Actional-Structural Graph Convolution Networks Applied to Martial Arts, Dancing and Sports Datasets for Action Recognition

By Adam Russell 12711142
Supervisor: Cen Wan

Contents

1. Abstract

This proposal outlines the application of Actional-Structural Graph Convolution Networks (AS-GCN) to Martial Arts, Dancing and Sports Datasets for action recognition. Initially the AS-GCN pipeline (Li, et al., 2019) will be deployed on a cloud deep learning machine image (DLMI) and trained on the NTU RGB+D 120 (Liu, et al., n.d.) data set referenced in the paper. The trained model will then be applied to the Martial Arts, Dancing and Sports (MADS) (Zhang, et al., 2017) which contain sequences of Tai-chai, Karate, Jazz style dance, Hip-hop style dance and other sports such as football, tennis and basketball. The recognition head will be assessed as a tool for temporal localization of component action recognition e.g. time stamping of kicks in a martial arts sequence. The predication head will be assessed for its accuracy in predicting the last frames in a sequence. Finally, the MADS data set will be used to train a new AS-GCN model either as a binary classifier or with the 30 action sequence labels as the output to the network.

2. Introduction

2.1. What constitutes an action?

Action recognition seeks to classify sequences of human body part motions. What constitutes as an action is very broad even with this posterior statement. In this proposal we regard an action as all the four subcategories of physical human activity as defined below (Jegham, et al., 2020):

'Gesture: It is a visible bodily action representing a specific message. It is not a matter of verbal or vocal communication but rather a movement made with the hands, face or other parts of the body such as Okay gestures and thumbs up.

Action: It is a set of physical movements conducted by only one person like walking and running.

Interactions: It is a set of actions executed by at most two actors. At least one subject is a person and the other one can be a human or an object (hand shaking, chatting, etc).

Group activities: It is a mixture of gestures, actions, or interactions. The number of performers is at least two plus one or more interactive objects (playing volleyball, obstacle racing, etc).'

2.2. Why work with Sequences?

Of course, many actions can be recognized from a single frame using now well-established deep learning convolution neural networks (CNN). These will often detect subtle features of the image around humans such as objects and environment (Konushin, 2017). For example, a model may classify an image of a human playing football by detecting the features of the football on a green field. Further many actions share similar poses as part of the sequence. For instance, an image of footballer with the ball at their foot may be of them receiving a pass or attempting a pass. Therefor we further elaborate that action recognition in the context of this proposal is purely based on human pose sequences. All environmental and objects are to be filtered out. Processing sequences of frames in theory mitigate the ambiguity of processing single static frames.

2.3. Challenges working with Sequences

Working with sequences of frames directly is unfeasible as the size of the inputs to the neural network increases rapidly with the number of frames in a sequence (Konushin, 2017). Various filtering techniques have been developed to reduce the input size such as foveated architectures, optical flow and dense trajectories. This project will implement structural graphs whereby a graph with edges for bones and vertices for joints is extracted for each human in each frame for example

Figure 1. The resulting lower resolution, sparse space time volumes are significantly more efficient to compute and store.

