Hand gestures classification in crowded environments

Classification of gesture phases in a crowded social setting recorded from top-view angle

Alexandru Grigore[1]
Supervisors: Hayley Hung, Ivan Kondyurin, Zonghuan Li
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Alexandru Grigore
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Ivan Kondyurin, Zonghuan Li, Mark Neerincx

Abstract

Hand gestures play a crucial role in communication, especially in social interactions. This research investigates the viability of using coding schemes to describe hand gestures and how accurately they can be classified in crowded environments by using fine-tuned **visual transformers** such as VideoMAE. The dataset used during training is based on the Conflab dataset and contains top-view video recordings of social interactions in a crowded social setting. The videos are manually annotated for gesture phases (preparation, hold, stroke, recovery) and gesture units. The two classifiers obtain high accuracies after fine tuning, with an overall accuracy of **95%** for the gesture phase classification and **93%** for classifying whether a clip is a gesture unit or not. These findings suggest that the proposed approach is effective in crowded environments and can be adapted for real-time applications.

1 Introduction

Human communication is a complex process, and words are only one aspect of it. The link between non-verbal communication and human interactions is a field which is still researched to this day, but studies indicate that a significant part of human interaction is conveyed through non-verbal cues [16].

Navarro [14] is just an example of books which try to provide a map for human body language. Considering that the main aim of the behavior analysis field is to deduce the semantic meaning of gestures by analyzing patterns in human interactions, it appears to be a scenario where Machine Learning models could excel.

This paper addresses the research question: "How accurately can we classify distinct gesture phases within a Gesture Unit, using as input a top-view video footage captured in a crowded social setting?"

Accurately classifying hand gesture phases in a crowded environment based on a specific coding scheme provides a foundation for further research into behavior analysis in natural settings. While extensive research exists on hand gesture classification, most of the studies focus on a setup that involves a single person being recorded from a front view angle. The purpose of the research is to determine whether the single person approaches can be adapted for a crowded environment setup, recorded from a top-view angle.

If gesture phase classification proves to be effective, the classifier can be used to further build a recognizer. Being able to recognize gesture phases in a crowded, natural setup opens the door for further analysis of interactions between people. For example, some hand gestures, such as shaking hands or giving a thumbs-up, often occur as reactions to the interlocutorâs gestures. This responsiveness can allow for deeper semantic research into gesture interactions, such as studying how a handshake can signify agreement or closure, or how a thumbs-up can indicate approval.

Additionally, we can consider the benefits that gesture recognition can bring to the field of video understanding. A crucial task in video understanding involves comprehending human actions and interaction which are depicted in videos, making behavior analysis techniques essential for achieving effective video understanding and interpreting the content [28]. Demonstrating that coding schemes can be adapted for crowded environments with suboptimal camera angles opens the door to implement them to comprehend human actions depicted in videos.

Identifying and understanding people's gestures during a social interaction can provide major insights in the way humans interact. The semantic meaning of gesture can help us understand nuances in speech without actually hearing what is being communicated. This would contribute to the ongoing work on privacy-sensitive approaches to social behavior analysis that rely on body language and less on audio, where the recording of private conversations can be considered too invasive [24].

2 Background and Related Work

The following section provides an overview of the fundamental concepts and previous research that form the basis for our study on gesture classification and analysis. This includes various approaches to representing gestures, the frameworks used for annotating and categorizing these gestures, and methodologies for training models to recognize and classify gestures accurately.

2.1 Gesture Representation

To classify or recognize gestures, a clear definition of what constitutes a gesture is needed. Scholars have developed multiple approaches to describe them.

McNeill [12] describes 4 categories of hand gestures: beat, diectic, iconic and metaphoric. Beat gestures refer to rhythmic movements of the hand, diectic movement regards gestures in which you indicate something, iconic gestures bear a close relationship to the semantic content, and lastly, metaphoric gestures which symbolize an abstract idea or concept [12]. While this method provides an ability to capture the rich and varied nature of hand gestures in communication, its implementation becomes complex since multiple gestures are usually overlaid upon one another.

Lausberg and Sloetjes [10] introduces the concept of Neurogens, which consists of three modules: kinetic, bimanual relation and functional. The kinetic module describes the dynamics and trajectory, bimanual relation coding focuses on the relation between the two hands, and the functional module focuses on the practical aspect of the gesture, such as to describe emotion, to emphasize and so on [10]. While this method is a comprehensive approach to analyzing hand gestures, the ambiguous separation between the three modules can make the annotation process labor-intensive and time-consuming.

Ekman and Friesen [3] takes an approach in which they describe a non-verbal behavior by taking into consideration the origin, usage, and coding. Thus, Ekman and Friesen proposed categorizing gestures in five different categories: emblems, illustrators, manipulators, regulators, and emotional expressions. While the detailed nature of the framework allows for a nuanced understanding of behavior from a linguistic perspective, it does not appear to be the best candidate automated solution for gesture classification due to the overlapping nature and potential ambiguity of the 5 gesture categories.

Rohrer et al. [20] employs a multimodal methodology to dissect the significance of a gesture into three distinct dimensions: form, prosody, and meaning. The form aspect focuses on the nature of the movement, the prosody regards the organization structure of a gesture, and the latter regards the semantic and/or pragmatic meaning associated with the movement. An important advantage of this structure is the fact that there is little overlap between the three classes.

2.2 Gesture Classification Methods

Video understanding is heavily researched due to the rapid growth of online video content and the need for advanced tools to interpret it [21]. One of the most crucial tasks in video understanding involves comprehending human actions depicted in videos, making behavior analysis techniques essential for achieving effective video understanding and interpreting the overall video content.[28] Thus, the machine learning models considered as a possible solution for the classification of hand gesture phases in this report are some of the most popular approaches in the video understanding field.

