CNN-LSTM Architecture for Human Action Recognition Using Skeletal Representation

Human Activity Recognition (HAR)

HAR is the problem of identifying actions, activities or events that are performed by humans.

HAR may be divided into

a) segmented recognition, segmenting the video into input clips that contain exactly one activity.

b) continuous recognition, wherein the goal is to detect and classify actions within a video, wherein several parts may not contain action while starting and ending points of action should be detected.

Approaches for HAR include many different sensorial inputs, like RGB cameras and Depth cameras.

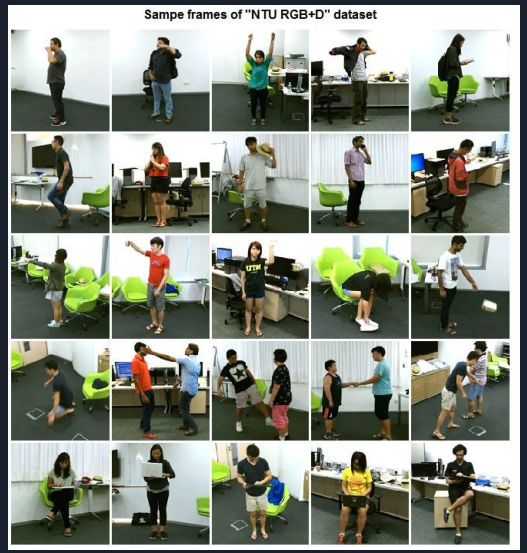
Skeletal Joints information is extracted from RGB or/and Depth information.

NTU RGB+D Dataset

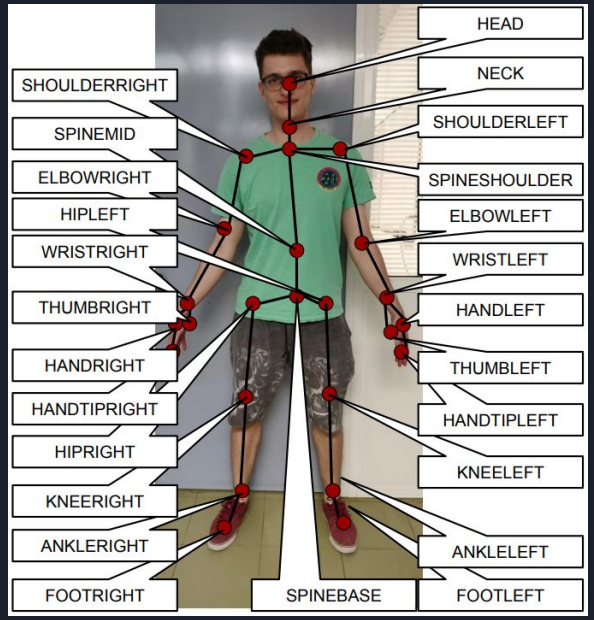
For network training and evaluation, our method was tested on the 60 classes from the NTU RGB+D

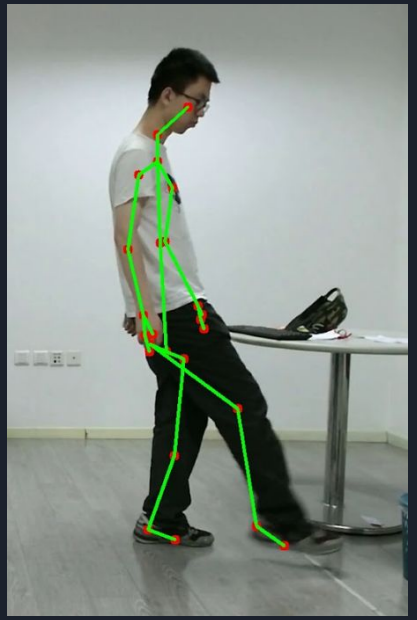
NTU RGB+D is a large-scale benchmark dataset for 3D Human Activity Analysis. RGB, depth, infrared, and skeleton videos for each performed action have been also recorded using the Kinect v2 sensor.

It contains around 67K videos of actions from multiple angles and multiple people.