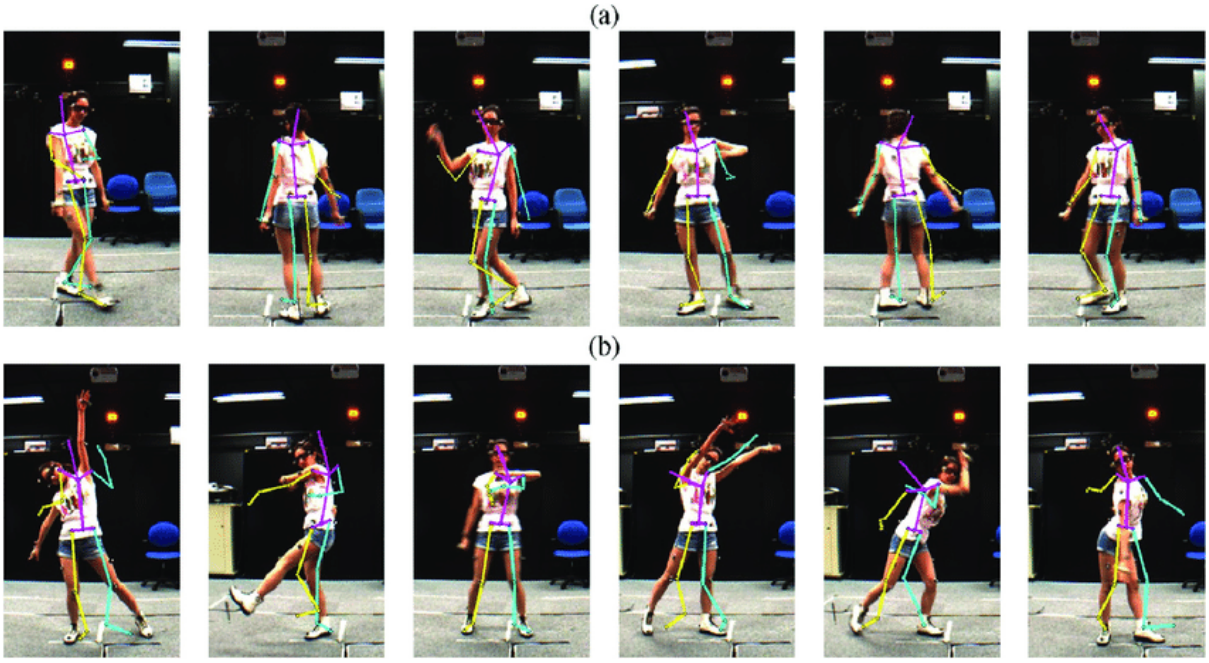

Figure 1 Example of skeletal graph extraction (Zhang, et al., 2017)

2.4. Applications of Action Recognition

Human action recognition is an active area of computer vision research for both real time and recorded data applications (Ni, n.d.). Deploying action recognition models to existing CCTV infrastructures have high value for all emergency services. Police, military and security organisations seek automated detection of crime and terrorism events. Fire services seek to minimise response times to road traffic accidents. Automated recognition of trauma/causality type (for example stroke/heart attack) would improve triage and save valuable time in deploying the correct treatment. Automated Factory's and warehouses seek to monitor the safety of human employees from the hazardous of collaborative robots (cobots). Self-driving car manufacturers would see large gains in pedestrian safety if they implemented action recognition models that detect if a pedestrian was about to step out into the street. Film, television, sports and media seek to use this technology to automatically time stamps actions of interest in sequence such as headers in a game of football or handshakes at a political conference for a news video archive indexing and retrieval.

3. Literature Review

3.1. Actional-Structural Graph Convolution Network

(Li, et al., 2019) propose that actions depend on richer joint dependencies than fixed skeletal graphs portray between joints that are far apart. Even basic actions such as walking or sitting mobilize many far apart joints and appendages in complex ways. They propose these Actional Links (A-Links) supplement traditional structural links of a fixed skeletal graph as shown in Figure 2. To generate these A-links the first stage of the Actional-Structural Graph Convolution Network (AS-GCN) proposed by (Li, et al., 2019) is the A-link inference module (AIM) consisting of an encoder and decoder. Skeletal space time volumes are input. The encoder 'exacts link features from 3D joint position and then converts the link features to the linking probabilities' . As information is propagated between joints and links the module learns link features. Concurrently Structural-links

(S-Links) are extracted at high order polynomials than the original to increase the set of 'local' neighbours of each joint/link.

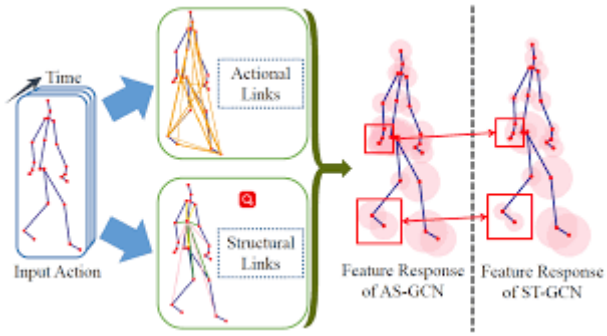

Figure 2 Graph Development (Li, et al., 2019)

The Actional-Structural Graph Convolution Block combine the actional graph convolution (AGC) and the Structural Graph Convolution (SGC) for each time stamp. A hyper parameter on the AGC trades off the importance between structural features and actional features. There are nine blocks of AS-GCN followed by one layer of Temporal Convolution to capture inter-frame action features along the time axis. This is the backbone of the network.

