AUTHOR: Alexandru Grigore
SUPERVISORS: Hayley Hung, Ivan Kondyurin, Zonghuan Li

TU Delft

Hand Gestures Classification in Crowded Environments

1. INTRODUCTION

Human communication is a complex process where non-verbal cues, such as hand gestures, play a crucial role.

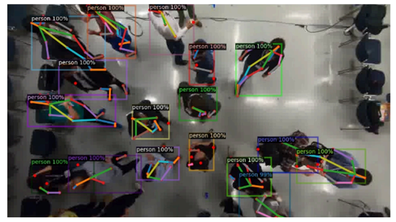
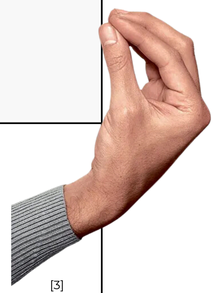


Figure 1. Conflab Dataset visualized [2]



[3]

RESEARCH QUESTION

How accurately can we classify:
- **Gesture Phases**
- **Gesture Units**

using as input:
- **top-view video footage**
- **crowded social setting**

WHY?

Accurate classifiers provide:
- A foundation for **gesture recognizers**
- progress in the **video understanding** field
- The ability to **interpret** conversations **without audio**

2. BACKGROUND INFORMATION

- **Coding schemes =** the practical description of a gesture
- **Gesture Phases =** the stages of a complete gesture
  - **preparation**
  - **hold**
  - **stroke**
  - **recovery**
- **Gesture Unit =** the sequence of gesture phases from when the hands leave rest to when they return to rest.

3. METHODOLOGY AND EXPERIMENT

- **Video Dataset**
  - Based on the **Conflab dataset** [2]
  - Top-view recordings of a 16-minute conference, 5 different camera angles
- **Gesture Representation**
  - based on the M3D framework [4]
  - regards only the prosodic dimension of a gesture
- **Annotation Tool**
  - VGG Image Annotator (VIA) [1]
  - **temporal annotations** for:
    - gesture phases (Figure 2)
    - gesture units (Figure 3)
  - **spatial annotations** for:
    - the area where the person makes the gesture (Figure 4)

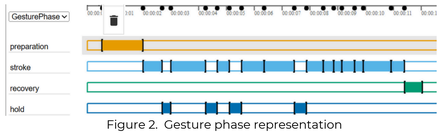


Figure 2. Gesture phase representation

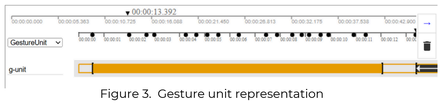


Figure 3. Gesture unit representation

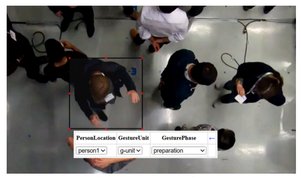


Figure 4. Bounding box of the person

METHODOLOGY AND EXPERIMENT

- **Gesture Classification Model:**
  - VideoMAE [5]
    - pre-trained Vision Transformer
    - designed to work on small datasets
  - Dataset split in train-val-test
    - 70-15-15 split
  - Dataset is unbalanced

- **Gesture Phase classification**
  - Fine-tuning of VideoMAE
  - **Labels**:
    - 4 gesture phases
    - Unknown = arbitrary non-annotated video sections of variable length
  - Strokes are abundant because they repeat within a single gesture unit.

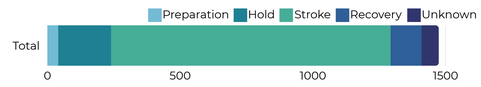


Figure 5. Gesture Phase classifier dataset distribution

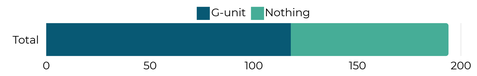


Figure 6. Gesture Unit classifier dataset distribution

- **Gesture Unit classification**
  - Fine-tuning of VideoMAE
  - **Labels**:
    - G-unit
    - Nothing = arbitrary non-annotated video section of variable length

4. RESULTS

Gesture Phase classifier

- High overall accuracy
- Good accuracy for 4/5 labels

- "Preparation" is often confused with "Stroke".
  - Possible causes:
    - The model does not consider the earlier phase
    - "preparation" dataset is too small

Baseline	Accuracy
Uniform Distribution	20%
Majority Class	70%
Our Model	**95%**

Table 1. Phase classifier compared with baselines

Label	Accuracy
Hold	1.0
Preparation	0.43
Recovery	0.89
Stroke	0.98
Unknown	0.93

Table 2. Per label accuracy for phase classifier

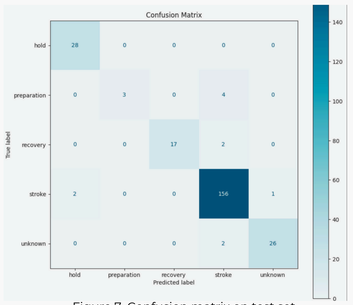


Figure 7. Confusion matrix on test set

Gesture Unit classifier

- High overall accuracy

- Some "G-units" are labelled as "Nothing". Cause:
  - G-units only contain movements that convey meaning

- training results indicate that the dataset might need to be extended

Baseline	Accuracy
Uniform Distribution	50%
Majority Class	60%
Our Model	**93%**

Table 3. Classifier compared with baselines

Label	Accuracy
G-unit	0.86
Nothing	1.0

Table 4. Per label accuracy for unit classifier

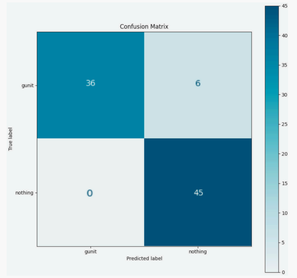


Figure 8. Confusion matrix on test set

5. CONCLUSIONS

- The high accuracies prove the potential of using vision transformers for gesture classification
- The success of the experiment shows the potential of using coding schemes to describe gestures in ML scenarios

FUTURE WORK

- Expand the dataset of both classifiers to verify our assumptions
- Build a recognizer for gesture units and gesture phases
- Expand the recognizer to consider all three dimensions of a gesture described by the M3D framework.

REFERENCES

[1] Abhishek Dutta and Andrew Zisserman. "The VIA Annotation Software for Images, Audio and Video". In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19. Nice, France: ACM, 2019, pp. 2276–2279. isbn: 978-1-4503-68896/19/10. doi: 10.1145/3343031.3350535. url: https://doi.org/10.1145/ 3343031.3350535.

[2] Chirag Raman et al. "ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild". In: Advances in Neural Information Processing Systems. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 23701–23715. url: https://proceedings.neurips.cc/ paper_files/paper/2022/file/95f9ad2e251e90146975890374 50f9bb-PaperDatasets_and_Benchmarks.pdf.

[3] Crace, J. (2010, December 12). What hand gestures mean - in pictures. The Guardian. https://www.theguardian.com/science/gallery/2010/dec/12/what-hand-gestures-mean

[4] P. L. Rohrer et al. The MultiModal MultiDimensional (M3D) Labeling System for the Annotation of Audiovisual Corpora: The Gesture Labeling Manual. Version 2. 2023. url: https://doi.org/10.17605/OSF.IO/ANKDX.

[5] Zhan Tong et al. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. 2022. arXiv: 2203.12602 [cs.CV].