B. THOMAS GOLISANO COLLEGE OF COMPUTING & INFORMATION SCIENCES

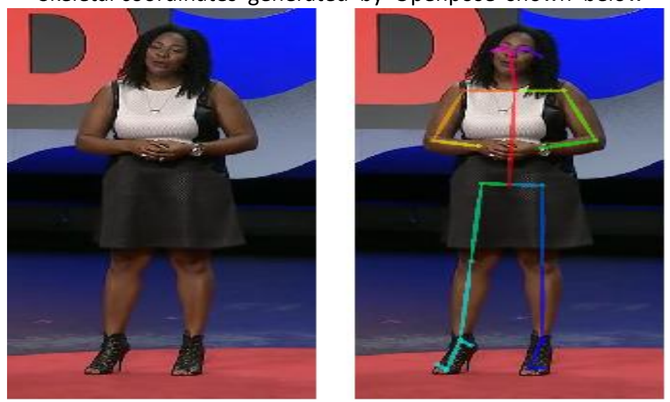
Feature Exploration for Compelling Talks

Vaibhav Kalpesh Joshi – vj3470@rit.edu Advisor: Prof Ifeoma Nwogu - ion@cs.rit.edu

RIT Rochester Institute of Technology

Introduction

- **Goal: To determine if human postures and gestures influence the popularity (likes/views) of a YouTube video**
- **Scope :** Develop a proof-of-concept model using Deep Learning algorithms to classify Ted Talk videos into Low, Medium and High likes using their skeletal coordinates. Analyze results to determine the key body parts (knee, shoulder, elbow) influencing the result.
- Skeletal coordinates generated by Openpose shown below



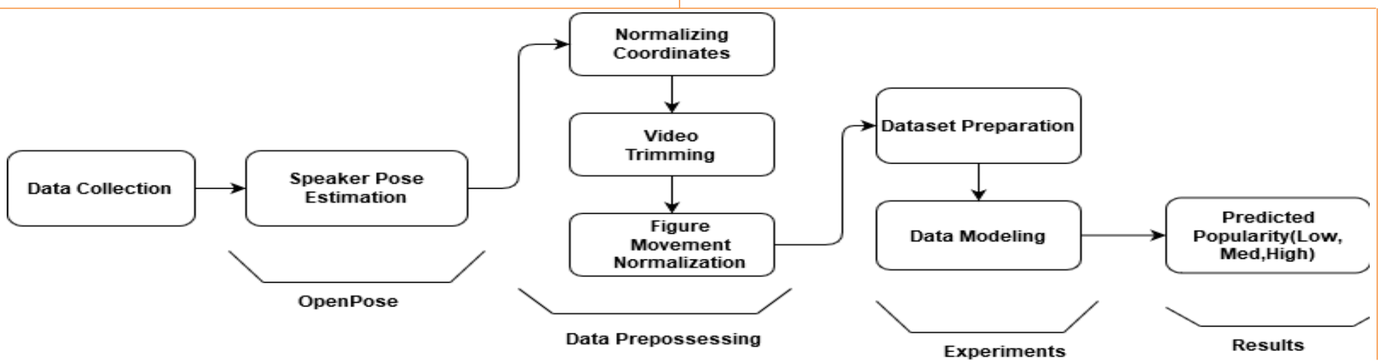
Dataset Collection

For this dataset, 100 Ted Talk videos were collected keeping into consideration the following:-

- Likes/Views across a distributed range (low-medium-high)
- Videos posted before 2017 to account for maturity.
- Videos where the camera is focused on the speaker for most of the frames. This helps to maximize the datapoints.
- Equal distribution between male and female speakers.

