Deep Learning based Action Recognition System using Skeletal Data

Fatema Jannat
University of North Carolina at Charlotte
fjannat@uncc.edu

Abstract

Action recognition is a very popular area of work in the Computer Vision field and is widely used in different HCI systems, video surveillance systems, and so on . In recent days, the use of skeleton data instead of RGB videos in action recognition tasks is increasing since it shows remarkable improvement even in complex backgrounds. The privacy concern is the main motivation behind the analysis of skeletal data for action recognition task, since unlike the RGB data the Skeleton data is basically a sequence of frames of a set of points and each point represents coordinates of human joints. Converting the raw skeleton data from NTU dataset to a numpy tensor array format, I have applied image based classification algorithm to recognize the actions. For this I have used virtual radar to generate the spectrograms from the skeleton data. My goal is to develop a model with optimized parameters to classify medical related actions with higher accuracy.

1.1. Development Plan

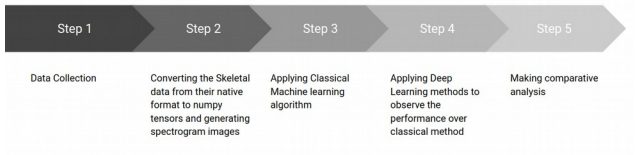

Fig: The flow chart of the development plan

I have organized my flow of work as follows,

1st Step: Getting the dataset and applying data preprocessing techniques on it

2nd Step: Converting the skeleton data to spectrogram images since I am going to use image based deep learning algorithms.

3rd Step: Applying classical machine learning algorithms like SVM to observe the performance

4th Step: Applying Deep Learning methods to observe if there is any significant difference between the classical and deep learning one.

5th Step: Applying different methods to boost the performance result such as-

- Image resizing
- Augmentation
- Using different network
- Tweaking network

1.2. Dataset

Dataset is taken from Rapid-Rich Object Search Lab. "NTU RGB+D" has in total 60 action classes and 56,880 video samples. These contain RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos for each sample. Here I have taken only the skelatal data. This NTU dataset has total 60 action classes, among those 9 are medical related, which are filtered from the whole dataset.

A41: sneeze/cough	A44: headache	A47: neck pain
A42: staggering	A45: chest pain	A48: nausea/vomiting
A43: falling down	A46: back pain	A49: fan self

Table: Medical related action list

After choosing only the medical related action classes, Total Training Sample: 5678 , Total Validation Sample: 2488 , Total number of classes : 9

1.3. Tools

I have been working on the python environment to develop the model. Pytorch is very popular library for computer vision application field and widely used for

image based classification tasks. I am using this Pytorch library for my project. All my scripts are written in Jupyter IDE.

1.4. Methodology

Here virtual radar technique is used to generate spectrograms. The skeleton data that are taken from NTU can be represented by spheres and ellipsoids. Calculating the spheres and ellipsoids the radar signal propagation can be obtained. The ellipsoids equation ,

$$\left(\frac{x-x_0}{a}\right)^2+\left(\frac{y-y_0}{b}\right)^2+\left(\frac{z-z_0}{c}\right)^2=1$$

x0, y0, z0 are the center of ellipsoid and a, b, c are the semi principal axes length. The radar signal reflection which is defined as Radar Backscatter (RCS) is expressed as,

$$RCS=\frac{\pi a^2b^2c^2}{((a\sin\theta\cos\phi)^2+(b\sin\theta\sin\phi)^2+(c\cos\theta)^2)^2}$$

Here θ is the angle between the ellipsoid's z axis and the radar's receiver direction. φ is the angle between the ellipsoid's x-axis and the radar's receiver direction.

Radar data's complex value is expressed as,

$$Phase=\sqrt{RCS}\times e^{\frac{-j4\pi d}{\lambda}}$$

The distance between the center of ellipsoid and the radar is d, and λ is the wavelength. These two parameters are tuned in the validation set to find the optimal value.
Then Short-Time Fourier Transform (STFT) is applied to the radar signal to generate the spectrograms.

The edges from the skeleton data is converted to ellipsoids, then the reflection of ellipsoids are calculated and summed up. Then STFT is applied to generate the spectrogram.

The generated spectrogram can be considered as images so thereby image based classification can be applied.

2. Work Flow

2.1 Generated Spectrogram

Applying STFT the spectrogram is generated which is viewable as image. Here the x-axis is the object velocity and y-axis is the time. And the pixel intensity explains the energy of the object (the amount of signal that is reflected by the object). Figure1 shows the generated spectrogram for each of 9 classes.