A possible option for the classification problem is the usage of 3D Convolutional Neural Networks (3D-CNNs). Hedegaard and Iosifidis [8] introduces the concept of Continual 3D Convolutional Neural Networks where videos are being processed frame-by-frame rather than by clip. One of the possible usage scenarios mentioned for this approach is real time surveillance camera processing, which is similar to the top-view footage scenario used in the paper. Molchanov et al. [13] uses 3D-CNNs for hand gesture recognition and shows them to be performant for such use cases. Thus, considering the successful experiments conducted using 3D-CNNs in the two cases, they can be considered a plausible option for the scenario of hand gesture recognition in crowded environments.

Yu, Qin, and Zhou [27] proposes a different approach, one where 2D convolutional Neural Networks are combined with feature fusion to achieve dynamic gesture recognition. While 3D-CNNs consider both spatial and temporal features, the complexity of the networks can lead to low efficiency of algorithm. To address the limitation of 2D-CNNs in understanding temporal dynamics, they incorporate optical flow key-frames, thus the model captures motion information. Therefore, the solution they propose consists of using optical flow key-frames and dual-channel 2D CNNs with Squeeze-and-Excitation blocks. The solution seems promising, achieving high accuracy while ensuring low network complexity on the Cambridge Hand Gesture dataset and Northwestern University Hand Gesture. Nevertheless, their experiment does not consider at all a crowded environment as in our use case. Thus, the viability of this solution for hand gesture recognition in a crowded environment remains an assumption until proven effective.

Tong et al. [22] introduces VideoMAE, where the concept of masked autoencoders is expanded from images to videos. The model is designed to be a data-efficient learner for self-supervised video pre-training, achieving strong performance even on relatively small datasets. Performance-wise, the paper shows that VideoMAE achieves 87.4% on Kinetics-400 and 75.4% on Something-Something V2, without using any additional data beyond the training set.

3 Approach

In the previous subsection, a few of the possible solutions to the three subproblems of the hand gesture classification problem were presented. This section will explain the chosen approach to solving each problem.

3.1 Gesture Representation

For the purpose of the project, M3D was chosen as a gesture scheme. It's little to no overlap between the three dimensions allowed our research group members to split them one each and research its viability to be used to build a hand gesture classifier. An important factor that contributes to choosing M3D as our gesture scheme is the existence of an extensive training program on how to annotate each aspect of a gesture depending on the dimension. This proved very helpful in building a dataset with consistent annotations.

My chosen dimension was prosody, due to its focus on the structure of the gesture itself rather than its interpretation. The prosodic dimension consists of the following concepts: Gesture Units, Gesture Phases and rhythmic properties. Due to the time constraint of 9 weeks to conclude the research, I choose to focus on classifying gesture phases and gesture units. Thus, Rohrer et al. [20] lists the following 4 phases which can be found in a gesture unit: preparation, hold, stroke and recovery. Intuitively, preparation refers to the movement done to reach the start of the stroke, the stroke represents one hand movement purposed to be a gesture, hold represents a small pause during a gesture unit, and recovery refers to the movement done to go back to the rest position. A Gesture Unit starts at the moment when the person breaks the rest position until the moment it goes back to rest.

3.2 Annotation Tools

One of the most popular annotation tools for temporal segments is ELAN. ELAN is focused on providing the user the possibility to add any textual annotations to audio and/or video recordings [25]. Rohrer et al. [20] recommends the usage of this framework for a single person and a front view video recoding scenario, as in the examples provided in the M3D annotation training.

Covfee is another annotation tool focused on providing a proper experience for continuous video annotation. By continuous annotation, they mean offering the possibility to annotate while watching the video. Covfee specializes in continuous media annotation, but it lacks some basic features of image annotation like bounding box or polygon annotation, which can be hard to do it as a user on a live video feed [18].

Dutta and Zisserman [2] proposes VGG Image Annotator (VIA) built for video and audio annotation. Even though the M3D training provided by Rohrer et al. [20] recommends using the ELAN annotation tool, I considered VIA to be more suitable for the gesture classification use case due to the nature of the video recordings from the dataset. While ELAN can be promising for the annotation of gestures on videos where only one person is being recorded, VIA allows describing gestures through temporal labels, while also associating them to a spatial region described by a bounding box.

The bounding boxes can help solve two possible issues during training. The first issue is that to train some ML models, videos have to be cropped so that only the person doing the gesture remains in the clip. Thus, the bounding boxes can be used to ease the clip creation procedure. The second issue is that if the full frames are used for training, with multiple people who are doing multiple gestures, the model might need a large amount of data to properly understand whose person's movement is annotated. Thus, by associating each gesture to a bounding box, the training process could be improved. Bellow, **Table 1** presents the key similarities and differences between the three annotation frameworks mentioned throughout the paper.

While the two issues could be solved using the skeleton annotation provided by the Conflab dataset, I considered the usage of bounding boxes more accessible if the dataset will be expanded further using automated or semi-automated solutions. The skeleton annotations lead to accurate person identification but are complex to manually annotate. The bounding box annotation of the persons could be done by using an ML model such as YOLO [9]. Yang [26] showcases how YOLOv5 can be used to detect the bounding box of people in a

Feature	**ELAN**	**Covfee**	**VIA**
Multi-tier Annotations	✓	×	✓
Audio-Video Synchronization	✓	✓	✓
Live Annotation	✓	✓	✓
Time Segment annotation	✓	✓	✓
Spatial segment annotation	×	×	✓

Table 1: Comparison of annotation frameworks: ELAN, Covfee, and VIA

multi-person video recording. Therefore, even though bounding boxes need to be manually annotated for the experiment, demonstrating that bounding boxes are enough for accurate gesture classification will ease future work, enabling the use of automated methods for larger datasets. Additionally, we opted for manual annotation because the created dataset is small, and we considered that it was not worth the time and effort to configure and fine-tune YOLO for this initial phase.

