# Gesture Recognition for Enhanced Meeting Analysis

## Segmenting and Tracking Hand Movements During Human Interaction

**Atanas Semov**

**Supervisors: Stephanie Tan[1], Edgar Salas Gironés[1]**

[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
January 26, 2025

Name of the student: Atanas Semov
Final project course: CSE3000 Research Project
Thesis committee: <Responsible Professor>, <Supervisor>, <Examiner>

## Abstract

The gesture recognition in the collaborative setting, like a meeting, is a very unique challenge due to the complex and dynamic nature of hand movements. This work identifies, annotates, and classifies the gesture phases-preparation, stroke, retraction, and neutral-using a multi-step approach: integrating Segment Anything Model for accurate hand segmentation, using ELAN software for manual annotation, and the VideoMAE model for classification of gesture phases. Our approach effectively separates the hand motions from the background clutter and annotates the gesture phases, thus enabling the VideoMAE model to capture the temporal dynamics in gesture recognition. The results bring out the variability in phase durations and demonstrate the model's accuracy in classifying gestures in meeting environments with different settings. This work contributes to an increasing trend of developments in the field of automated gesture recognition, laying a foundation for futures studies to explore issues such as hand orientation and fusion of gesture with speech data.

## 1 Introduction

Effective meetings are crucial for collaboration and decision-making in workplaces, yet they are often plagued by misunderstandings among participants. When individuals leave with differing interpretations of the discussion, this can result in poor outcomes, including miscommunication and wasted time. Addressing these challenges requires a deeper understanding of both verbal and non-verbal elements of communication during meetings.

Modern automated systems for meeting analysis predominantly focus on tasks such as speech-to-text conversion and text summarization, utilizing platforms like IBM Watson or Google Cloud Speech.

These systems suffer, however, because they are unable to account for other aspects of communication which are important in the big picture of collaboration, for example gestures[2]. Certain innovations, for example the Segment Anything model [3] of Meta offer possibilities to counter the issues. The model not only does segmentation of objects but also tracking of such objects that are part of a video; this model opens doors to quantitative analysis of movement as a channel of communication in meeting situations as recordings of all video dynamics can now be analyzed in detail.

This paper addresses the important research question: Why and how are hand movements and gestures characterized for the purpose of informing meeting participants? In order to cover such a significant question, the study addresses three major sub-questions:

How can gestures be recognized in close proximity to other objects [6]? How can the phases of a gesture be identified, annotated, and analyzed? What are effective methods for transitioning between gesture phases during classification in automated systems? The significance of these questions in this research is to measure how new hand-tracking systems are designed for more challenging tasks, but also how these systems are adjusted for use in meeting settings.

The contributions of this research are twofold. The contributions of this research are twofold. First, it examines gestures in complex environments, such as crowded meeting settings, and investigates the identification, annotation, and analysis of their phases—preparation, stroke, retraction, and neutral. By studying the durations and variations of these phases, the research provides a detailed understanding of the temporal dynamics of gestures, enabling more precise modeling and interpretation.

Second, it explores strategies for detecting transitions between gesture phases in automated systems. This includes testing different methods and settings to determine how factors like frame selection and temporal context impact the system's ability to accurately detect and classify these transitions.

These contributions collectively advance the field of automated meeting analysis by integrating gesture-based insights into existing frameworks. By focusing on the functional roles, temporal characteristics, and classification challenges of gestures, the research aims to enhance the accuracy and applicability of systems that analyze non-verbal communication in collaborative settings.

The organization of this paper is as follows. In Section 2, we review the related work on gesture classification and explain how our approach improves upon them. Section 3 discusses the methodology for identifying, annotating, and analyzing gesture phases, as well as the study of phase durations. The tool selection is also explained here. In Section 4, we present the experimental setup and results, followed by a discussion of hypothesis testing. Section 5 reflects on the ethical aspects of the research. Finally, Section 6 concludes the paper and outlines future work in this area.

## 2 Background and related work

In this section, we will present the works that our research builds upon, highlighting the sections that are the most relevant for our approach.