These constrains help to negate any potential overfitting in future

System Architecture



Models

Dataset preparation :-

- Split the videos into smaller segments to compensate for less data
- Associate label of video to each of its sub segment
- Ensure videos appearing in training set do not appear in testing

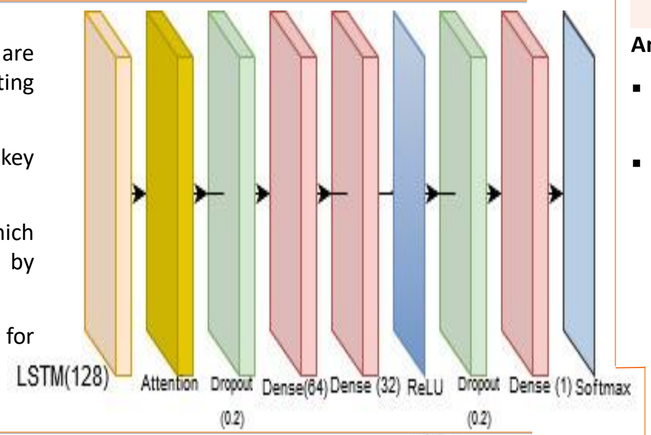
Segment size	Training Samples	Testing samples
30	60722	16021
100	22369	5593

Random Forest

- We create a feature vector of length 85 corresponding to each segment. We compute pairwise distances and individual statistics of points 0,2,3,4,5,6,7.
- We use 10-fold validation for training. We use Scikit-Learn Random Forest implementation; number of trees = 64.

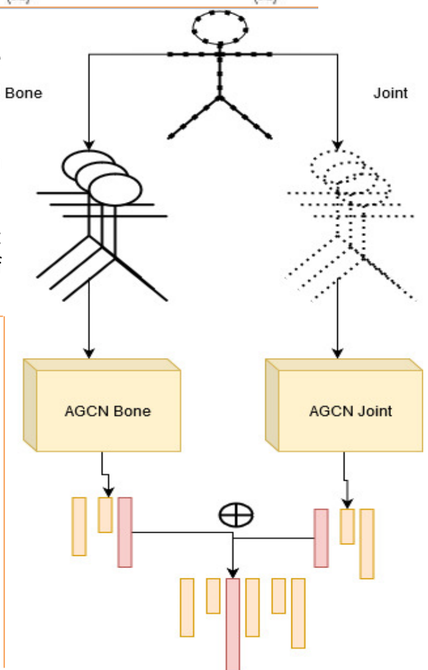
Long Short-Term Memory (LSTM) Network

- LSTM is used as the main layer. LSTMs are useful in sequential data and for accounting gaps in spatio-temporal data.
- Attention layer added for tracking the key joints
- A Dense layer reduces the output to 1D which corresponds to the popularity predicted by the network
- Dropout layers are added to account for overfitting
- Optimizer is set as Adam; batch size is 32



Adaptive Two Stream Graph CNN (AGCN) [1]

- Uses Graphical CNN block on the first order (joint) data and the second order (bone data). Uses two blocks of AGCN for each data stream
- Calculates individual softmax predictions and ensembles them as the final result
- We use the github implementation as a black box and we just change our data structure to meet the input specification of AGCN



Results and Analysis

Accuracy numbers for a 3-class (Low, Med, High) problem

Algorithm	Input Segment Size	Accuracy
Random Forest	30	38.45%
	100	40.40%
LSTM + Attention	30 (timesteps)	39.08%
	100 (timesteps)	38.58%
AGCN	30	43.13%
	100	44.11%
	Whole video	47.62%

Analysis :-

- For the random forest algorithm, we removed specific features out of the preselected feature vector points and reported on the accuracy.
- Accuracy was seen to decrease upon removal of all points or all pairs of points. Shows that individual points may have lesser influence as compared to collective

Joint / Joint Pair	Accuracy after removal
Joint - 2	35.11
Joint Pair - (2,4)	36.03
All individual joints	31.13
All pairwise joints	32.63

Future Work

- Working with a larger dataset (possibly 1000 videos)
- Ablation study on LSTM, AGCN for time steps, compelling features.
- Using Transformer networks.

Conclusion

- In this project, we determine the popularity of Ted Talk videos via classification based on the skeletal coordinates of the speaker.
- We implement random forest, LSTM with attention and use a state-of-the-art Adaptive Graph Convolutional Network.
- We achieve decent accuracy measures with AGCN giving the best result for our dataset. We preliminarily analyze the points that contribute the most to the final output using the random forest results.

References

[1]. Shi, Y. Zhang, J. Cheng, H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in CVPR (2019)