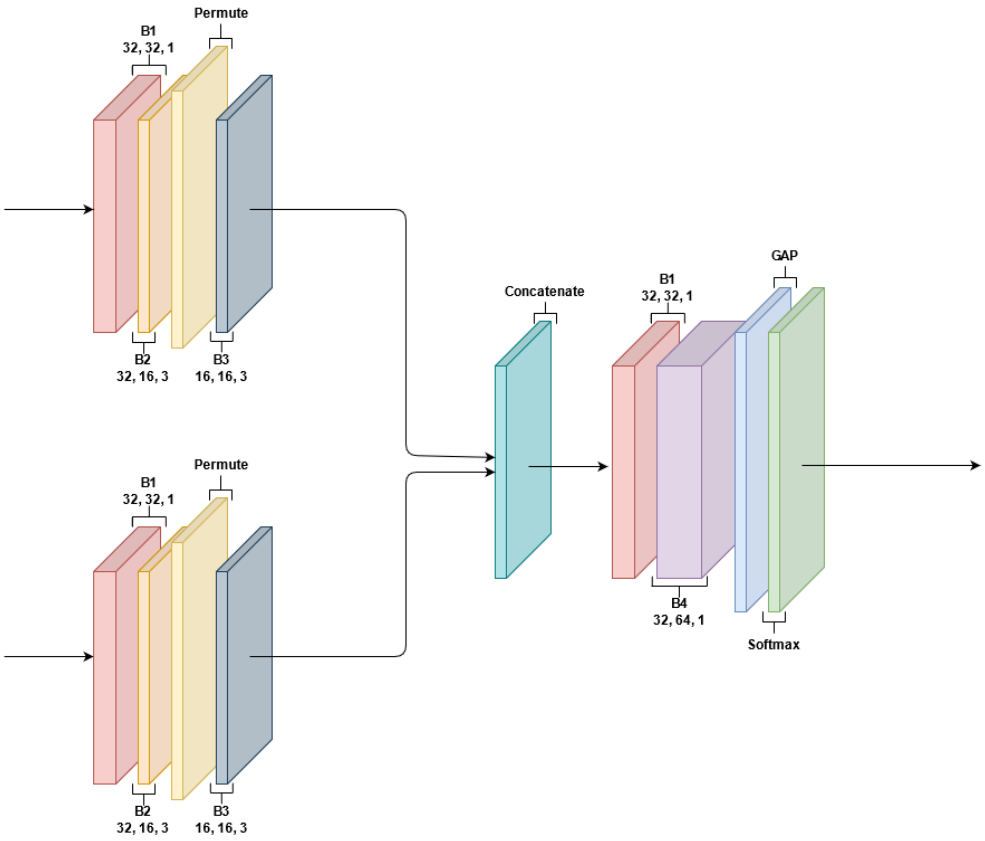

Figure 3: Illustration of the network architecture. A total of 8 basic blocks. The three numbers of each block represent the number of input channels, the number of output channels, and the stride. GAP represents the global average pooling layer.

3.4 Action Recognition

After extracting both point-level and globally co-occurrence features in both the skeleton representation and the temporal skeleton motion, the network applies a global average pooling layer. This is performed at this stage to pool feature maps of different samples to the same size. For the final recognition task, a SoftMax function is used to convert the real vector feature map to a categorical probabilities vector used for predictions.

3.5 Explanations Module

Given the current proposed model for an action recognition CNN, let

$$S = \Sigma_{f=1}^{F}(s), \qquad s_f = \{J_1, J_2, \dots, J_{25}\}, \qquad \mathrm{J} = (\mathrm{x}, \mathrm{y}, \mathrm{z})$$

represent a skeleton sequence of F frames with joint J 1-25 with 3D coordinates (x,y,z). The method ends on a function σ that maps the skeleton sequence to a SoftMax probability $\sigma_c(x)$ for a class with the index of c within all C classes. The explainable system's proposed goal is to derive a sequence of importance maps that map identified features to frames in the skeleton sequence S.