Current gesture analysis approaches frequently employ deep learning models for feature representation of raw RGB or depth data, as well as for gesture detection and classification. A pioneering approach by Molchanov et al.[4] uses a 3DConvolutional Neural Network (CNN) on depth and RGB data and incorporates Connectionist Temporal Classification (CTC) to predict an "in progress" gesture from video segments. Typically, the class content of these segments features a "silent gesture that depicts an action or an object".

Recent advancements in gesture detection have moved beyond simple binary classification approaches to capture the sequential and contextual nature of gestures. Ghaleb et al. [1] proposed a novel framework that treats gesture detection as a multi-phase sequence labeling problem, addressing the dynamic phases of gestures—preparation, stroke, and retraction. Their method processes skeletal movement sequences over time, utilizing Transformer encoders to capture contextual embeddings and Conditional Random Fields (CRFs)

for sequence labeling. Evaluated on a large dataset of co-speech gestures in task-oriented dialogues, their framework demonstrated superior performance compared to traditional binary classification models, particularly in detecting gesture strokes. Notably, the integration of Transformer-based contextual embeddings improved the detection of gesture units, highlighting the framework's ability to model the fine-grained dynamics of gesture phases effectively. This work underscores the importance of modeling gestures as inherently sequential phenomena, paving the way for more nuanced analysis of co-speech gestures in real-world interactions.

# 3 Methodology

To address the research questions, this study employed a multi-step approach designed to explore the recognition and classification of hand gestures and their phases in meeting scenarios. The methodology was specifically tailored to provide both theoretical and practical insights into gesture dynamics in collaborative contexts.

## 3.1 Instrumentation and Systematic Tool Integration

**Introduction to Tool Selection**

In any computational research project, the selection of appropriate tools forms the backbone of the methodology, directly influencing the accuracy, reliability, and efficiency of the outcomes. In this project it was required to have tools capable of handling tasks such as precise segmentation of hand movements, accurate annotation of gesture phases, and robust spatiotemporal modeling of gesture dynamics. Each tool was evaluated based on its functionality, compatibility with other components of the research pipeline, and ability to address the challenges posed by complex multi-person interactions.

The criteria for tool selection extended beyond basic functionality; scalability, ease of integration, and adaptability to domain-specific challenges were key considerations. For instance, the segmentation tool needed to perform reliably even in the presence of overlapping objects or occlusions, while the annotation software had to support detailed temporal and semantic labeling of gestures. These requirements ensured that the tools not only facilitated individual tasks but also contributed to a cohesive and efficient workflow.

**Role of Each Tool in the Workflow**

The tools employed in this project were integral to the three primary phases of the research: data preparation, annotation, and model training. Each tool was selected to fulfill a distinct role in addressing specific research challenges:

- **Segmentation Tool: Segment Anything Model (SAM)**
  The Segment Anything Model was employed to generate pixel-level segmentation masks for hands in video recordings. Its advanced capabilities in handling object segmentation across diverse and cluttered environments ensured precise isolation of hand movements. The ability of SAM to generalize across different contexts without requiring extensive domain-specific training made it an optimal choice for this project.
- **Annotation Tool: ELAN Software**
  ELAN software was used for the manual labeling of gesture phases, including preparation, stroke, retraction, and neutral states. ELAN's support for multi-tier annotations and temporal alignment made it ideal for associating gesture phases with time-sequenced data.
- **Model Training and Fine-Tuning: VideoMAE**
  VideoMAE was selected as the backbone model for spatiotemporal learning. Its design, which excels in capturing motion dynamics across video frames, made it particularly suitable for recognizing and classifying gestures based on their phases. The model was fine-tuned using the manually annotated dataset, enabling it to generalize across new, unseen videos.

## 3.2 Research Implementation Workflow

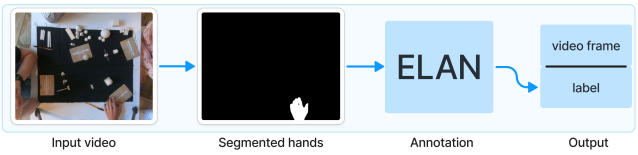



Figure 1: **Video Processing**

The workflow followed to address the research question consists of three parts. As shown in Figure 1, first, masks of the hands are generated from a series of videos using the Segment Anything Model (SAM). Next, selected segments of these videos are manually annotated according to the M3D principles, using ELAN software.

