Deciphering the Meaning of Gestures In the Wild

Author: Irene Aldabaldetrecu; Responsible Professor: Hayley Hung; Supervisors: Ivan Kondyurin and Zonghuan Li
Contact: I.AldabaldetrecuAlberdo@student.tudelft.nl

TU Delft

1. Background

- McNeill distinguishes four mutually exclusive types of gesture meaning: iconic, metaphoric, deictic and beat/ non-referential [2].
- This definition of gesture meaning should be expanded to include **pragmatic meaning** (the function of a gesture) and allow the classification of gestures into multiple types.
- The **M3D labelling system** fulfils this criteria and has additional benefits such an online manual for annotation [4].
- Most gesture annotation studies involve collecting data in the lab. It is yet to be explored whether gestures produced in **densely crowded social settings** can be annotated and classified.
- Research has shown that **transformer-based ML models** such as ViViT [1] and VideoMAE [6] can outperform deep convolutional networks in video classification tasks. VideoMAE performs favourably with a high masking ratio (90-95%) and when trained on small datasets (around 3k-4k videos).

2. Research Question

Can we accurately annotate and classify gestures produced in **densely crowded social settings** using the **M3D labelling system** and **VideoMAE**?

3. Data Annotation

- **ELAN** [7] was used to annotated video data from the ConfLab [3] dataset. The M3D system has a publicly available template that can be directly loaded into ELAN.
- The ConfLab dataset provides privacy-protecting low-frequency audio. Without access to co-occurring speech, many **strong assumptions** had to be made. These assumptions include:
  - Rotational hand and arm movements are annotated as metaphoric.
  - Counting gestures are annotated as iconic.
- The M3D system includes 23 pragmatic functions. Due to time constraints and lack of speech content, **pragmatic meanings** were not annotated.
- Each gesture is annotated as having **one label instead of multiple**. This does not comply with the M3D guidelines, but facilitates adapting the dataset for the VideoMAE model used for fine-tuning. The model, which is pre-trained on the UCF101 dataset [5], only outputs one label per input.

4. Data Pre-Processing

- The dataset used for fine-tuning contains **1119 clips** of unique gesture instances. See the dataset distribution in Figure 1.
- The ConfLab videos were manually cropped to fit a single person. **Separate clips for each individual gesture** in each video were generated by trimming them according to the start and end times specified in the ELAN annotation file.
- The dataset was split as follows: 70% for the train set, 15% for the validation set and 15% for the test set.
- The training set was transformed using **uniform temporal subsampling, pixel normalisation, random cropping and random horizontal flipping.**
- Only uniform temporal subsampling and pixel normalisation were applied to the test and validation sets.

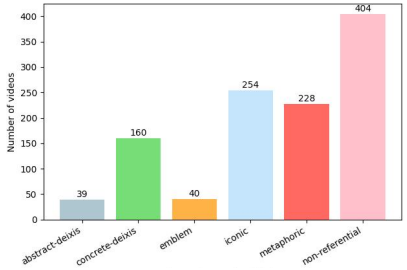

Figure 1. Dataset distribution

Configuration of parameters:
Epochs: 10; Batch Size: 8; Learning rate: 1e-4

5. Results

The model achieved an overall accuracy of 49% on the test set and 48% on the validation set. Figure 2 shows the accuracy of the model per label on the test set. Figure 3 presents the confusion matrix computed on the results obtained from the test set.

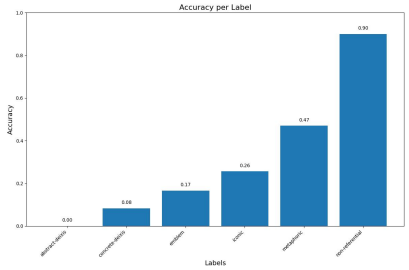

Figure 2. Accuracy per label on the test set.

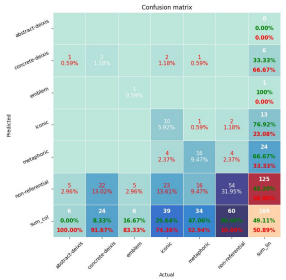

Figure 3. Confusion matrix on the test set.

Due to the imbalanced dataset (see Figure 1), only 41 out of the 169 instances from the test set were classified as categories other than non-referential, and none were predicted to be abstract deictic. The precision, recall, F1-score and ROC AUC metrics are shown in Table 1.

Table 1. The precision, recall, F1-score and ROC AUC metrics computed on the test set.

Class	Precision	Recall	F1-score	ROC AUC	Support
Abstract deixis	0.00	0.00	0.00	0.59	6
Concrete deixis	0.33	0.08	0.13	0.72	24
Emblem	1.00	0.17	0.29	0.88	6
Iconic	0.77	0.26	0.38	0.81	39
Metaphoric	0.67	0.47	0.55	0.79	34
Non-referential	0.43	0.90	0.58	0.75	62
Accuracy			0.49		
Macro average	0.53	0.31	0.32	0.76	169
Weighted average	0.55	0.49	0.44	0.77	169

6. Conclusions

- This approach does not account for a **proper definition of gesture meaning**, as gestures were annotated as having a single label solely based on semantic meaning. Thus, the results only partially answer the research question.
- The classifier has a **strong bias towards non-referentiality** due to the imbalanced dataset.
- This imbalance might have been caused by the strong and potentially wrong assumptions made during the annotation process.
- If **high-quality audio or transcripts** of co-occurring speech were made available, multiple gestures would have been annotated differently.
- The **ambiguity** of gestures, the **visual similarity** between gestures with different semantic meanings and the **small size of the dataset** seem to negatively impact the model's performance.

7. Future Work

- Future work could explore privacy-preserving approaches to **recording high-frequency co-occurring speech**.
- **Pragmatic meaning** should be annotated. Since the semantic and pragmatic dimensions of M3D are independent from one another, two separate classifiers could be trained.
- This task should be approach as a **multi-label, multi-class** classification problem.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. "ViViT: A video vision transformer", 2021.

[2] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[3] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. "Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild", 2022.

[4] Patrick Rohrer, Ingrid Vila-Gimenez, Julia Florit-Pons, Nuria Esteve-Gibert, Ada Ren-Mitchell, Stefanie Shattuck-Hufnagel, and Pilar Prieto. *The MultiModal MultiDimension (M3D) labelling system*, 2023.

[5] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild", 2012.

[6] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training", 2022.

[7] PeterWittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 2006.