3.3 Gesture Classification Model

Throughout this project, the chosen model to fine-tune my gesture classifier is the one introduced by Tong et al. [22], VideoMae. Besides the high accuracy on video understanding datasets, the vast documentation on this visual transformer was another reason to choose it for the experiment. MCG-NJU [11] provides a Jupyter notebook which showcases how VideoMae can be fine-tuned on an openly available video understanding dataset [22]. Thus, the open-source code available on GitHub, along with a detailed guide on the fine-tuning process, made VideoMAE an optimal model choice for this project.

Compared to 3D-CNNs, VideoMAE has some clear benefits. 3D-CNNs are good at capturing local features in videos and are commonly used for action recognition, but they often need large sets of labelled data and preprocessing. On the other hand, VideoMAE uses transformers to understand global context and long-range patterns, making it effective even with smaller datasets. Considering that the gesture annotation is done manually, opting for a data-efficient option seems the proper choice. However, it should be noted that VideoMAE needs more computational power and is more complex to set up compared to 3D-CNNs, which are usually simpler and faster to train. Nevertheless, since the model was only fine-tuned, the computational power required was not a significant concern.

4 Experiment

The following section describes how the previously described methods were employed for the training of the classifiers. The code developed for annotation parsing, data preprocessing and fine tuning is uploaded on Github at Grigore [5].

4.1 Conflab Dataset

The project is built around the Conflab dataset provided by Raman et al. [19]. This dataset contains video recordings of a 16-minute conference from five cameras placed at different corners of the room, all showing top-down views. Around 50 people come and go during the recording, talking to each other. The dataset also includes skeleton annotations of the people and low-frequency audio recordings of their conversations. Figure 1 shows a frame from the footage with the skeleton annotations.

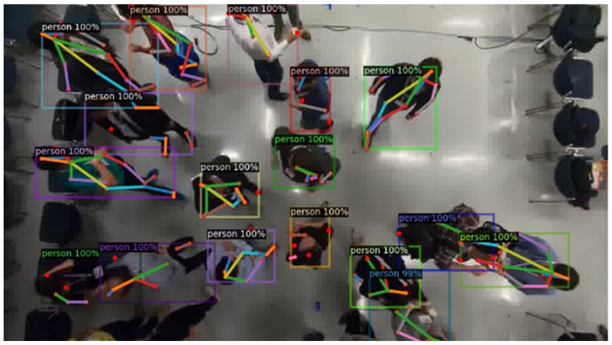


Figure 1: A top-view image from the Conflab dataset with skeleton annotation [19].

4.2 Data Annotation

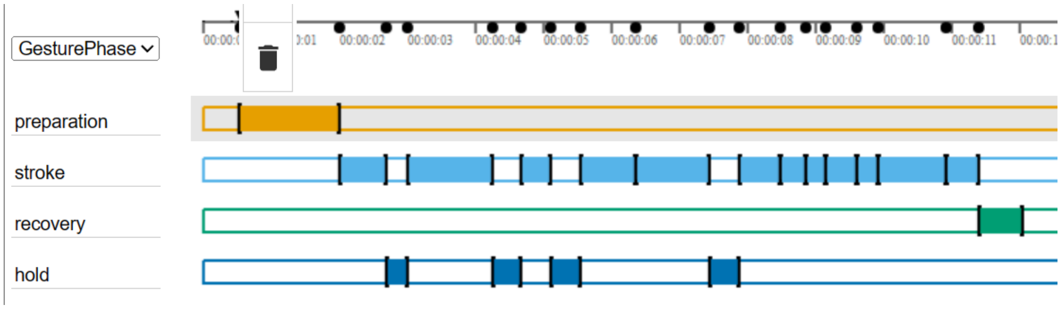


Figure 2: Temporal segment annotation for gesture phases

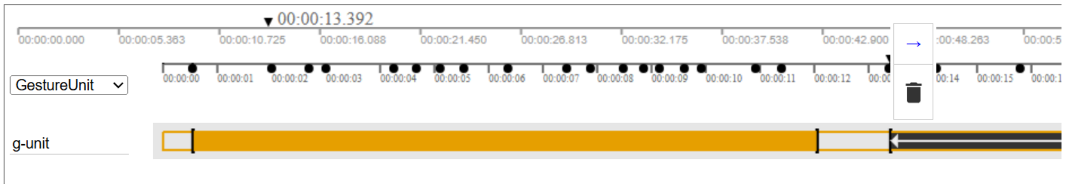
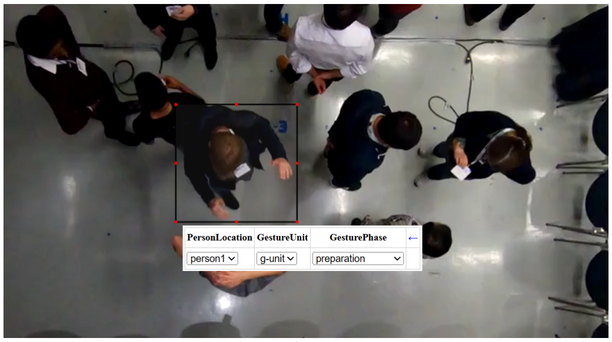


Figure 3: Temporal segment annotation for a gesture unit

The dataset creation for the fine-tuning of VideoMAE is done manually. Using VIA, three aspects are annotated: the space region where the person starts doing the gesture, the gesture phases and the gesture units. Each gesture phase that is annotated is linked to a bounding box that selects the area where the person is moving. Consecutive gesture phases that do not end in a rest position and that happen at a time distance of maximum 2 seconds are grouped in the same Gesture Unit. Additionally, it might be decided to split gesture phases in different gesture units if it makes more sense from a semantic perspective. Bellow, Figure 2 represents an example of gesture phase annotation of a gesture unit that lasts for about 11 seconds. It contains 1 preparation, 12 strokes, 4 holds and 1 recovery. Figure 3 presents an example of how that exact gesture unit is annotated as a single time segment.