The recognition head of the network classifies actions using global averaging pooling and a SoftMax classifier with a standard cross entropy loss function. The Prediction head features the back-bone network in reverse with a standard l2 loss function for prediction. The entire pipeline is shown is shown in Figure 3.

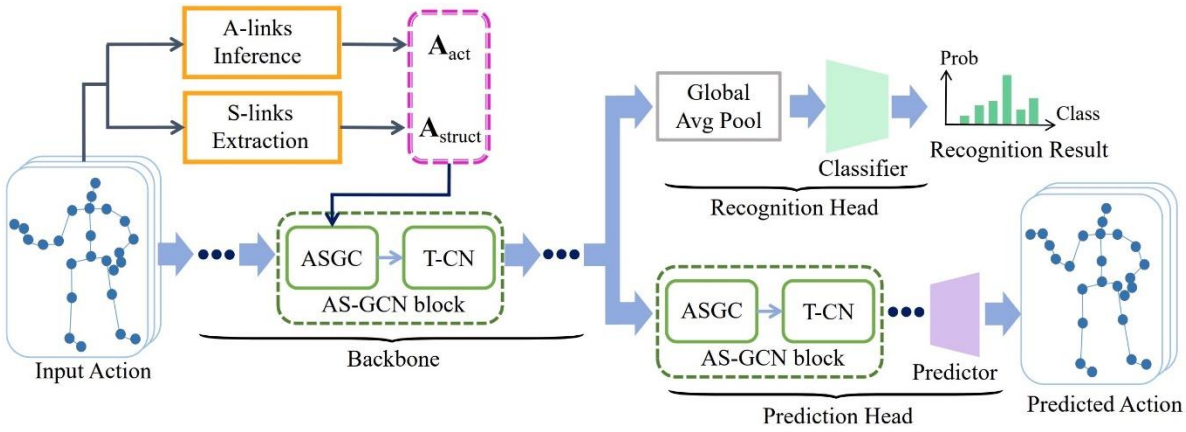

Figure 3 AS-GCN Pipeline (Li, et al., 2019)

3.2. NTU RGB+D 120 Data Set

The NTU RGB+D 120 (Liu, et al., 2019) data set features 114480 samples and 120 action classes sub categorised into 82 Daily Actions Table 1, 12 Medical Conditions Table 2 and 26 Mutual Actions/Two Person interactions Table 3. This dataset has been captured using three Kinect V2 cameras concurrently. The three kinetic v2 cameras were all set at the same height at -45, 0, +45 horizontal angles with the actors asked to perform each action towards both the left and right cameras resulting in 6 perspectives per action per set up. The samples are available in a variety of formats (RGB video, IR, Full depth maps etc.) but on 3D skeletons (body joints) will be required for training.

