Body Language Recognition: An Artistic Implementation of Deep Learning

Keyin Wu, Jingxian Xu, Yiming Huang

New York University Shanghai

Author Note

This research was supported by the Dean's Undergraduate Research Fund of New York University Shanghai.

Abstract

Throughout recent years, the world has witnessed a tremendous number of cases where methods of machine learning have successfully bridged the gap of communication between individuals. This study is a preliminary effort to build upon such a trend and establish a well-functioning and dynamic model for body language recognition. Looking past the complex recognition of gesture classes, this study focuses on the implications of the gestures. In this study, we screened, analyzed, and recategorized multiple large-scale human action video datasets to create a suitable motion dataset to use in the training of a Long Short-Term Memory model. Training results of the model demonstrated both possibilities and limitations in the recognition of the implications of commonly used body language. Considering the diversity and complexity of human communication, patterns of the model's output were analyzed and incorporated into a separate artistic visualization which can be used for further presentation.

Keywords: Body Language, Gesture Recognition, Recurrent Neural Networks

Body Language Recognition: An Artistic Implementation of Deep Learning

Introduction

Human action recognition is an actively pursued topic in the applications of human-computer interaction. A wide range of techniques have been developed and used to accurately capture motion information, including skeletal pose tracking models [9] and depth motion sensors such as the Perception Neuron and Microsoft Kinect. Multiple successful efforts have been made to achieve and optimize the recognition of a large variety of gestures [1][2][3][4]. However, in contrast, the recognition of body language is a relatively untouched field. Body language is unique from other methods of communication in that it is universal yet diverse at the same time. Unlike spoken languages, there is no fixed dictionary of definitions for body language. Therefore, connecting body language to an algorithm may not be as intuitive as connecting gesture classes (e.g. sit, stand, wave). Prior works have also shown that interpretation of body language relies both on the movement itself as well as facial expressions. The congruency and incongruency between movement and expressions may result in different interpretations [7]. Such factors can create obstacles in achieving objective recognition results. As we believe that under the circumstances of our research, it is unlikely to achieve genuine objective judgement throughout the study, we choose to work within a broader and more general scope, experimenting with more idealized assumptions.

In this preliminary study, we attempt to explore the possibilities of using recurrent neural networks to interpret versatile body language, relying on 2D RGB videos as input and focusing solely on movement alone. Building upon available large-scale multi-action RGB video datasets [1][2][3][4][8], we create a new dataset of skeletal joint coordinates for commonly used gestures in human interactions. We screen and divide the RGB videos into three categories: *habitual,*

communicative, and *offensive.* The categories are chosen to represent the different degrees of tension and need for attention involved in the movement. Categorized videos are processed using a skeletal pose tracking model [9] to obtain frame-by-frame joint coordinates for each movement sequence. Then the processed data is used as input for our training model. To tackle sequence to sequence recognition, we choose to use a Long Short-Term Memory model that supports feedback connections. The training results show stable and relatively high accuracy in both training and validation. We believe that the results demonstrate promising potential in body language recognition. However, we also foresee many issues that need to be solved. In the following sections we will introduce more details about the dataset, the processing methods, the training process, and our discussions.

A Sequential Skeletal Joint Coordinate Dataset

Source of Data

Multi-modal Gesture Recognition Challenge: This challenge [8] was organized by ChaLearn in conjunction with ICMI 2013. providing RGB images, depth images, skeleton information, joint orientation and audio sources recorded using Kinect. For our study we collected the 2D RGB gesture sequence data.