Kinect V2 Skeletal Information





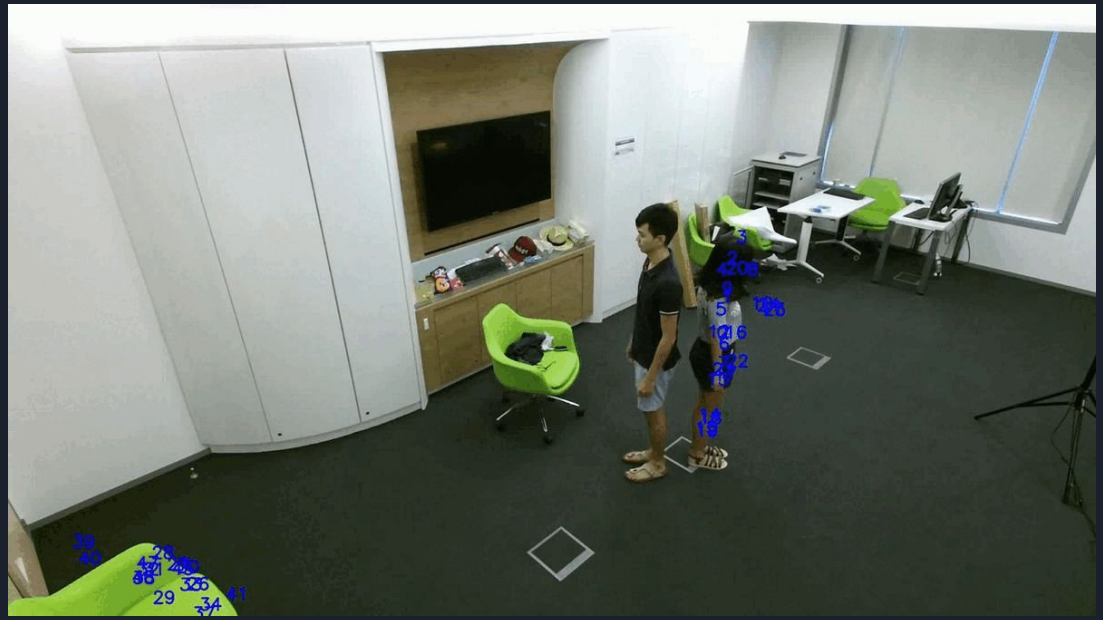
Pre Processing

Firstly a check if each file contains completed information, by checking each line in a file to look if there are lines with zeros, and if a threshold of 15 zero lines if passed we remove the file.

As a second step, we check the points between two persons if there's any joint that passes between two persons.

Finally a step to center the beginning of the action by setting the Spinbase ta 0-0-2.5 and move the rest of the joints relative to the first.

Problematic File



Smoothing (Failed)

A step to smooth the skeleton movement by using the bezier curve was tried.

Before

After

Skeletal Representation

Every joint has an 3D information.

To Represent this into an image we allocate each dimension into the 3 colors of RGB. The three dimensions are allocated as such, x will be the Red Color, y the Green and z the Blue.

To address the problem of temporal variability between actions and between users, a zero padding step was impose to fill the image to 200 frames, this covered the ~99% of the dataset.For actions that had more than 200 frames, a linear interpolation step used to manually reduce the frames to 200.

We have N joints for each of 200 frames. Every Color on each pixel is calculate as follows.

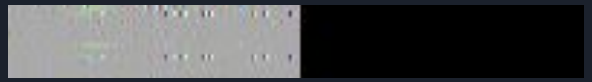
$$C(i,n) = xi(n + 1) - xi(n) \text{ where } i = 1 \dots N \text{ and } n = 1 \dots 200$$

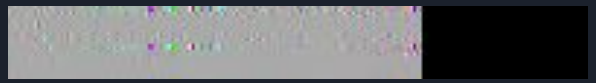
The final image has size of N x 199

The method first introduced by Vernikos et al., 2019a. [Vernikos, I., Mathe, E., Papadakis, A., Spyrou, E., and Mylonas, P. (2019a). An image representation of skeletal data for action recognition using convolutional neural networks.]