2.2 Classical Method (SVM)

Support Vector Machine (SVM) is a very classical method for regression and classification. SVM is a discriminative classifier that outputs an optimal hyperplane to categorize the labeled data points.

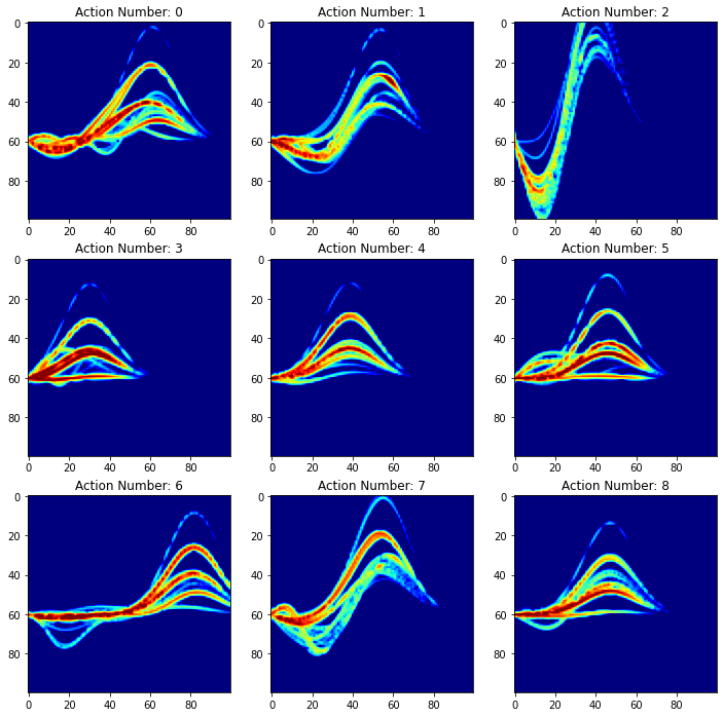

fig: Generated spectrogram for each of nine classes

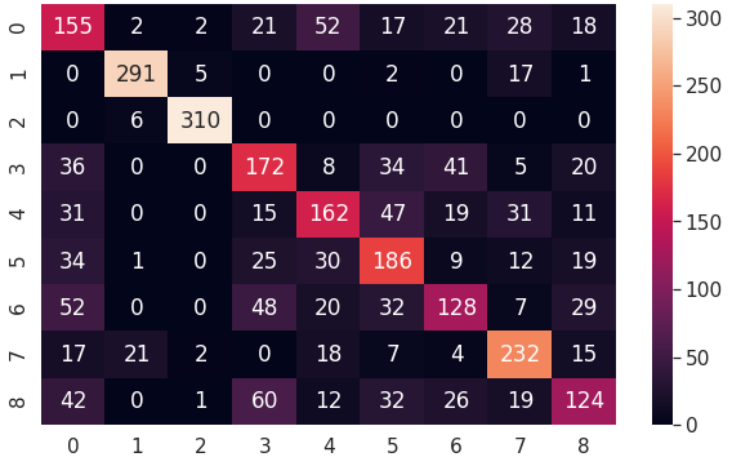

figure: The confusion matrix after applying SVM

2.3 Deep Learning Method (Resnet Architecture)

Resnet-18 is a convolutional network that has 18 deep layers. It is very populr since it reduces the vanishing gradient problem. It is trained with millions of images from ImageNet database and can classify 1000 image categories. So it is already learnt with huge rich representation. This takes 224x224 images as inputs.

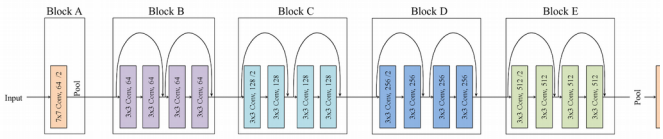

figure: Resnet 18 Architecture

reference: https://www.mdpi.com/2072-4292/11/12/1417/htm

2.4 Data Augmentation

Deep learning is a data hungry system. It requires a lot of data to train the data. In this case the data augmentation technique plays a very supporting role to increase the dataset by applying random transformation to images, thereby increasing the size of data set.