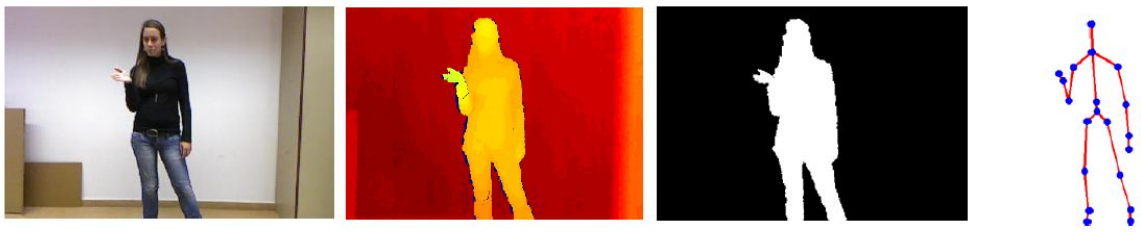

Figure 1. Sample frames from the Multi-modal Gesture Recognition Challenge

The LIRIS Human Activities Dataset: This dataset [4] contains RGB and depth videos showing people performing various everyday activities including making phone calls, chatting, opening and closing doors, etc. For our study, we focused on the RGB video data involving only one individual.

Figure 2. Sample frames from the LIRIS Human Activities Dataset

"NTU RGB+D" Dataset and "NTU RGB+D 120" Dataset: These two datasets [2][3], created by the Rapid-Rich Object Search Lab, contain both RGB videos, depth and skeletal data, and infrared (IR) videos for each sample, concurrently captured by three Kinect V2 cameras. Samples included daily actions, medical conditions, and mutual actions. We selected some RGB video samples but not all from the action classes within the daily action category.

Figure 3. Sample frames from the "NTU RGB+D" Dataset and "NTU RGB+D 120" Dataset

HMDB51 Human Motion Database: This database [1] is a collection of videos obtained from movie clips, public databases, YouTube, etc., containing a total of 6849 video clips which were divided into 51 categories. We screened the clips based on suitable categories and selected those that included clear human body movement of a single individual.

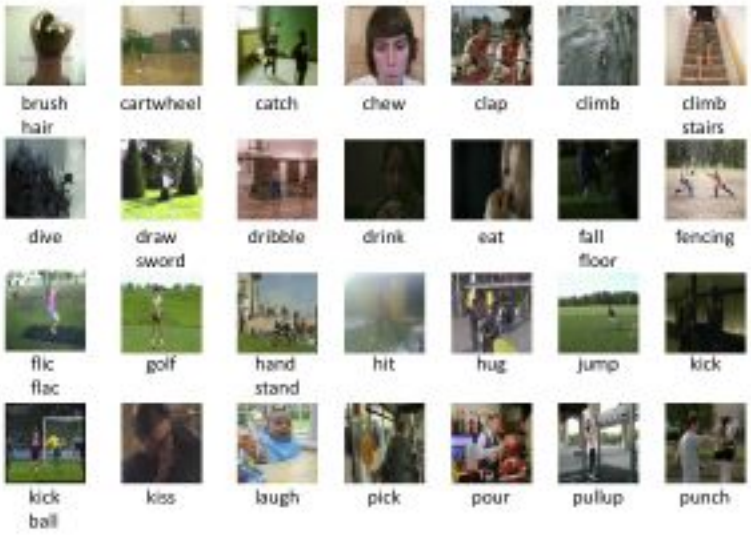

Figure 4. Sample frames from the HMDB51 Human Motion Database

Data Categorization

We selected a total of 3159 RGB video clips and categorized them into three categories under the labels: *Habitual, Communicative, Offensive.* The number of clips for each category is shown in Table 1.

Table 1

Number of Samples for Each Category

	Habitual	Communicative	Offensive
Number of Samples	1050	1332	777

Considering the versatility and subjective nature of body language, we chose to avoid defining meanings of gestures under our circumstances. Instead we decided to utilize a broader way of categorization that investigates the tension and need for attention involved in the gesture. If a gesture can be seen as a common everyday action (e.g. yawning, stretching), it is classified as *Habitual,* which suggests a low level of tension. If a gesture is recognized as a signal or an active medium for communication (e.g. waving, cheering), it is classified as *Communicative*, suggesting a need for further attention in order to better receive the information being transferred. If a communicative gesture shows signs of aggression or violence, it is classified as *Offensive*, representing high tension and serving as a warning to take defensive measures during interactions. This method of categorization not only creates three classes that include diverse yet connected data samples for the model to build its knowledge upon, but also establishes a preliminary way of understanding body language that encourages and supports further interpretation.