As shown in Figure 4, the proposed method for implementing an explanations module is to take output feature maps from intermittent points in the CNN model, input them into the SoftMax probability function and map this result with the frame information. These results are then compared against the final output (prediction). This will identify the changes in prediction as more features are learned as the network progresses. For example, in skeleton S, given that the result of σ with the first taken output m^1 is c_1 (class 1) we have its current feature map containing identified point-level features. At this point, we can say S is classified to c_1 due to features in m^1. At the following taken result of σ with output m^1 the classification is now c_2. Continuing through each sub-output of the classification model into the explanation module, the network can produce a picture of ongoing feature learning and its effect on the classification.

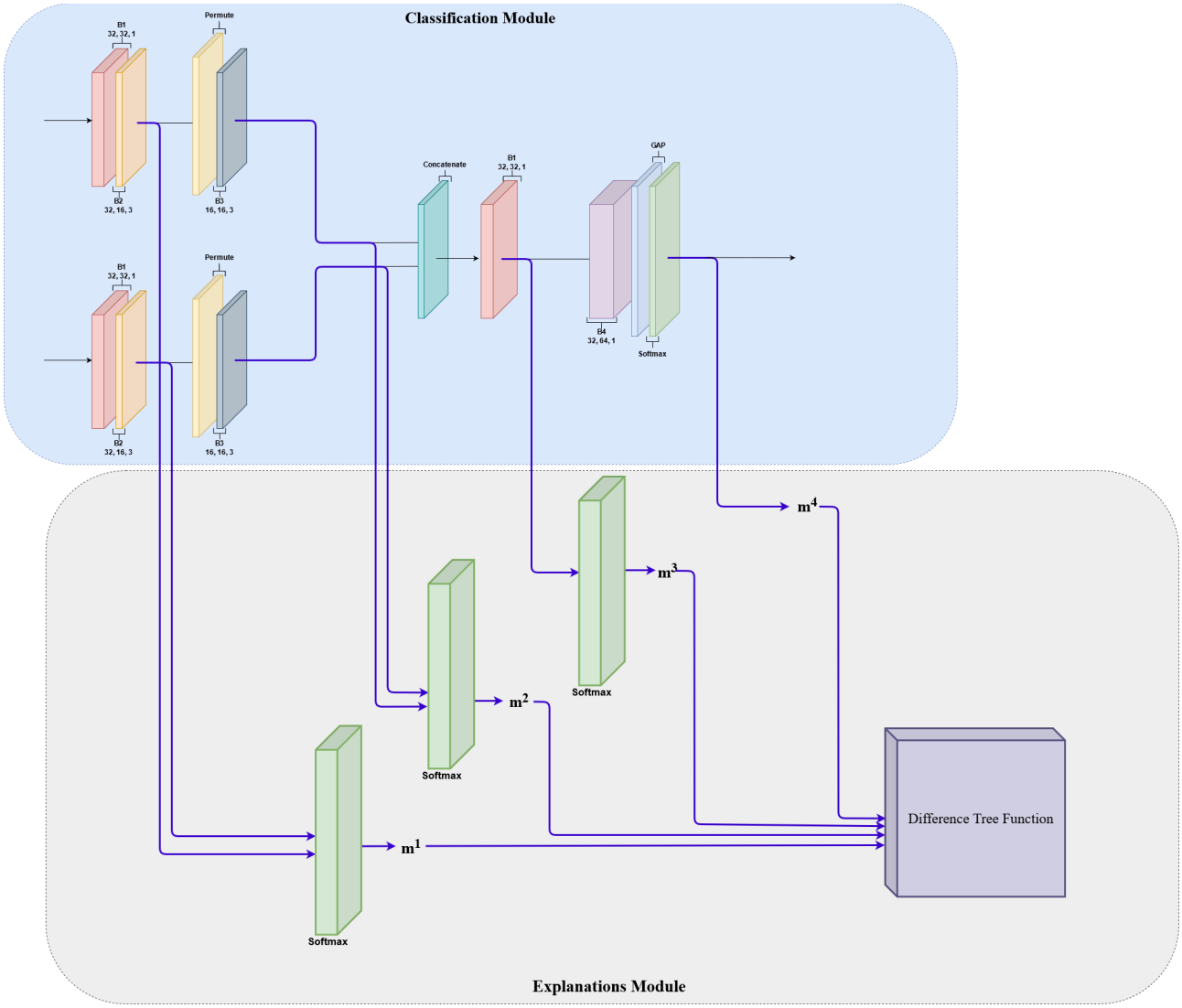

Figure 4: Illustration of the network architecture, including the highlighted Explanations Module.