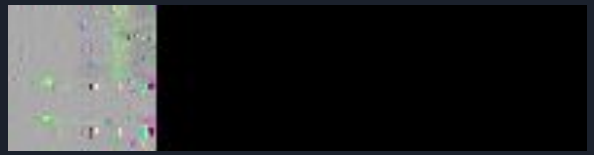
Skeletal Representation

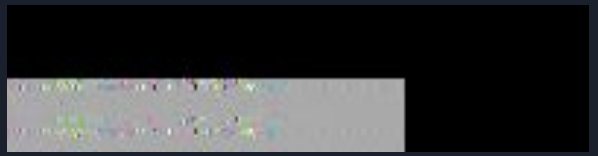
A) Original Design





B) Zero padding

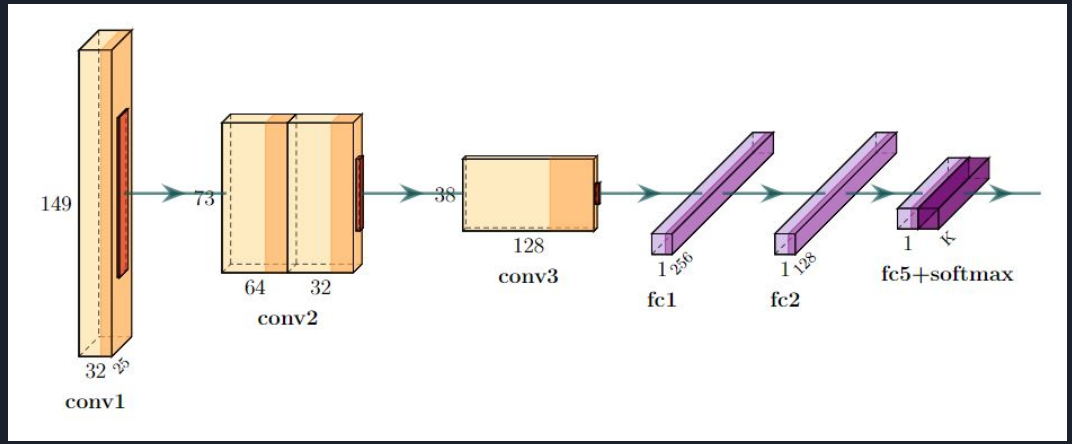




C) Zero padding 2 Persons

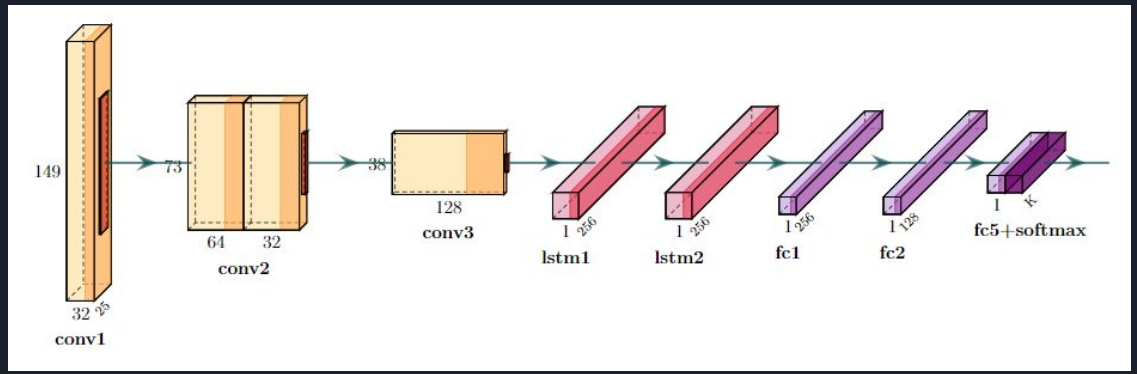
Convolutional Neural Network

I. Convolution, 32 Kernels
II. Max Pooling, 2x2
III. Convolution, 64 Kernels
IV. Convolution, 64 Kernels
V. Max Pooling, 2x2
VI. Convolution, 128 Kernels
VII. Max Pooling, 2x2
VIII. Dense, 256Kernels
IX. DropOut 0.5
X. Dense, 128 Kernels
XI. Dense, Output



Convolutional Neural - Long Short-term Memory Network

I. Convolution, 32 Kernels
II. Max Pooling, 2x2
III. Convolution, 64 Kernels
IV. Convolution, 64 Kernels
V. Max Pooling, 2x2
VI. Convolution, 128 Kernels
VII. Max Pooling, 2x2
VIII. TimeDistributed Flattening
IX. LSTM 256 Kernels
X. LSTM 256 Kernels
XI. Flattening
XII. Dense, 256 Kernels
XIII. DropOut 0.5
XIV. Dense, 128 Kernels
XV. Dense, Output



Results

Model	CrossView	CrossSubject
CNN	54.2	51.3
CNN - LSTM (Proposed Method)	66.2	57.0
CNN - LSTM 2 Persons(Was broken)	60.6	53.4

Accuracy was used for the evaluation.

CNN-LSTM Architecture for Human Action Recognition
Using Skeletal Representation

Thank You
Dimitrios Koutrintzes