Data Processing

Using a real-time pose estimation model, PoseNet [9], we were able to transform the categorized RGB video samples into frame-by-frame skeletal joint coordinates. We adjusted the

model to enable video input, video slicing and JSON output for each single-person sample on a local web server. The model provided coordinate output for 17 individual joints: *nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle.* In our study we focused on video samples that contained only one individual rather than those that involved mutual interactions. Therefore, we did not use multi-person pose estimation which relies on a relatively more complex and heavy-weight algorithm.

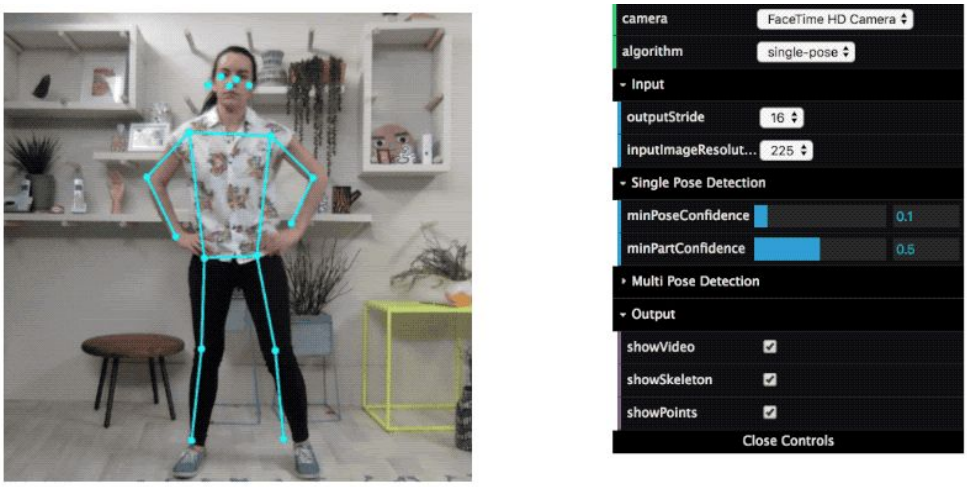

Figure 5. Examples from PoseNet

Training and Evaluation of an LSTM Model in Body Language Recognition

For the training model, we chose to build upon a LSTM architecture, which is a Recurrent Neural Network that utilizes feedback connection, making it highly applicable to projects using sequential data, which in our case is sequential (frame-by-frame) motion coordinates. The model allows long term dependencies to be considered when inputting a JSON sample that contains a sequence of single-frame joint coordinates and provides a predicted label: *Habitual, Communicative, Offensive,* as its output. The 3159 data samples were shuffled and

divided into a 80% training set and a 20% validation set. The baseline gives approximately a 0.32 accuracy rate. After multiple rounds of training with a training epoch of 100, we noticed that the training accuracy and validation accuracy stabilized near 0.68 after approximately 30 passes, which is significantly higher than the baseline. Figure 6 represents the 4th round of training which shows that the training accuracy tends to 0.710928319623718 and the validation accuracy tends to 0.6914191419141914.