Finally, the labeled data, consisting of sequences of frames paired with their corresponding phase labels, is analyzed and used to fine-tune and test a pre-trained video processing tool, VideoMAE. The training data for VideoMAE is provided in the format of temporal frame sequences with per-frame gesture phase annotations, enabling the model to learn phase-specific features effectively.

The first research question: How can gestures be recognized in close proximity to other objects? This question required the use of a robust segmentation method. The Segment Anything Model (SAM) was chosen for its proven capability to segment objects with high precision, ensuring that hand regions could be accurately isolated from visually complex backgrounds typical of meeting environments. This step was essential to create a clean dataset that focuses on hand movements, allowing subsequent steps to operate on clearly defined inputs.

To address the second research question: How can the phases of a gesture be identified, annotated, and analyzed in terms of their durations and variations between different gesture types? Manual annotation was conducted following the guidelines established in the M3D framework[5]. This decision was motivated by the need to incorporate domain-specific expertise in gesture analysis. The annotation process categorized gestures according to their phases (preparation, stroke, retraction, and neutral). By integrating these estab-

lished criteria, the dataset ensured a rich representation of both the structural and semantic aspects of gestures.

Finally, the third research question: What are the most effective methods for transitioning between gesture phases during classification in automated systems? To address this, I used the videoMAE model. This model was selected for its capability to capture spatiotemporal relationships in video data, making it ideal for analyzing gestures as sequences that unfold over time. By testing and different settings and evaluating the results I found what was optimal when classifying the gestures.

This methodology was chosen for its ability to combine advanced machine learning techniques with human expertise. By using SAM for precise segmentation, manual annotation for capturing nuanced gesture semantics, and videoMAE for dynamic classification, the approach ensured a comprehensive pipeline capable of addressing the challenges outlined in the research questions. This careful alignment of methods to research objectives highlights the study's commitment to both accuracy and interpretability in gesture recognition and analysis.

## 4 Experimental Setup and Results

In this chapter, I outline the experimental process and present the results of the proposed methodology for gesture analysis in meeting scenarios. The section is divided into two parts: details of the experiment and the corresponding results.

### 4.1 Experiment Details

To evaluate the effectiveness of the proposed methodology for gesture recognition in meeting scenarios, a structured experimental pipeline was designed. The experiment involved three main stages: segmentation, annotation, and model fine-tuning, each carefully tailored to ensure the robustness of the system.

**Hand Segmentation**

The first step in the pipeline utilized the Segment Anything Model (SAM) to generate segmentation masks for hand regions in video footage of meetings. The videos consisted of diverse interactions with varying lighting conditions, participant and objects configurations, and hand movements. To be able to perform this step the videos had to be split into frames at a rate of 30 frames per second (FPS). This frame rate was chosen based on a careful consideration of multiple factors to ensure an optimal balance between temporal resolution, computational efficiency, and gesture phase identification.

One critical factor in selecting 30 FPS was the need to capture the different phases of gestures—preparation, stroke, retraction, and neutral—accurately. Gestures often involve quick and dynamic movements, and a lower frame rate could miss key transitions between these phases, leading to a loss of important temporal details.

At the same time, computational efficiency was a significant consideration. While higher frame rates, such as 60 FPS, could offer even finer temporal resolution, they would result in a considerably higher number of frames to process, increasing both storage requirements and processing time. Opting for 30 FPS provided a practical trade-off, allowing faster processing and reduced computational demands without compromising the ability to capture gesture nuances.

SAM was applied to each video frame to isolate hand regions from the background, and its segmentation performance was evaluated. Deficiencies in the results, such as incomplete masks or false positives, were addressed to ensure the quality and accuracy of the outputs.