A1: drink water	A2: eat meal	A3: brush teeth	A4: brush hair
A5: drop	A6: pick up	A7: throw	A8: sit down
A9: stand up	A10: clapping	A11: reading	A12: writing
A13: tear up paper	A14: put on jacket	A15: take off jacket	A16: put on a shoe
A17: take off a shoe	A18: put on glasses	A19: take off glasses	A20: put on a hat/cap
A21: take off a hat/cap	A22: cheer up	A23: hand waving	A24: kicking something
A25: reach into pocket	A26: hopping	A27: jump up	A28: phone call
A29: play with phone/tablet	A30: type on a keyboard	A31: point to something	A32: taking a selfie
A33: check time (from watch)	A34: rub two hands	A35: nod head/bow	A36: shake head
A37: wipe face	A38: salute	A39: put palms together	A40: cross hands in front
A61: put on headphone	A62: take off headphone	A63: shoot at basket	A64: bounce ball
A65: tennis bat swing	A66: juggle table tennis ball	A67: hush	A68: flick hair
A69: thumb up	A70: thumb down	A71: make OK sign	A72: make victory sign
A73: staple book	A74: counting money	A75: cutting nails	A76: cutting paper
A77: snap fingers	A78: open bottle	A79: sniff/smell	A80: squat down
A81: toss a coin	A82: fold paper	A83: ball up paper	A84: play magic cube
A85: apply cream on face	A86: apply cream on hand	A87: put on bag	A88: take off bag
A89: put object into bag	A90: take object out of bag	A91: open a box	A92: move heavy objects
A93: shake fist	A94: throw up cap/hat	A95: capitulate	A96: cross arms
A97: arm circles	A98: arm swings	A99: run on the spot	A100: butt kicks
A101: cross toe touch	A102: side kick	-	-

Table 1 Daily Actions (82) (Liu, et al., n.d.)

A41: sneeze/cough	A42: staggering	A43: falling down	A44: headache
A45: chest pain	A46: back pain	A47: neck pain	A48: nausea/vomiting
A49: fan self	A103: yawn	A104: stretch oneself	A105: blow nose

Table 2 Medical Conditions (12) (Liu, et al., n.d.)

A50: punch/slap	A51: kicking	A52: pushing	A53: pat on back
A54: point finger	A55: hugging	A56: giving object	A57: touch pocket
A58: shaking hands	A59: walking towards	A60: walking apart	A106: hit with object
A107: wield knife	A108: knock over	A109: grab stuff	A110: shoot with gun
A111: step on foot	A112: high-five	A113: cheers and drink	A114: carry object
A115: take a photo	A116: follow	A117: whisper	A118: exchange things
A119: support somebody	A120: rock-paper-scissors	-	-

Table 3Mutual Actions / Two Person Interactions (26) (Liu, et al., n.d.)

3.3. Martial Arts, Dancing and Sports (MADS)

The MADS contain sequences of Tai-chai, Karate, Jazz style dance, Hip-hop style dance and other sports such as football, tennis and basketball performed by a single actor at a time. It features >53000 video frames and 30 action classes sub categorised into Tai-chi, Karate, Jazz, Hip-Hop and sports each with 6 actions. The video was shot in a studio environment using point grey bumble-II cameras. Each sequence is 60 or 80 seconds long with frame rates of 15 or 10 or 20.

4. Project Description, Aims and Objectives

Initially the AS-GCN pipeline proposed by (Li, et al., 2019) will be deployed on a cloud deep learning machine image (DLMI) and trained on the NTU RGB+D 120 data set referenced in the paper. The computation time shall be recorded for bench marking. Concurrently the skeletal models of MADS data set will be investigated for compatibility as inputs to the AS-GCN trained model. If the provided skeleton models are not direct compatible OpenPose toolbox (Hidalgo, et al., n.d.) may be implemented to estimate skeleton data.

The trained AS-GCN model recognition and prediction head model shall be applied to the MADS data set for temporal localization of component action recognition. The training data action classes are different from MADS, therefore NTU RGB+D 120 action classes shall manually be applied to MADS footage to assess the accuracy of the pre-trained AS-GCN model.

Finally, the MADS data set will be used to train a new AS-GCN model either as a binary classifier or with the 30 action sequence labels as the output to the network. This will require the author to deeper understand the model architecture of AS-GCN (Li, et al., 2019). The MADS data set may require pre-processing to; extract compatible skeletal data, reshape the data for the AS-GCN architecture and split the sequences out into shorter sequences.