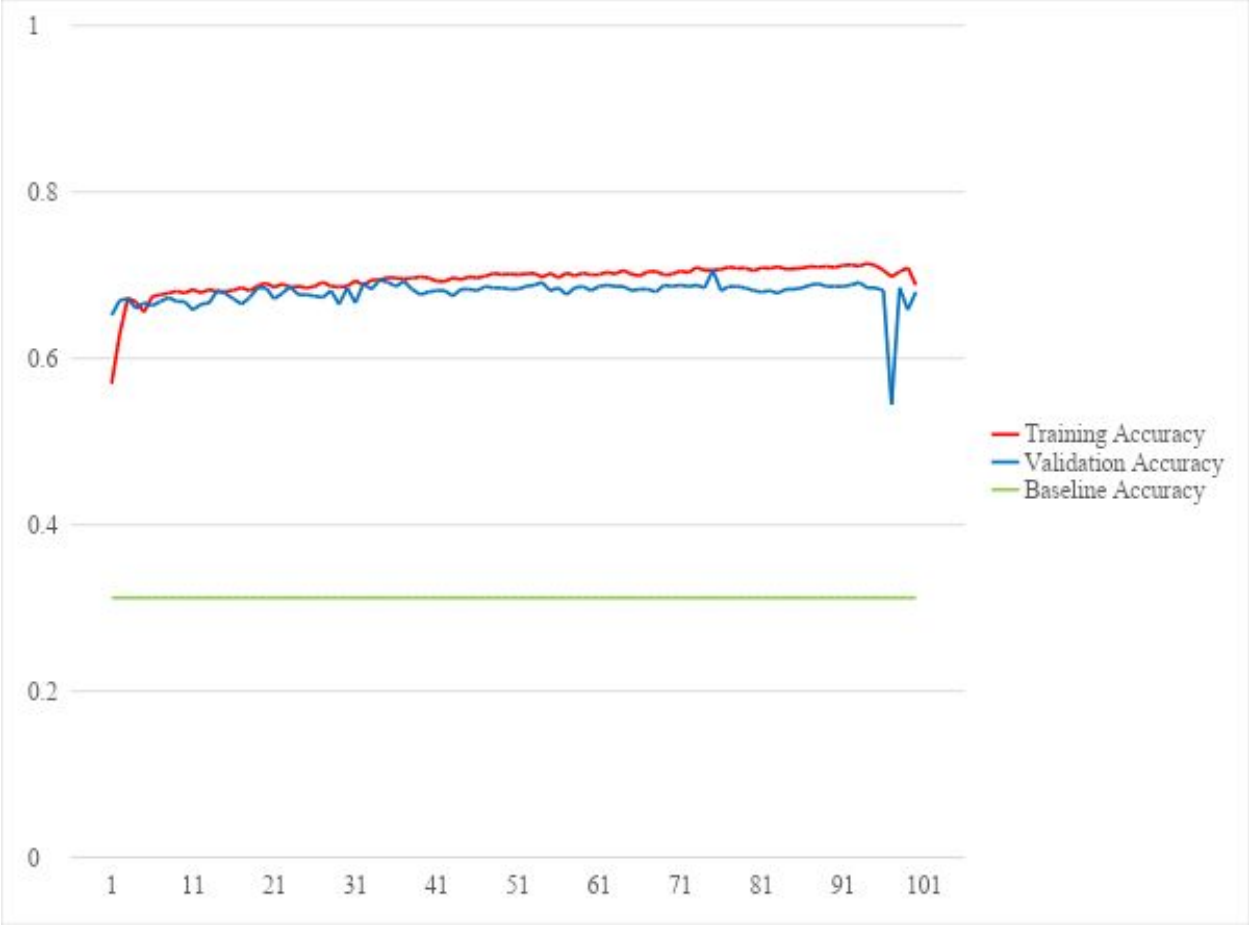

Figure 6. Model Accuracy Rates in the 4th round of training

Visualization

We believe that translating recognition into a vibrant form of art by building a connection between the output of the model and moving visuals adds a valuable touch to our research by giving our findings more vitality and making the findings more accessible and more intuitive to a broader audience. Figure 7 shows sample frames of our visual presentation.