The videos contain multiple people performing gestures concomitantly. Each gesture phase is associated with a bounding box representing the person that performs it. Thus, Figure 4 shows how the preparation phase is associated with the bounding box around Person 1, while also associating it to the gesture unit it is part of.



PersonLocation	GestureUnit	GesturePhase	←
person1	g-unit	preparation	

Figure 4: Bounding box annotation at the beginning of a preparation phase

4.3 Preprocessing

The project consists of training two classifiers, one that classifies the phases of a gesture and one that classifies if a clip segment represents a gesture unit or not.

4.3.1 Classifier of Gesture Phases

The first classifier is for gesture phases, and it has the following labels: preparation, stroke, hold, recovery and unknown. While the first four classes are annotated, there is no annotation for the unknown label.

Therefore, the dataset of unknown gestures is built by selecting clips whose lengths are random values, in seconds, from the interval [*5, min(10, T)*], where T is the length of the gap segment in seconds between two consecutive gesture units.

The clips should not contain frames which are part of a gesture unit. The length is decided randomly in the given interval to prevent bias while training. If there is a gap of at least 5 seconds between two gesture units, then a clip is generated from the end of the first gesture. If the gap is longer than 7 seconds, then another clip is generated from the end of the second gesture. The length of the additional clip is a random value, in seconds, from the interval [*5, min(10, T-1)*], where T is the length of the gap segment in seconds between two consecutive gesture units.

The -1 comes into play to ensure that the two clips are not identical. In order to get a clip that only contains the person doing the unknown gesture, the video is cropped around a new bounding box, designed so that it fits both the bounding boxes of the prior and ulterior gesture phases.

The clips for preparation, stroke, hold and recovery are generated by using the annotated time segments. By using the bounding box associated with each gesture phase, the videos are cropped so they only show the person doing that exact gesture.

4.3.2 Classifier of Gesture Units

The second classifier is used to determine whether a clip represents a gesture unit (g-unit) or not. Thus, we have two labels for the model: g-unit and nothing. The g-unit clips are generated by cropping the annotated time segments, while the non-g-units are generated by cropping the gap segments between two gesture units. The gap segments should be at least 2 seconds long so that they can have a similar length to a g-unit.

4.4 Model Training

After generating the clips for both classifiers, the dataset is built into training data sets using the 70-15-15 rule as follows: 70% training, 15% validation and 15% testing.

4.4.1 Classifier of Gesture Phases

The dataset consists of 1455 clips, which are split into 1015 for training, 216 for validation and 224 for testing. Figure 5 presents the label distribution over the dataset. As we see in the figure, there is an imbalance between labels due to two factors. Firstly, the strokes are the only mandatory gesture phase for a gesture unit, so a gesture unit might not contain preparation, recovery or hold. Secondly, strokes are much more frequent than other phases, as can be seen in the dataset distribution presented in Figure 5.

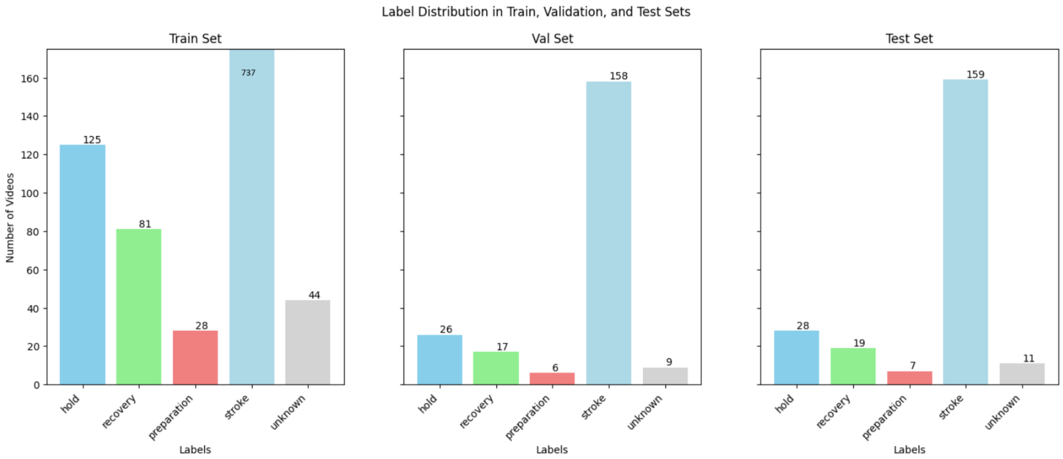


Figure 5: The video distribution on train validation and test subsets, for each label

The training set is augmented using the following transformations: Random Short Side Scale, Random Crop and Random Horizontal Flip.

4.4.2 Classifier of Gesture Units

The dataset consists of 196 clips, which are split into 136 for training, 28 for validation and 32 for testing. The label distribution over the dataset is also quite unbalanced, with g-units being more predominant than the nothing clips, a fact shown in Figure 6. The difference between the two classes is caused by the fact that gesture units can follow each other after less than 2 seconds, which will lead to a gap segment not fit to be labeled as "nothing".