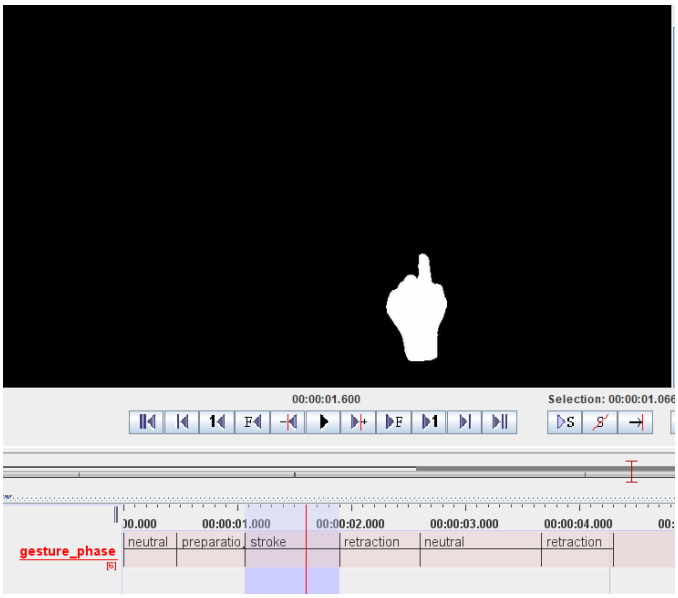



Figure 2: **ELAN Software**

**Data Labeling**

Once the segmentation masks were obtained, the videos were imported into the ELAN software for manual gesture annotation. As shown in the example in Figure 2, the phases are annotated to specific segments of the video. I utilized the frame-by-frame annotation option, which allows me to accurately determine the start and end frames of each gesture phase. To ensure the accuracy, consistency, and objectivity of the labeling process, the M3D framework was employed. This framework offers well-defined guidelines for identifying and annotating the different phases of gestures. A gesture typically unfolds in the following sequence: preparation, stroke, and retraction. However, in some cases, either the preparation or retraction phase may be absent. The neutral phase, representing a pause or inactivity in hand movement, can occur at any point between or within other phases, providing flexibility in the gesture's overall flow. The annotations also captured the temporal progression of gestures, ensuring that the sequential nature of gesture phases was preserved.

A key principle used during annotation was that a gesture is defined by the presence of exactly one stroke phase. The stroke phase, being the most meaningful and emphasized part of a gesture, is pivotal in distinguishing one gesture from another. This definition provided a clear method for determining the number of gestures in a video.

To distinguish the preparation phase from retraction, a

guiding principle was applied: preparation is defined as the movement of the hand from a rest position to the location where the stroke begins, while retraction occurs after the stroke and represents the movement of the hand back toward the rest position. This distinction ensures that the transitions between gesture phases are clearly marked and temporally accurate.

Additionally, the video segments were carefully selected to minimize inactive periods, ensuring that gestures (ranging between 5 and 10 per segment) occurred in rapid succession, one after another. This selection criterion helped to focus the annotation on continuous, dynamic gestures without interruptions, preserving the natural flow of movements.

In the output from the ELAN software, each gesture is annotated with its respective phases and corresponding start and end times. The file containing the annotated data follows the structure:

- Each distinct gesture phase is represented in three lines.
- The gesture phase is indicated in the first line.
- After it the `begin` and `end` times for each phase are provided in seconds, marking the temporal boundaries of the gesture phase within the video.

This structure allowed me to easily transform the times into frames of the video. By knowing the frame rate of the video, which was constant throughout the recording, I was able to compute the corresponding frame numbers for each `begin` and `end` time. Specifically, for a given time $t$, the corresponding frame number $f$ can be calculated as:

$$f = t \times \text{frame\_rate}$$

Where frame_rate is the number of frames per second (FPS) of the video. This process allowed me to represent the gesture phases in terms of discrete frames, facilitating the integration of the data with the frame-level analysis performed by the gesture recognition model.