Chapter 4: Experiments

4.1 Datasets and Implementation Details

The NTU RGB+D 120 dataset is currently the largest and most commonly used action recognition dataset. It contains over fifty thousand skeleton sequences, which are classified as one of 120 action classes. In the experiments conducted for this work, any two-person actions have been removed from the training and testing data. This leaves 93 potential action classes for one person actions. The sequences are performed by 40 volunteer actors of different ages ranging from 10 to 35 years old. Within the dataset are 3D joint locations of each frame detected by the Kinect depth sensor. There are 25 joints for each subject, with each video containing no more than two subjects. There are two recommended evaluation protocols outlined in the original paper [Liu et al., 2019]: (1) Cross-subject (CS): sequences of 20 subjects are used for training, and sequences of the remaining 20 subjects are used for validation. (2) Cross-view (CV): Samples are split by camera views, samples from two camera views are used for training, and samples from the final camera view used for validation.

For the model's training, a subsection of the video sequence is selected by creating a random range ratio that is then used to find a start point for the subsequence. The selected input sequence is always normalised to 32 frames as not all actions are equal in duration across the dataset. As detailed in Section 3.3, the experiments on the network have been limited by the hardware available. If this were not to be the case, then each skeleton sequence would be normalised to the maximum length of 300 frames. This would allow for a greater view of the whole action rather than a cropped subsection of the action. In order to reduce overfitting, dropout layers with a ratio of 0.1 are implemented after both final convolution layers and after the first fully connected layer. In this paper's experiments, the model has been trained for only 10000 iterations with a batch size of 64. A stochastic gradient descent (SGD) (with momentum) is applied as the optimisation strategy with a learning rate of 0.01 and momentum of 0.9. Categorical cross-entropy is selected as the loss function.

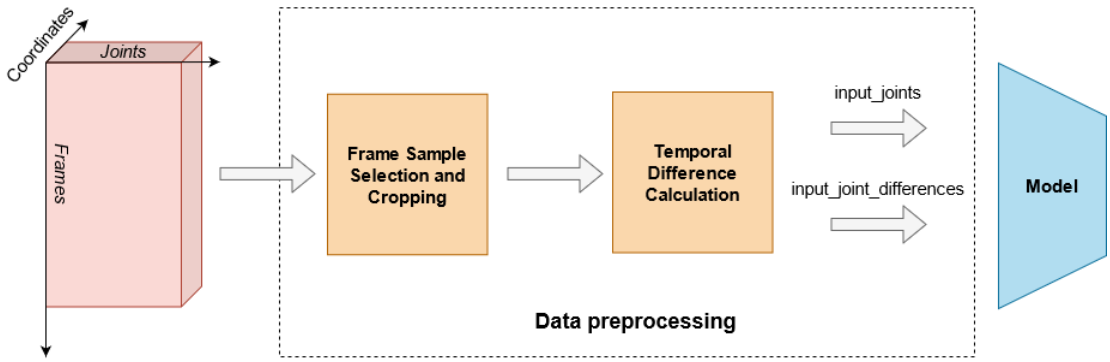

Figure 5: Illustration of data selection and preprocessing to enable a two-stream input.

4.2 Ablation Study

4.2.1 Point-level and Global Feature Learning

In order to better understand the behaviour of the global co-occurrence feature learning, an ablation study is performed. The 'global model' is the proposed version of the model where joint features are aggregated global across the channels dimension. The 'local model' is a modified version of the model where the permutation has been removed so that the 3D joint coordinates are the feature in the channels.

