INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY
AND
HONDA RESEARCH INSTITUTE

Translating motion sequences to Vocab Sequences using 3D Convolutional Networks

by

Pranav Sankhe

Supervisor: Dr. Heike Brock

July 2018

0.1 Abstract

Around 5% of the world population suffers from hearing loss. Engaging differently abled people who have hearing problems in everyday conversations requires action based language which we call as sign language. Deaf people use Sign Language to communicate with each other which is often native to the country. Japanese Sign Language is used by around 317000 people. However there are lack of interfaces to translate the Sign Language to **NL** to enable conversation with deaf people and the hearing ones. The hearing people can't grasp the sign langauge with the normal signing speed and it would greatly benefit the deaf people to have the translations of **NL** to sign language. Hence it is essential to have a bi-directional system which translates Sign Language to **NL** and vice-versa. However, **SL** is not a regular language, it doesn't use voice but actions of arms, fingers and facial expression which increases the complexity of the recognition and generation task. developed to recognize SL words or to generate SL animations generally do not account for all aspects of a signed sentence, such as facial expression, natural signing speed, transitions between words and temporal and spatial context information [1], what makes them incomplete and hard to interprete. Hence the recognition and generation of **SL** hasn't been implemented with good enough quality that it can be practically implemented. To represent the multi-dimensional aspects of SL and solve the issues related to its continuous movements, we investigate the use of deep Machine Learning (ML) models, useful for many domains. The translation of Sign Language to Vocab Sequences has been implemented here based on the neural machine translation architecture developed initially for natural language translation. The encoder of the sequence-to-sequence model is inspired from the models developed for action recognition(classification) tasks and modified accordingly for sequence prediction task. The analysis proves that with an extended dataset, the designed trainable model can be put to use to aid the deaf people. Over the course of this two month internship, we investigated two major models for translation of **SL** to **NL**. The overall framework of the two models is a sequence to sequence model designed for **NL** translation. The two models investigate the idea of encoding the motion data[SL] into a fixed representation which can be further used to predict the sentences in **NL**.

This report extensively elaborates on the data analysis, the model architectures, data processing for the input data and details of the implementation in `TensorFlow`. Special attention has been devoted to explain the code for further reproducibility.

Contents

Abbreviations

SL	**S**ign **L**anguage
NL	**N**atural **L**anguage

Chapter 1

Data Analysis

In order to get familiar with the dataset, we tried some pre-processing and visualizations of the data. The data analysis enabled us to select the data format for our architectures.

1.1 Variance based Analysis

We had the motion data captured using the Vicon System which was converted to the `BVH` format. The BVH format ahd the hierarchical structure which represented the human skeletal structure. The joints were constructed and ordered in the form of kinematic chains and each joint had six motion parameters. The first 3 represented the lengths of the individual segments and the other three denote the rotation angles of the joints. We performed varaince based analysis to determine the motion of each joint and which joint is contributing substantially to the overall motion. Fig:1.1 plots variances of all markers througout the signing of one sentence. The indvidual variances of each of the 6 motion parameters are plotted seperately.

1.2 C3D data

The motion capture data was captured using the Vicon system which gives output in the `C3D` format. The `C3D` format shows

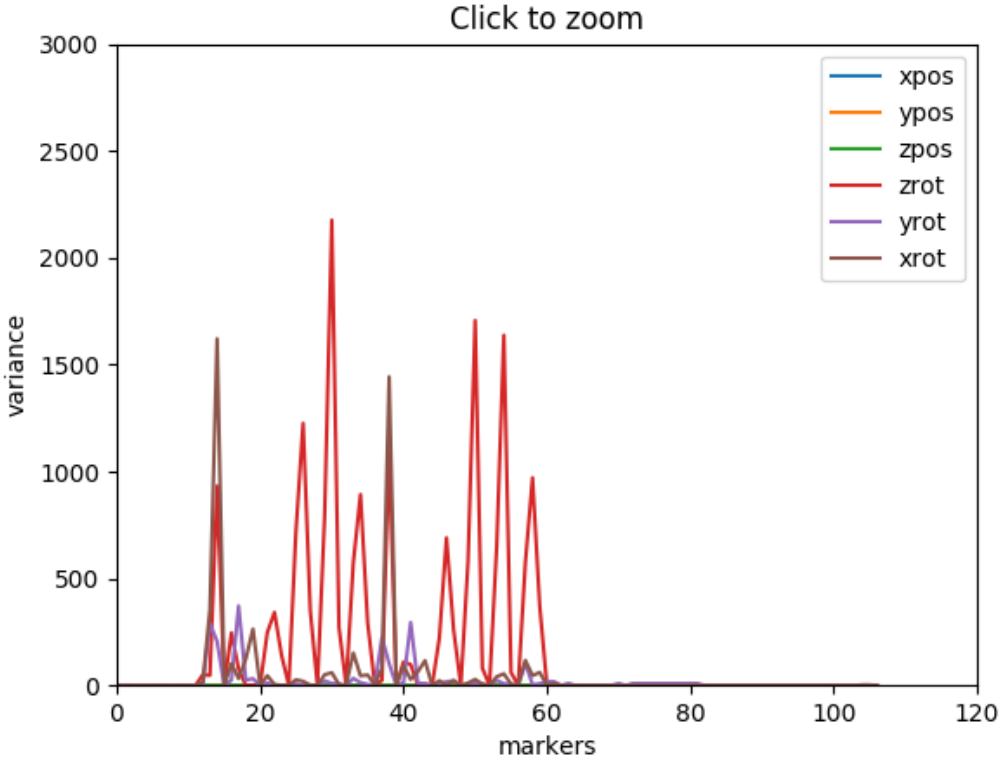

FIGURE 1.1: Variance of individual joints

Chapter 2

Two Stream RNN CNN based Encoder

2.1 Architecture

2.2 Input Data

2.3 Pre-Training

2.4 Results

Chapter 3

Two Stream RNN CNN based Encoder

3.1 Architecture

3.2 Input Data

3.3 Pre-Training

3.4 Results

Chapter 4

Optical Flow Computation

4.1 What is Optical Flow

4.2 Motivation to use Optical Flow

4.3 Optical Flow Code

Chapter 5

Final Model

5.1 Initial Architecture

5.2 ResNet based Architecture

5.3 Training

5.4 Inference

Chapter 6

Future Work

Appendix A

Pre-Training

Appendix B

Sequence to Sequence Model

Bibliography

[1] M. Huenerfauth, *Generating american sign language classifier predicates for english-to-asl machine translation,* Ph.D. dissertation, University of Pennsylvania, 2006.

[2] Sebastian Stober, Avital Sternin, Adrian M. Owen and Jessica A. Grahn *Towards Music Imagery Information Retrieval: Introducing the OpenMIIR Dataset of EEG Recordings from Music Perception and Imagination.* In: Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR15), pages 763-769, 2015.