**Model Training and Testing**

- **Fine-Tuning the videoMAE Model:** After generating the annotated dataset, the videoMAE model was fine-tuned using the labeled videos. The input to the model consisted of the segmented hand regions and their corresponding phase annotations. The fine-tuning process was designed to enable the model to learn spatiotemporal features of gestures, improving its ability to classify gestures dynamically over time. To fine-tune the model, the data needed to adhere to a specific format. The videos were resized to 224x224 pixels resolution and segmented, ensuring each segment contained frames consistently labeled with the same gesture phase. Two segment length configurations were tested: one with a maximum of 12 frames per segment and the other with 16 frames. Each sequence of frames was paired with its corresponding label, ensuring that the temporal continuity of the gesture phase was preserved. All segments were then grouped into a list, which was subsequently fed into the model for fine-tuning. This approach enabled the model to learn the temporal patterns of each gesture phase, as it was not limited to processing a single frame at a time. Instead, it received a sequence of frames, allowing the model to capture the dynamics and transitions within each phase, leading to a more robust understanding of the gesture's progression.
- **Testing Scenarios:** To evaluate the system, additional videos containing unseen gestures were processed through the pipeline. The videos were split in the same manner as the training data, with each segment consisting of a sequence of frames corresponding to a single gesture phase. The same approach was applied, where each list of frames was passed through the model to predict a single label for the entire segment. This process allowed the evaluation of the model's performance on previously unseen gestures, providing insights into its generalization capability.
- **Evaluation Metrics:** The performance of the system was evaluated using several metrics, including accuracy, precision, recall, and F1-score, which are essential for measuring the effectiveness of gesture recognition and classification:
  - **Accuracy**:
    $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
    Accuracy measures the overall proportion of correct predictions (both true positives and true negatives) to the total number of predictions. It provides a general indication of the model's performance but may not be reliable in imbalanced datasets.
  - **Precision**:
    $$\text{Precision} = \frac{TP}{TP + FP}$$
    Precision calculates the proportion of positive predictions that were actually correct. It is particularly useful when false positives have a high cost, as it reflects how many of the predicted positive gestures were correct.
  - **Recall**:
    $$\text{Recall} = \frac{TP}{TP + FN}$$
    Recall measures how many actual positive instances were correctly identified by the model. It is important when false negatives are costly, as it reflects the model's ability to identify all relevant gestures.
  - **F1-Score**:
    $$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
    The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both metrics. It is useful when the data is imbalanced or when both false positives and false negatives are important to minimize.

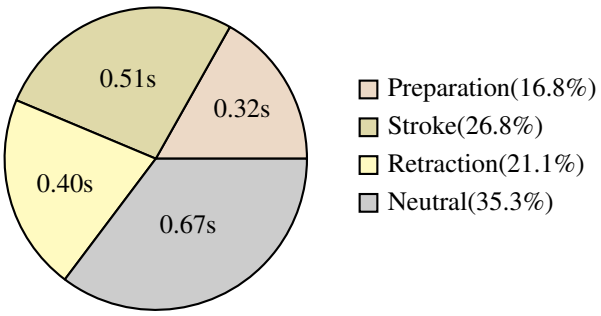
### 4.2 Results

This section presents the outcomes of the experiments conducted during the study. The results are presented using tables and figures to ensure clarity and accessibility, with observations and conclusions drawn directly from the data.

**Phase Distribution Analysis**

Figure 3 summarizes the average percentage of time each phase—preparation, stroke, retraction, and neutral—occupies within a gesture. These percentages were calculated across the annotated dataset to provide insights into the temporal dynamics of gesture phases. The dataset was curated to include only portions with active gestures and hand movements, while segments without gestures or hand activity were excluded. This preprocessing approach ensured a focus on dynamic hand motion, but it also directly influenced the time allocation of the neutral phase, as periods of inactivity were minimized.

Figure 3: Average Gesture Duration by Phase




**Observations:**

- The stroke phase consistently occupies the largest portion of a gesture, aligning with its role as the most meaningful segment.
- The preparation and retraction phases exhibit notable variability, reflecting differences in gesture initiation and conclusion among participants.
- The neutral phase appears intermittently and occupies varying durations, depending on pauses in hand movements.