It is shown that the global model consistently outperforms the local version of the model in both the cross-subject and cross-view evaluations. The results show a significant difference in cross-subject evaluation than in cross-view. This observation suggests that with global co-occurrence, the variation of action caused by variation in actor can be minimised.

Method	**Cross-Subject**	**Cross-View**
Local Model	69.9	75.1
Global Model	74.0	77.3

Table 1: Comparison of local and global model classification accuracy on the NTU RGB+D dataset.

4.2.2 Two Stream Input

To review the impact of a two-stream input classification accuracy, we test the complete model's performance and both input streams' performance independently. 'Complete Model', 'Joint Location', and 'Temporal Joint Motion' representing the model functioning as proposed, just the 3D coordinate location of joints and the temporal difference in joint locations. The results of which can be found in Table 2. It can be seen clearly that the two-stream input method outperforms each of the single-stream methods. For single-stream methods, the Joint Location performs slightly better than the Temporal Joint Motion for the cross-subject evaluation, whereas, in the cross-view evaluation, the result is reversed. This demonstrates the complementary nature of the two modalities. When combing the two streams, a notable increase in performance is observed as expected.

Method	**Cross-Subject**	**Cross-View**
Joint Location	62.5	65.3
Temporal Joint Motion	60.8	65.9
Complete Model	74.0	77.3

Table 2: Comparisons of the classification accuracy with different input sources on the NTU RGB+D dataset.

4.3 Comparison to Existing Solutions

Model	Parameters	Cross-Subject	Cross-View
[Liu *et al.,* 2016]	-	55.0	57.9
[Caetano *et al.,* 2019a]	-	67.9	62.8
[Caetano *et al.,* 2019b]	-	67.7	66.9
[Yan *et al.,* 2018]	3.10	70.7	73.2
[Song *et al.,* 2019]	6.21	74.6	75.3
[Shi *et al.,* 2019]	6.94	82.5	84.2
[Li *et al.,* 2017]	4.20	83.2	89.3
This Model	0.035	74.0	77.3

Table 3: Comparison with SOTA methods on NTU RGB+D 120 dataset in parameter number (million) and accuracy (%): cross-subject and cross-view.

As seen in table 3 the proposed model is significantly smaller in total parameters than the current STOA methods. It can be observed that the proposed model achieves similar accuracy to that of [Song *et al.,* 2019] and a measurable increase in accuracy to that of [Yan *et al.,* 2018]. More recent and advanced models that utilise a graph convolutional networks significantly outperform the proposed model. However, this is to be expected as they are created with the intention of accuracy above anything else. The approach adopted in this study returns a significant increase in accuracy over the previously standard LSTM-based methods [Liu *et al.,* 2016; Caetano *et al.,* 2019a; Caetano *et al.,* 2019b]. Compared with a similar two-stream CNN method [Li *et al.,* 2017], the accuracy is 9.2% worse for cross-subject evaluation and 12% worse for cross-view evaluation.

Chapter 5: Conclusion

5.1 Review of project aims and objectives

Action recognition of skeleton sequences is currently at the forefront of computer vision, with a rich and extensive body of work committed to pushing the limitations of accuracy and efficiency. Explainable networks aim to alleviate many black-box models' issues to allow for a better understanding of results and decisions.

This paper presents the design and implementation of a two-stream hierarchical co-occurrence feature learning model for skeleton-based action recognition. The work demonstrates the capabilities of the CNN to learn both point-level and global co-occurrence features of the skeleton joints and skeleton motion. Through experiments and comparison, it is proven that the proposed addition of a temporal joint positions stream allows the network to learn more details of local and global features, which allows for greater classification accuracy. Experimental results show that the presented model achieves similar classification accuracy to current SOTA CNN-based models while requiring significantly fewer training parameters. With the hardware-reliant changes proposed, an additional increase in accuracy is predicted with no additional training parameters requirement. Although the presented model performs well compared with CNN models, it is still behind the new generation of GCNs created for action recognition tasks. In addition to the presented model is a design and outline of an explanation module for use in the network. The module is described to a high-level view and offers insight as to implementation for future work.