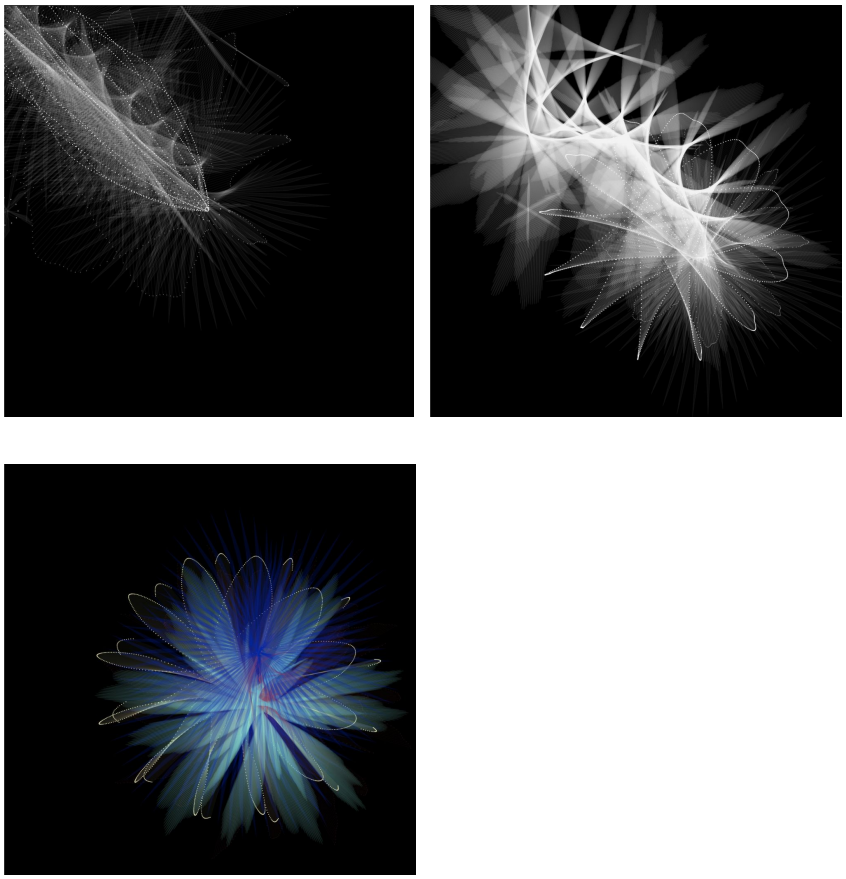

Figure 7. Sample frames from visualization demo

Discussion

Based on the results of this study, we believe that models that are based on RNN architecture, such as the Long Short-Term Memory model show high potential in gesture sequence recognition. Even when working with broad categories that contain a diverse variation of subclasses, the results were quite promising. The challenge, however, in what we experienced

was the categorization of body language itself. We chose a different angle of analyzing body language, focusing on the nature of its use and the level of tension involved. We believe this is meaningful because it can serve as a preliminary signal or trigger for further interpretation of the meanings behind a certain gesture or action. Our choice allowed us to avoid the issue of subjective judgement and interpretation of meanings involved with body language. As widely recognized, body language is by nature a diverse and versatile medium of communication. It is quite difficult to determine a fixed meaning for it and restrict it to a certain framework. Though we initially hoped to explore the limitations of deep learning models when faced with such a difficulty, we have found that, in fact, as humans we have yet to find a good solution, if there even ought to be any, to this issue. We do believe, however, that there is much room for further experimentation, and we hope that our preliminary study can support this exploration.

Acknowledgements

We would like to express our sincere gratitude and appreciation to Professor Jung Hyun Moon for providing us with guidance and encouragement throughout our project.

Portions of the research in this paper used the NTU RGB+D (or NTU RGB+D 120) Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

References

1. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. “HMDB: A Large Video Database for Human Motion Recognition”. *ICCV*, 2011.
2. Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.
3. Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, Alex C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAM*I*)*, 2019.
4. C Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, B. Sankur, “Evaluation of video activity localizations integrating quality and quantity measurements.*” Computer Vision and Image Understanding* vol 127, 2014, pp.14-30.
5. Parton, Becky Sue. “Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence.” *Journal of Deaf Studies and Deaf Education*, vol. 11, no. 1, 2006, pp. 94–101. *JSTOR*, www.jstor.org/stable/42658794.
6. Chaquet, Jose & Carmona, Enrique & Fernández-Caballero, Antonio. “A survey of video datasets for human action and activity recognition.” *Computer Vision and Image Understanding*. vol. 117, 2013, pp. 633-659. 10.1016/j.cviu.2013.01.013.
7. Hanneke K. M. Meeren, et al. “Rapid Perceptual Integration of Facial Expression and Emotional Body Language.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 45, 2005, pp. 16518–16523. *JSTOR*, www.jstor.org/stable/4152458.

8. Multi-modal Gesture Recognition (n.d.) Retrieved from https://www.kaggle.com/c/multi-modal-gesture-recognition/rules.
9. Pose Detection in the Browser: PoseNet Model (n.d.) Retrieved from https://github.com/tensorflow/tfjs-models/tree/master/posenet