**Model Accuracy with different settings**

We evaluated the model's accuracy using segments of 12 frames and 16 frames to analyze their impact on recognizing gesture phases. Segments of length 12 performed better in recognizing preparation and retraction phases, as these phases typically occur more quickly and benefit from the finer temporal resolution. Conversely, segments of 16 frames provided a broader context, which was more suited for recognizing longer phases like strokes and neutral gestures. This highlights the importance of segment length in balancing temporal granularity and contextual understanding for gesture phase recognition.

## 5 Discussion and Ethical Considerations

**Data Gathering and Ethical Considerations**

As gesture recognition technologies are applied in real-world scenarios, it is crucial to consider their ethical implications. While this research aims to improve gesture recognition models in complex environments, such as meetings, it is important to acknowledge that such technology may not always be entirely reliable. Errors or inaccuracies in recognizing gesture phases can lead to misinterpretations or unintended outcomes in automated systems, potentially misleading users or decision-makers. We stress the importance of using this research responsibly and advocate for transparency in the development and deployment of AI systems.

The data used for this study was collected with the full consent of the participants involved, and the anonymity and confidentiality of their information were prioritized.

**Reproducibility of the Research**

We aimed to ensure that our findings could be independently verified and built upon. We have provided a detailed description of our experimental setup, including the datasets (gesture phase annotated dataset), the architecture of the model (VideoMAE), and the metrics for evaluation (accuracy, precision, recall, F1-score). We have documented the key hyperparameters, such as clip length (16 frames) and frame size (224x224). Our methodology and evaluation procedures are designed for reproducibility, with clear documentation on the tools and models used.

## 6 Conclusions and Future Work

This paper addresses the challenges of gesture recognition in meetings, focusing on the identification, annotation, and analysis of hand gestures within complex environments. Our main research question was how to effectively recognize gestures in close proximity to objects and how to identify, annotate, and analyze the phases of these gestures in a way that is applicable to automated systems.

To answer these questions, we employed a multi-step methodology that integrated the Segment Anything Model (SAM) for segmentation, ELAN software for manual annotation, and VideoMAE for model training. The segmentation process effectively isolated hand movements from background clutter, while the annotation process provided detailed labeling of gesture phases: preparation, stroke, retraction, and neutral. These phases were then used to fine-tune the VideoMAE model, enabling it to recognize and classify gestures based on temporal dynamics.

A key direction that this research can lead to is the exploration of additional factors that influence the interpretation of gestures, such as hand orientation. While our current approach focuses on gesture phase identification and temporal dynamics, hand orientation plays a crucial role in understanding the intent behind certain gestures. For instance, the direction in which a hand is held or the specific orientation of the palm could provide important context, especially in distinguishing between gestures with similar movements but different meanings.

Another promising area for future research is the integration of gesture recognition with audio processing. Ges-

ture analysis alone provides valuable information about non-verbal communication, but combining it with speech data could enhance the interpretation of the gestures. For example, the meaning of a hand gesture could be better understood when contextualized with the verbal content of the conversation.

In conclusion, while the current work advances the field of automated gesture recognition in meetings, more research is needed to tackle the limitations identified, improve system robustness, and address ethical concerns. This paper provides a foundation for future studies aiming to enhance the accuracy and applicability of gesture recognition technologies in collaborative settings.

## References


[1] Esam Ghaleb and Ilya Burenko. Co-speech gesture detection through multi-phase sequence labeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024.

[2] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, UK, 2004.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, A. Rolland, L. Gustafson, and R. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[4] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, 2016.

[5] Patrick Louis Rohrer, Ulya Tütüncübasi, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve Gibert, PeiLin Ren, Stefanie Shattuck-Hufnagel, and Pilar Prieto. The multimodal multidimensional (m3d) labeling system. Project, August 2020. DOI: 10.17605/OSF.IO/ANKDX, [Online]. Available: http://m3d.upf.edu.

[6] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan. Vision-based hand-gesture applications. *Communications of the ACM*, 54(2):60–71, 2011.