5.2 Future Work

Moving forward with this body of work, it would be improved and extended in two main parts: (1) Implementation of the proposed Explanations Module: building upon the design and outline presented in this work, an explanations module would allow for further investigation into decision making within the network. (2) Extending the frame amount: as detailed in this study, a second more complete version of the skeleton sequence could be used as input for the network. This is expected to produce a significant increase in classification accuracy. As well as this adapting the network to work for two person actions that have been excluded from the dataset in this work.

5.3 Final Remarks

Through completion of this work I have learned a considerable amount about both the areas explored and my own abilities in research and time management. The study successfully produced a lightweight two-stream CNN for action recognition and detailed possible implementation for an explanation module. The presented network is fairly accurate in classification and particularly simple in design and size. It has the potential to be improved with small amendments outlined throughout the work and built upon in the future.

References

[Liu *et al.*, 2019] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, Alex C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. In *CVPR*, 2019.

[Simonyan and Zisserman, 2014] Karen Simonyan, Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *CSVR*, 2014.

[Liang *et al.,* 2019] Duohan Liang, Guoliang Fan, Guangfeng Lin, Wanjun Chen, Xiaorong Pan, Hong Zhu. Three-Stream Convolutional Neural Network With Multi-Task and Ensemble Learning for 3D Action Recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[Du *et al.*, 2016] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition*, 2016.

[Ke *et al.*, 2017] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A New Representation of Skeleton Sequences for 3D Action Recognition. In *CVPR*, 2017.

[Kinect, 2014] Microsoft. docs.microsoft.com/en-gb/previous-versions/windows/Kinect. 2014.

[Zhang *et al.,* 2017] Songyang Zhang, Xiaoming Liu, Jun Xiao. On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017

[Kanehira *et al.,* 2019] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, Tatsuya Harada. Multimodal Explanations by Predicting Counterfactuality in Videos. In *CSCV*, 2019.

[Hiley *et al.,* 2019] Liam Hiley, Alun Preece, Yulia Hicks. Explainable Deep Learning for Video Recognition Tasks: A Framework & Recommendations. In *CSLG,* 2019.

[Hendricks *et al.,* 2016] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell. Generating Visual Explanations. In *CSCV,* 2016.

[Nourani *et al.,* 2020] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, Nicholas Ruozzi, Vibhav Gogate. Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition. In *CSHC,* 2020.

[Liu *et al.,* 2016] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatiotemporal LSTM with trust gates for 3d human action recognition. In European Conference on Computer Vision (ECCV), 2016.

[Caetano *et al.,* 2019a] Carlos Caetano, Francois Bremond, and William Robson Schwartz. Skeleton Image Representation for 3D Action Recognition Based on Tree Structure and Reference Joints. In SIBGRAPI Conference on Graphics, Patterns, and Images, 2019.

[Caetano *et al.,* 2019b] Carlos Caetano, Jessica Sena, Franeois Bremond, Jefersson A. Dos Santos, and William Robson Schwartz. SkeleMotion: A New Representation of Skeleton Joint Sequences based on Motion Information for 3D Action Recognition. In IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019.

[Yan *et al.,* 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI Conference on Artificial Intelligence, 2018.

[Song *et al.,* 2019] Yi-Fan Song, Zhang Zhang, and Liang Wang. Richly activated graph convolutional network for action recognition with incomplete skeletons. In IEEE International Conference on Image Processing (ICIP), 2019.

[Shi *et al.,* 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In IEEEConference on Computer Vision and Pattern Recognition (CVPR), 2019.

[Li *et al.,* 2017] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In ICMEW, 2017.

[Liu *et al.,* 2020] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In CVPR, 2020.

[Shi *et al.,* 2020] Lei Shi, Yifan Zhang, Jian Cheng, Hanqing Lu. Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action Recognition. In CVPR, 2020.

[Cheng *et al.,* 2020] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, Hanqing Lu. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.