Analysing Hand Gestures in Real-World Interactions

Employing gesture coding schemes and machine learning to predict physical features of hand gestures in video footage from a crowded social setting.

Author - Franciszek Latała (F.J.Latala@student.tudelft.nl) / Responsible Proffesor - Hayley Hung / Supervisors - Zonghuan Li and Ivan Kondyurin

Introduction

Hand gestures in social interactions:

- Fundamental form of nonverbal communication [1].
- Hand gestures can assist speech. For instance by emphasising points or referencing objects or concepts [1].
- Hand gestures can also convey meaning of their own, such as (but not limited to) emotions, instructions, semantic meaning (eg. thumbs-up geature) [2].
- Hand gestures play a role in structuring interactions [2].

Motivation:

- Being able to classify and quantify physical features of hand gestures could help in uncovering patterns and relationships with their other aspects, such as the meaning or emotions that they convey.
- Automating such predictions by utilising ML techniques could be a promising approach with potential to accelerate gesture and interaction reserach.
- Gesture coding schemes can capture the multi-facted and complex nature of hand gestures [4]. Thus they were used for gesture coding in this study.
- Gesture classification research generally uses front-facing lab footage not representative of real world social settings. To adress that top-view footage from a social setting was used (ConfLab [3] dataset).

Reserach questions:

- **Main reserach question:**
  - Would it be possible to train a machine learning model to classify complex physical features of hand gestures, using video footage captured in a crowded social setting and annotated according to a gesture coding scheme?
- **Sub-questions:**
  - What physical features of hand gestures to predict?
  - What coding scheme would best fit that task?
  - How to annotate the data?
  - What kind of machine learning approach could be followed?

Approach

Data preparation:

- Out of multiple considered schemes, the M3D [4] gesture coding scheme was selected for video footage annotation, in combination with the ELAN [5] annotation framework.
- Selected video footage from the ConfLab [3] dataset was annotated. More specifically, the trajectory direction of hand gestures (as proposed by M3D [4]) was coded. Due to time constraints slight simplifications had to be made and in he end the following labels were assigned for the left and right hands separtely: ‘up’, ‘down’, ‘forward’, ‘self’, ‘left’, ‘right’, ‘no movement’, and ‘rotation’. Figure 1 presents an example annotation.

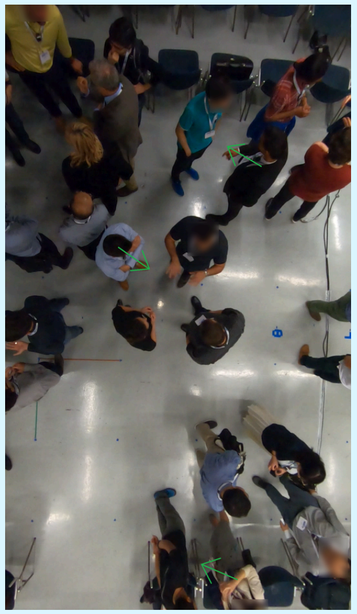


Figure 2: Example frame from the ConfLab [3] dataset, with arrows highliting the varied positioning of participants.

Figure 1: Example annotation in ELAN [5] framework. The right hand is annotated with the ‘right’ label, and the left hand with ‘no movment’ label.

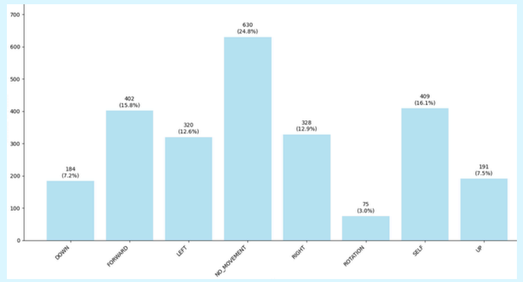


Figure 3: Label distribution in annotated data, irrespective of the hand.

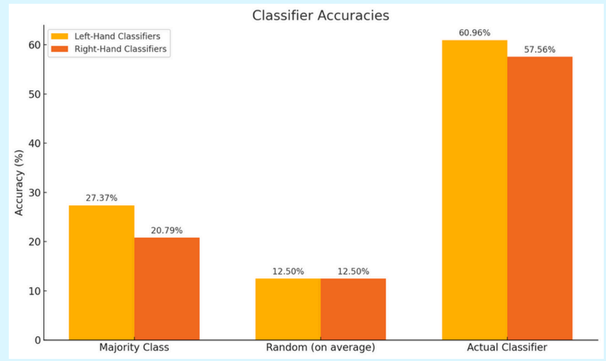
Pre-processing and model training:

- Based on the annotations from the ELAN tool [5], 2 datsets were created - one for the left hand and one for the right hand. As depicted by Figure 3, significant imbalance between the classes in the data was observed. Finally, the left-hand datset consisted of 1346 annotated video segments and the right-hand dataset consisted of 1193 segments. The sets were divided into training, validdation, and test (70%-15%-15%).
- After applying random flips and random shifts to the data, two classifiers based on the VideoMAE [6] video transformer were trained, one for the left and one for the right hand.

Findings

Results and analysis:

- Both classifiers significantly outperformed the baseline models - a majority-class classifier and a random classifier (on average, given 8 classes). This demonstrates that the proposed approach of using ML models to predict complex physical features of hand gestures is indeed feasible.



- An imbalance in the per-class accuracies was observed, with the ‘rotation’, ‘up’ and ‘down’ classes achieving the lowest accuracies. This corespodnds to the imbalance in the prepared datasets.
- Some interesting confusion patterns were observed. First of all there was a lot of ambiguity in the ‘up’ and ‘down’ labels, which corresponds to the simplifications made, since due to time constraints labels for more detailed annotation of ‘up’ and ‘down’ motions were disregarded.
- Almost no confusions between ‘right’ and ‘left’ occured, indicating that the models can distinguish between these two directions no matter the direction a person is facing.

Conclusion:

It is indeed possible to use machine learning and coding schemes to capture physical features of hand gestures in social scenarios. Future research could expand on this approach by mitigating data imbalance and ommiting the simplifications that reduced the granularity of annottations. Such improved approach could be very useful in various areas of gesture and interaction research.

[1] D. McNeil. Hand and mind: What gestures reveal about thought. University of Chickago Press, 1992.
[2] Adam Kendon. Gesture: Visible action as utterance. Cambridge University Press, Cambridge, UK, 2004.
[3] C. Raman, J. Vargas Quiros, S. Tan, A. Islam, E. Gedik, and H. Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. 2022.
[4] Rohrer et al. The multimodal multidimensional (m3d) labeling system. Open Science Framework, 2023.
[5] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: a professional framework for multimodality research. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06), Genoa, Italy, May 2006. European Language Resources Association (ELRA).
[6] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoen-coders are data-efficient learners for self-supervised video pre-training. arXiv preprint, 2022.