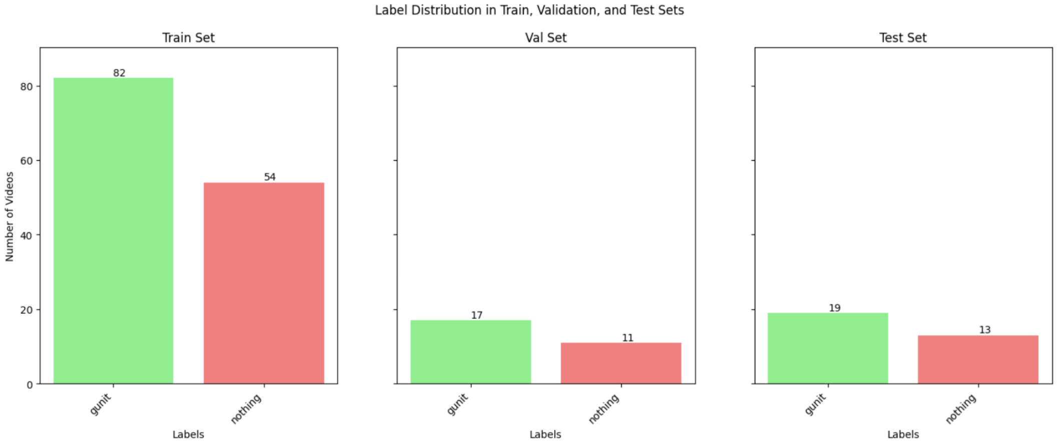


Figure 6: The video distribution on train validation and test subsets, for each label

Similarly to the gesture phase classification task, the training set is augmented using the following transformations: Random Short Side Scale, Random Crop and Random Horizontal Flip.

5 Result and Limitations

The following section will present and discuss the accuracies of both models, as well as their limitations.

5.1 Gesture Phase classification

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

Table 2: Per label accuracies for the classifier

For the Gesture Phase model, the fine-tuning process led to a model with an overall accuracy of 95% on the test set, with a per-label accuracy shown in Table 2. The model, its parameters and its test results are uploaded on Hugging Face and are available at Grigore [6]. This model was fine-tuned for 10 epochs, using a batch size of 8 and a learning rate of 1-e5. The evolution of the train loss and validation loss over multiple epochs is shown in Figure 7. The figure shows a steady decrease in both losses, indicating effective learning and generalization and suggesting that the model does not overfit.

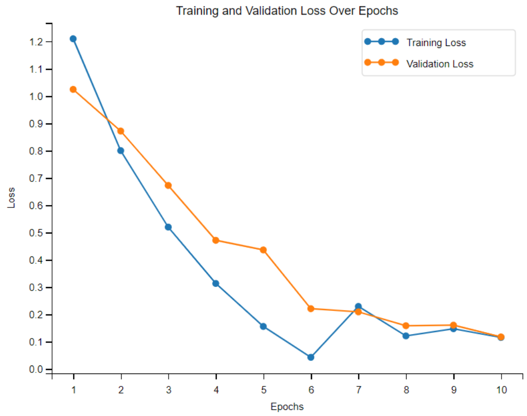


Figure 7: Validation and Training loss during the training of gesture phases classifier

With an average accuracy of 95%, the results of the model are above baselines such as Uniform Distribution Baseline, which is 20% or Majority Class Baseline, which is 70%, as shown in Figure 8. Uniform Distribution Baseline assumes that each class has a similar probability to be chosen, which leads to 20% accuracy. Majority Class Baseline is a model that always predicts that a clip is a stroke, and the accuracy of 70% is due to the imbalance between strokes and other gesture phases.

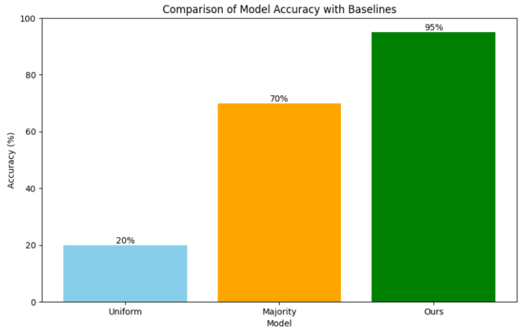


Figure 8: Accuracy of our model, Uniform Distribution and Majority Class

Based on these results, we can conclude that four out of the five gesture phases are classified with high accuracy. The preparation phase has a low accuracy of 0.43, but it still surpasses the 0.20 performance of the Uniform Distribution baseline. Bellow, Figure 9 contains the results of the model on the test dataset as a confusion matrix.

As it can be seen, preparation gestures are often recognized as strokes. A possible cause is the similarity between the movements. A movement for rest position towards upwards can be easily mistaken as a less powerful stroke movement, even by a human evaluator. The distinction between the two is that the preparation gesture is preceded by no other phase within the gesture unit it belongs to. The model does not take into account the previous gestures phases, thus it cannot distinguish them properly from one another.

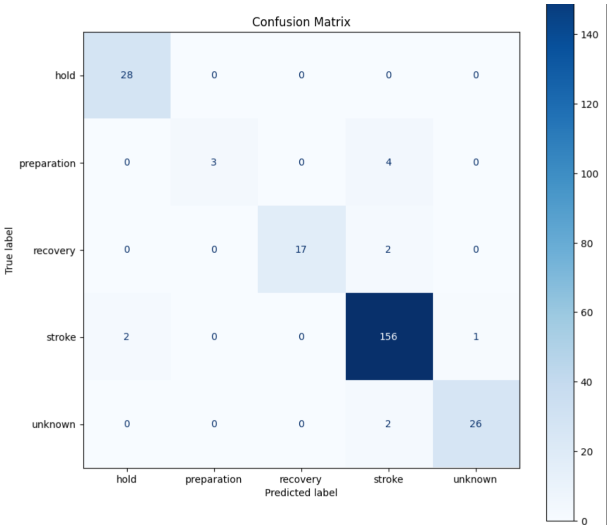


Figure 9: The confusion matrix results of the phase classifier

5.2 Gesture Unit classification

For the Gesture Unit model, the fine-tuning process led to an overall accuracy of 93% on the test set, with per-label accuracy shown in Table 3. All the details regarding the model are also available on Hugging Face at Grigore [7]. The model was built by fine-tuning VideoMae for 5 epochs, using a batch size of 8 and a learning rate of 5e-5.