In summary the objectives of the project are:

1. Deploy the AS-GCN model on cloud deep learning machine image and train
2. Manually assign labels actions from the ROSE dataset to the MADS dataset footage and assess the AS-GCN model accuracy
3. Pre-process the MADS for training a new AS-GCN with MADS action labels

5. Tools and Programming Languages

The accompanying code repository (Li, et al., 2019) for the AS-GCN model lists the following requirements:

- Python 3.6
- Pytorch 0.4.1
- pyyaml
- argparse
- numpy

The AS-GCN paper notes that models were trained using 8 GTX-1080Ti GPUs utilising the Pytroch CUDA integration. As comparable hardware is not directly available to the author, Amazon Web Service (AWS) cloud deep learning machine image (DLMI) shall be used for model training. Initial selection of DLMI and instances are shown below. The DLMI selected is includes Pytorch and NVIDIA CUDA frameworks and Conda environments. Amazon EC2 G4 Instances are entry level machine learning instances with NVIDIA T4 GPUs (Amazon Web Services, n.d.).

DLMI: Deep Learning AMI (Ubuntu 16.04), MXNet-1.6.0, Tensorflow-2.1.0 & 1.15.2, PyTorch-1.4.0, Keras-2.2, & other frameworks, configured with Neuron, NVIDIA CUDA, cuDNN, NCCL, Intel MKL-DNN, Docker & NVIDIA-Docker. (Amazon Web Services, n.d.)

Instance Size: g4dn.2xlarge, vCPUs=16, Memory (GB)=64, GPU=1, Storage (GB)=225, Network Bandwidth (Gbps)=Up to 25, EBS Bandwidth (GBps)=4.75.

6. Work Plan

The submission deadline for this project is September 14th, 2020. Below is a rough schedule for this project commencing after the final summer term exam June 12th, 2020 which roughly covers 13 weeks.

June 12th – June 28th; Objective 1, 2 weeks

June 29th – July 13th; Objective 2, 2 weeks

July 14th – August 30th; Objective 3, 6 weeks

August 31st – September 14th; Write up/Contingency, 2 weeks

7. References

Amazon Web Services, n.d. *Amazon EC2 G4 Instances.* [Online]
Available at: https://aws.amazon.com/ec2/instance-types/g4/
[Accessed May 2020].

Amazon Web Services, n.d. *Deep Learning AMI (Ubuntu 16.04).* [Online]
Available at: https://aws.amazon.com/marketplace/pp/Amazon-Web-Services-Deep-Learning-AMI-Ubuntu-1604/B077GCH38C
[Accessed May 2020].

Hidalgo, G. et al., n.d. *openpose.* [Online]
Available at: https://github.com/CMU-Perceptual-Computing-Lab/openpose
[Accessed May 2020].

Jegham, I., Khalifa, A. B., Alouani, I. & Mahjoub, M. A., 2020. *Vision-based human action recognition: An overview and real world challenges.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S174228761930283X
[Accessed May 2020].

Konushin, A., 2017. *Action classification with convolutional neural networks.* [Online]
Available at: https://www.coursera.org/learn/deep-learning-in-computer-vision/lecture/0R0ds/action-classification-with-convolutional-neural-networks
[Accessed May 2020].

Li, M. et al., 2019. *Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition.* [Online]
Available at: https://arxiv.org/abs/1904.12659
[Accessed May 2020].

Li, M. et al., 2019. *The model architecture of AS-GCN (for human action recognition).* [Online]
Available at: https://github.com/limaosen0/AS-GCN
[Accessed May 2020].

Liu, J. et al., 2019. *NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding.* [Online]
Available at: https://arxiv.org/pdf/1905.04757.pdf
[Accessed May 2020].

Liu, J. et al., n.d. *Action Recognition Datasets: "NTU RGB+D" Dataset and "NTU RGB+D 120" Dataset.* [Online]
Available at: http://rose1.ntu.edu.sg/datasets/actionrecognition.asp
[Accessed May 2020].

Ni, H., n.d. *Human action recognition.* [Online]
Available at: https://www.turing.ac.uk/research/research-projects/human-action-recognition
[Accessed May 2020].

Zhang, W. et al., 2017. *Martial Arts, Dancing and Sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation.* [Online]
Available at: https://www.sciencedirect.com/science/article/abs/pii/S026288561730046X
[Accessed May 2020].

Zhang, W. et al., 2017. *Martial Arts, Dancing and Sports Dataset.* [Online]
Available at: http://visal.cs.cityu.edu.hk/research/mads/
[Accessed May 2020].