I have applied two augmentation techniques,

- Time Masking
- Frequency Masking

Time Masking:
In time masking, a certain length of time is masked from the spectrogram. The start point is set randomly to make the variation of images and the width of the mask is chosen from 0 to a specific width which also set randomly.

Frequency Masking:
Like the time masking, a certain length of frequency is masked from the spectrogram. The width, start point is set just like the time masking technique.

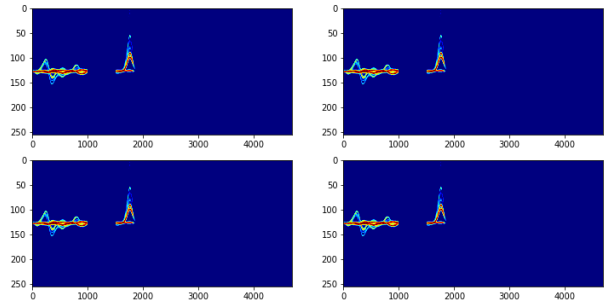

figure: Spectrogram after applying time masking

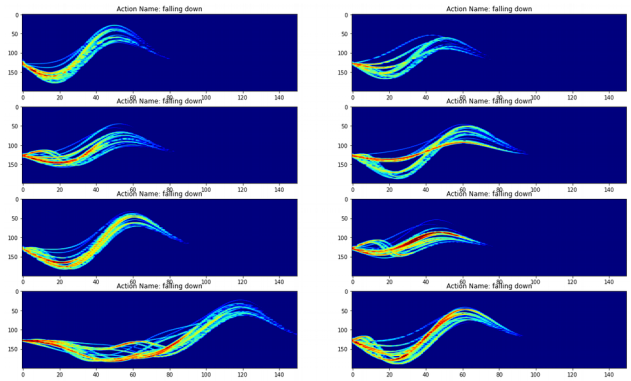

figure: Spectrogram for a single action after removing spaces and resized

2.5 Image resizing:

The image that was generated is 4800x255 in size, from that image the unwanted blank spaces are removed as they were not carrying any information and then resized to 200x200.

2.7 Resnet Fine Tuning:

Fine tuning is the process of utilizing the transfer learning technique where it takes a model that has been trained for one task, and utilizing the model for another similar task just tuning the model a little bit. It allows to use the learned model instead of building a new model from scratch.

Fine tuning can be done by removing any layer or by adding layers.

Freezing any layer is another thing that is done for fine tuning. Freezing any layer means, for that layer weights will not be updated. Weight will remain the same as it was for the previous task.

2.6 Parameter Tuning:

- num_pad_frames (number of interpolated frames that to be inserted between consecutive real frames for each of the data sample).
- sigma (sigma value of gaussian filter that is used to smooth the signal)
- Image shape
- Radar wavelength
- Learning rate
- Batch size

2.7 Label Smoothing

To prevent the over-fitting issue the label smoothing is applied.

	SVM	Resnet18	Resnet18 with transfer learning
accuracy	62%	67%	75%

Table: Accuracy for different approach

2.8 Experimental result

In the first approach where I have applied the classical SVM to classify the action classes, using rbf kernal and standardizing the dataset the accuracy was 62%.

Then moving to the deep learning approach, I have implemented the transfer learning technique, for which I took the advantage of the most popular Resnet18 network to train my dataset. Using the learning rate as 1e-1, batch size of 16 the validation accuracy was achieved as 67%.

Then applying the data augmentation technique, label smoothing and tweaking the hyper parameters, the highest accuracy that I obtained is 75%.

3. Conclusion & Future Work

The technique that I am using to generate the spectrogram from raw skeletal data is very novel. Though the accuracy so far is not quite impressive, but there's several corner where I can work to further improve the model. The frame rate can be tuned to observe the effect on accuracy. The network that I am using that can be fine tuned. I will work on these more to make the model more robust and will develop the optimized model for the action recognition.

References

[1] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. 2019

[2] Bag of Tricks for Image Classification with Convolutional Neural Networks; He et al., AWS, 2018

[3] http://rose1.ntu.edu.sg/datasets/actionrecognition.asp

[4] P. Janakaraj, K. Jakkala, A. Bhuyan, Z. Sun, P. Wang and M. Lee, "STAR: Simultaneous Tracking and Recognition through Millimeter Waves and Deep Learning," 2019 12th IFIP Wireless and Mobile Networking Conference (WMNC), Paris, France, 2019, pp. 211-218, doi: 10.23919/WMNC.2019.8881354.