Label	**Accuracy**
G-unit	0.86
Nothing	1.0

Table 3: Per label accuracy for the classifier

The 93% average accuracy obtained by the model is above baselines such as Uniform Distribution Baseline, which is 50%, or Majority Class Baseline, which is 60%, as shown in Figure 11.

Following those results, we can assess that the model manages to recognize if a video does not represent a gesture unit, but sometimes it labels gesture unit videos as nothing. Figure 10 contains the results of the model on the test dataset as a confusion matrix.

A possible cause is that a gesture unit is not only defined as a person doing a movement. It has to be considered as having a meaning for the conversation. Movements such as scratching their head or putting their had in the pocket are not considered a gesture phase, and the clips that only contain such movements are labeled as nothing. Thus, the relatively

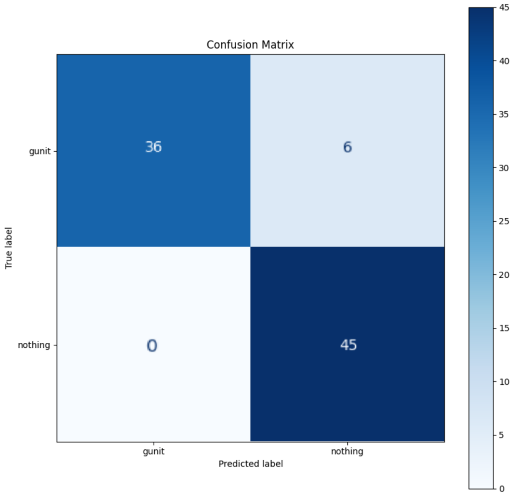


Figure 10: The confusion matrix results of the gesture unit classifier

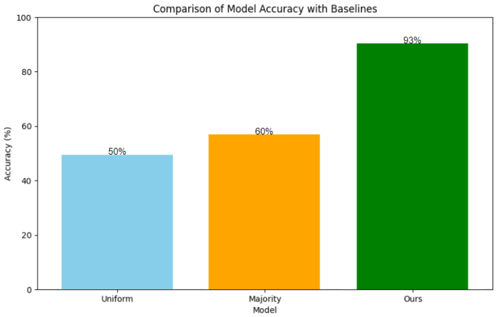


Figure 11: Accuracy of our model, Uniform Distribution and Majority Class

small dataset could make the model consider gesture units that only contain one or few strokes as just movements without meaning.

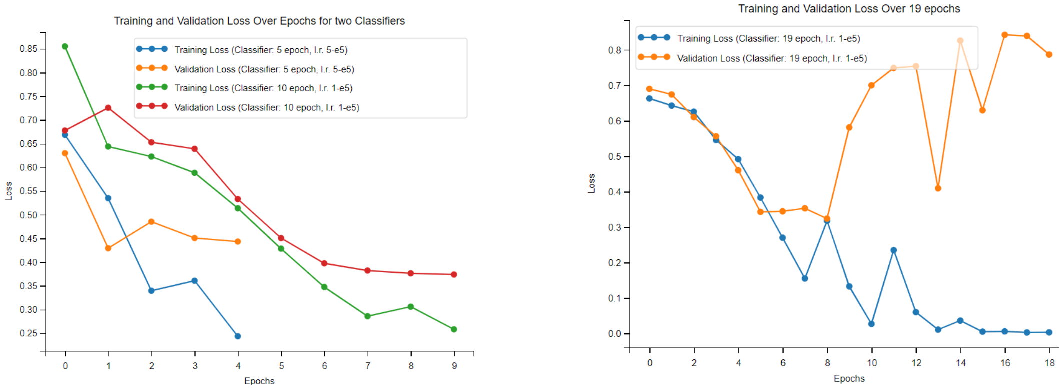
Figure 12a shows the evolution of the validation for the 5 epochs classifier and for another classifier trained for 10 epochs with a learning curve of 1-e5. Figure 12b shows the evolution of a configuration of 20 epochs, with a learning curve of 1-e5. The model trained for 10 epochs got an overall accuracy of 90% and the model trained for 19 epochs gives an accuracy of 95%. In Figure 12a, we can see that there is a discrepancy between the evolution of the training loss compared with that of validation loss. In both cases the final validation loss is higher than the training loss, which indicates that the model could be overfitting. To try to solve this problem, a 19 epochs classifier was fine-tuned, but, as it can be seen in Figure 12b, the validation is fluctuating, which indicates that the model is now severely overfitting. Considering this and that we have a classification problem between only two labels, a possible conclusion could be that the dataset is too small to consider the obtained accuracies reliable. Nevertheless, the first two classifiers indicate that there is potential in using the current technique for the gesture unit classification task.

5.3 Limitations

The project encountered two main limitations: the need to manually annotate the Conflab dataset and the dataset imbalance between classes for both classifiers.

Regarding building the dataset gesture phase and gesture unit datasets upon Conflab recordings, manual annotation was necessary due to the lack of a pre-annotated dataset. Besides the process taking several weeks, the amount of annotation that one person can produce is limited.

While the dataset for the gesture phase classifier appears to perform well on the current size of almost 1.5k videos, expanding the dataset could help mitigate the per-class imbalance problem. Strokes appear more often than any other phases. Thus, they represent a large part of the current dataset. If we have more videos, we can gather more samples for



(a) Validation and Training loss for two gesture unit classifiers

(b) Validation and Training loss for the gesture unit classifier trained for 19 epochs

Figure 12: Comparison of training and validation losses between three classifiers

underrepresented labels such as "preparation", "recovery" and "hold" and generate more "unknown" labeled clips. By doing so, we can assess if the assumption that the mislabeling of "preparations" as "strokes" is truly caused by the fact that we do not consider what phase precedes the gesture we try to classify. Another possible option could be that the "preparation" clips are just underrepresented, and by increasing their number, the problem will disappear.

Regarding the gesture unit classification task, as explained in the previous subsection, it appears that the dataset of gesture units is a bit too small to get a model that generalizes well on unseen data. Thus, expanding this dataset would offer a more accurate assessment of the performance of the classifier.

Another possible limitation could be that the Conflab dataset consists of 5 recordings of 16 minutes each from different camera angles, but from the same conference. Thus, to ensure that the classifier is able to generalize properly, the videos used to expand the dataset could be recorded in a different environment.

6 Conclusions and Future Work

The following are some of the conclusions that can be drawn from our experiments and a discussion on what are the direction in which we could proceed with future work.

6.1 Conclusions

The paper addresses the viability of classifying gestures, described through coding schemes and recorded in crowded environments. The project consists of two classifiers, both obtained by fine tuning the visual transformer VideoMae described in Tong et al. [22].

The first one classifies hand gesture phases, and the gesture coding is based on the prosodic dimensionality of a gesture described by Tutuncubasi et al. [23]. The accuracy obtained with this approach was high, of 95%, while also getting high per label accuracies for each phase, exception making the preparation phase. The adaptation of the Tutuncubasi et al. [23] one-person gesture coding to the crowded environment scenario was done by associating each gesture annotation to bounding boxes around the person doing it. Upon cropping around the bounding box, the problem was reduced to a one-person gesture coding annotation. With a high overall accuracy and a converging training loss and validation loss, the experiment appears to be a success.

The second classifier decides whether a video represents a gesture unit or not. It is also based on the prosodic dimensionality of a gesture described by Tutuncubasi et al. [23], where consecutive and continuous gesture phases are grouped into gesture units. The accuracy of the classifier was also high, of 93%, while also getting 100% in correctly classifying non-gesture units and 86% on gesture units. Similar adaptation from one-person gesture coding to multi-person was used in this task as well. Should be considered that the discrepancy between the validation loss and training loss described in Section 5 raised questions regarding the ability of the model to generalize on unseen data. Based on the results obtained, a possible solution could be to expand our gesture unit dataset.

Finally, it should be noted that a total of 1455 videos were used to build a gesture phase classifier, while 196 videos were used to build a gesture unit classifier. Both of them used a train-validation-test split approach of 75-15-15.

6.2 Future Work

One aspect for future work includes expanding both the dataset of hand gesture phases to correct the unbalanced class distribution. By doing so, our assumption described in Section 5 regarding the mislabeling of "preparation" gestures as "strokes" can be verified. On the note of expanding the dataset, it appears that it would also help with improving the performance of the gesture unit classifier, which appears to encounter problems in generalizing on unseen data.

Another aspect would regard experimenting with different models such as 3D-CNNs or some adaptations of 2D-CNN, which were described in previous sections. Due to the time constraints, it was not possible to implement and assess their performance, but the promising results obtained in Hedegaard and Iosifidis [8] and Molchanov et al. [13] make CNN a possible solution for the classification problem.

Focusing on the M3D annotation procedure, further research could be focused upon classifying gestures by the other two dimensions explained in Tutuncubasi et al. [23], more exactly the form of the gesture and its meaning. Additionally, the prosodic dimension has another interesting aspect which can be researched, the rhythmic properties of the gesture.

Lastly, after the promising results obtained by fine tuning VideoMae, a clear direction would be to build a recognizer for gesture phases and gesture units. Such a recognizer could be employed for live recognition in video streams.

Nevertheless, the directions in which the project can extend are endless. The high accuracies obtained by the two classifiers prove that there is potential in using coding schemes to identify and classify gestures. Thereby, this opens the opportunity for further research into gesture recognition based on coding schemes which are not only for hand motions, but also full body movements.

7 Responsible Research

This section outlines the practices withheld during the project to ensure that the research is conducted responsibly. The main focus is on ethical data collection, privacy, reproducibility, and the usage of language models.

7.1 Data Collection and Privacy

An important ethical aspect, considering the nature of the project, is ensuring that the dataset on which the model is trained is obtained ethically, with the consent of the participants. This matter was already addressed by the researchers who built the Conflab dataset. Additionally, the training of the model should be done on a platform whose data policies align with the informed consent given to the participants. In our case, the training is performed using the Delft Blue Supercomputer, which meets these criteria.[1]

7.2 Ethical Considerations

Ethical considerations are crucial in ensuring the integrity and fairness of research. To avoid racial bias in the machine learning models used for classification, it is essential to use a diverse dataset. The current version of the dataset addresses this issue when selecting the persons whose movements are to be annotated, but it should also be considered if and when the dataset is extended.

Participant identities are protected by using pseudonyms. Within a video, individuals are identified as "Person_index". The pseudonym of an individual changes from one video to another to further ensure anonymity.

7.3 Reproducibility

Reproducibility is fundamental to ensuring that results can be consistently obtained and verified by other researchers. Therefore, the entire code used in this project is openly available on GitHub. Additionally, the model, its parameters, and results are freely available on HuggingFace for anyone who wishes to test them. The only aspect of the project that is not publicly accessible is the Conflab dataset and the manual annotations built for the fine-tuning process, due to the agreement signed with the volunteers who were recorded.

7.4 Usage of LLMS

Throughout the project, LLMs were used as a way to gather information about diverse topics. Additionally, GitHub Copilot was also used for programming purposes [4]. The main tools used for information gathering are Perplexity and ChatGPT [17][15]. The scenarios in which these models were used are the following:

- Search for papers on a certain subject. An example of a prompt would be: "Can you find me some research papers that showcase coding schemes for hand gestures?"
- Explain concepts. A prompt example is: "Can you explain what data splits are optimal for the fine tuning of a visual transformer model?"
- Explain results from an experiment. A prompt example is: "Can you explain what it means if my accuracy on the training set is high, but on the testing set it is low?"
- Explain how certain python libraries work. A prompt example is: "Can you explain how I can train a model using PyTorch?"

Finally, it should be mentioned that for all the answers provided by LLMs where either manually tested or fact-checked from various sources to ensure that they are valid.

References

[1] TU Delft. *DelftBlue: the TU Delft supercomputer*. Accessed: 2024-06-10. 2024. URL: https://www.tudelft.nl/dhpc/system#:~:text=%23%20DelftBlue%3A%20the%20TU%20Delft%20supe.

[2] 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: 978-1-4503-6889-6/19/10. DOI: 10.1145/3343031.3350535. URL: https://doi.org/10.1145/3343031.3350535.

[3] Paul Ekman and Wallace V. Friesen. "Hand Movements". In: *Journal of Communication* 22.4 (Feb. 2006), pp. 353–374. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.1972.tb00163.x. eprint: https://academic.oup.com/joc/article-pdf/22/4/353/22387316/jjnlcom0353.pdf. URL: https://doi.org/10.1111/j.1460-2466.1972.tb00163.x.

[4] GitHub. *GitHub Copilot*. Accessed: 2024-06-23. 2024. URL: https://github.com/features/copilot.

[5] Alexandru Grigore. *Gesture-Classification-Thesis*. 2024. URL: https://github.com/AlexxGrigore/Gesture-Classification-Thesis.

[6] Alexandru Grigore. *Model Results on Hugging Face for Gesture Phase classification*. Accessed: 2024-06-10. 2024. URL: https://huggingface.co/alexgrigore/videomae-base-finetuned-good-gesturePhaseV11.

[7] Alexandru Grigore. *Model Results on Hugging Face for Gesture Unit classification*. Accessed: 2024-06-10. 2024. URL: https://huggingface.co/alexgrigore/videomae-base-finetuned-good-gestureUnitV11.

[8] Lukas Hedegaard and Alexandros Iosifidis. "Continual 3D Convolutional Neural Networks forÂ Real-time Processing ofÂ Videos". In: *Computer Vision â ECCV 2022*. Springer Nature Switzerland, 2022, 369â385. ISBN: 9783031197727. DOI: 10.1007/978-3-031-19772-7_22. URL: http://dx.doi.org/10.1007/978-3-031-19772-7_22.

[9] Glenn Jocher. *YOLOv5 by Ultralytics*. Version 7.0. 2020. DOI: 10.5281/zenodo.3908559. URL: https://github.com/ultralytics/yolov5.

[10] Hedda Lausberg and Han Sloetjes. "Coding gestural behavior with the NEUROGES–ELAN system". In: *Behavior Research Methods* 41.3 (2009), pp. 841–849. DOI: 10.3758/BRM.41.3.841.

[11] MCG-NJU. *VideoMAE: Video Pre-training and Fine-tuning*. Accessed: 2024-06-19. 2023. URL: https://github.com/MCG-NJU/VideoMAE/blob/main/engine_for_finetuning.py.

[12] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[13] Pavlo Molchanov et al. "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4207–4215.

[14] Joe Navarro. *The Dictionary of Body Language: A Field Guide to Human Behavior*. HarperCollins, 2018.

[15] OpenAI. *ChatGPT: A Conversational AI Model*. https://www.openai.com/chatgpt. Accessed: 2024-06-19. 2024.

[16] Mansi Patil, Vishal Patil, and Unisha Katre. "Unspoken science: exploring the significance of body language in science and academia". In: *European Heart Journal* 45.4 (Oct. 2023), pp. 250–252. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehad598. eprint: https://academic.oup.com/eurheartj/article-pdf/45/4/250/56432906/ehad598.pdf. URL: https://doi.org/10.1093/eurheartj/ehad598.

[17] Perplexity. *Perplexity AI*. https://www.perplexity.ai. Accessed: 2024-06-19. 2024.

[18] Jose Vargas Quiros. *covfee: Continuous Video Feedback Tool*. https://josedvq.github.io/covfee/. 2024.

[19] 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/95f9ad2e251e9014697589037450f9bb-Paper-Datasets_and_Benchmarks.pdf.

[20] 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.

[21] Yunlong Tang et al. *Video Understanding with Large Language Models: A Survey*. 2024. arXiv: 2312.17432. URL: https://arxiv.org/pdf/2312.17432.

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

[23] U. Tutuncubasi et al. *The M3D Training Program*. https://m3d.upf.edu. 2023.

[24] J. Vargas and H. Hung. "CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection". In: *Working Notes Proceedings of the MediaEval 2019 Workshop*. 2019, pp. 11–13.

[25] P. Wittenburg et al. "ELAN: a Professional Framework for Multimodality Research". In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. 2006.

[26] Fan Yang. *A Multi-Person Video Dataset Annotation Method of Spatio-Temporally Actions*. 2022. arXiv: 2204.10160 [cs.CV].

[27] Jing Yu, Ming Qin, and Shun Zhou. "Dynamic gesture recognition based on 2D convolutional neural network and feature fusion". In: *Scientific Reports* (). URL: https://doi.org/10.1038/s41598-022-08133-z.

[28] Yi Zhu et al. *A Comprehensive Study of Deep Video Action Recognition*. 2020. arXiv: 2012.06567 [cs.CV].