Deciphering the Meaning of Gestures In the Wild

Understanding the meaning of gestures in densely crowded social settings

Irene Aldabaldetrecu Alberdi[1]
Supervisor(s): Hayley Hung[1]**, Ivan Kondyurin**[1]**, Zonghuan Li**[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
June 23, 2024

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

Abstract

Recent studies have shown that gesture annotation schemes should account for the multidimensional nature of gestures and define their meaning in terms of referentiality and pragmatic meaning. However, accurately annotating gesture meaning in densely crowded social settings using such a coding scheme remains to be accomplished. This study uses the MultiModal MultiDimensional (M3D) labelling scheme and the EUDICO Linguistic Annotator (ELAN) tool to annotate video data from the Conference Living Lab (ConfLab) dataset. The ConfLab dataset contains 8 video recordings of standing conversations at a conference, captured from an overhead perspective, and low-frequency audio recordings of the conversations. A total of 1119 clips of individual gesture instances are generated. This data is then fed into a VideoMAE model pre-trained on the UCF101 dataset. The model achieves an overall accuracy score of 49% on the test set but shows a significant bias towards one class due to the imbalanced dataset. Due to the small size of the dataset and the similarities between gestures with different meanings, the model cannot identify different gesture types. The results demonstrate that high-frequency audio or transcripts of the conversations are vital to avoid strong and potentially incorrect assumptions when annotating gesture meaning. Further investigation is required into the annotation and classification of pragmatic meanings and Machine Learning solutions for multi-class, multi-label video classification problems.

1 Introduction

Gestures in communicative settings play a significant role in conveying meaning. Not only do they carry semantic meaning that may not be deciphered directly from speech, but also coordinate with it to visually assist the listener [19]. Traditionally, gesture meaning has been defined either in terms of its referentiality, such as McNeill's (1992) functional dimensions (iconicity, metaphoricity, deixis and beat [18]), or in terms of its pragmatic functions in discourse, which convey information about a speaker's communicative intent independent of their referentiality [22]. Nevertheless, modern frameworks for labelling gestures account for their multi-dimensional nature. Rather than being assigned a single category or function, gestures are classified as having multiple overlapping referential and pragmatic dimensions.

However, it is unclear whether gestures produced in densely crowded social settings, colloquially referred to as "in the wild", can be accurately classified. In previous studies, the annotated data was collected in the lab or in an equally controlled environment. Furthermore, they involved recordings of just one or two people at a time. As such, it is yet to be explored whether existing gesture labelling schemes can be used to annotate recordings of multiple, unrestricted interactions.

This study analyses whether VideoMAE [25] and the MultiModal MultiDimensional (M3D) labelling scheme [22] can be used to annotate and accurately classify gestures produced in densely crowded social settings. The initial step involves a literature survey on existing coding schemes for annotating gestures and Deep Learning models for video classification. After selecting the M3D labelling system, recordings of people from the ConfLab dataset are annotated via the ELAN annotation tool. Finally, a pre-trained VideoMAE model is fine-tuned on a dataset containing 1119 gesture clips. The results show that gesture meaning cannot be properly annotated without recordings or transcripts of co-occurring speech. Due to the bias caused by an imbalanced dataset and the visual similarity of hand gestures despite their different meanings, VideoMAE cannot effectively distinguish gesture types.

2 Background and Related Work

This section introduces previous work upon which this study is based. Section 2.1 discusses previous definitions and conceptions of gesture meaning. Section 2.2 presents various coding schemes and analyses their main strengths and weaknesses. Finally, Section 2.3 explains why transformer-based models were considered for the project and compares two state-of-the-art video classification models: ViViT [2] and VideoMAE [25].

2.1 The Meaning of Gesture

McNeill [18] distinguishes four types of gestures: iconic, metaphorical, deictic and beats (see definitions in appendix A.1). The first two types are described as 'representational' and encompass body movements that express or elaborate on some meaning communicated through co-occurring speech. Research has shown that this division of functional types needs to be expanded to better encapsulate the multi-dimensional nature of gesture. The first argument concerns identifying meaning beyond a gesture's referentiality—the quality of referring to some external entity. Several annotation schemes (see subsection 2.2) partially define gesture meaning in terms of its pragmatic function. Although no clear, widespread conception of 'pragmatic function' seems to exist within the field, we can understand it as the functional purpose of a gesture in an interactive setting, independent of its referential value.

Secondly, there is a general misconception concerning 'beat' or non-referential gestures. In fact, many studies have described them as basic, 'meaningless' movements that are rhythmically aligned with co-occurring speech. For example, Dimitrova et al. (2016) claim that beat gestures "represent rhythmic nonmeaningful hand movements" [7]. With the introduction of pragmatic functions, beat gestures can be recognised as carrying pragmatic meaning. Results from Kong et al.'s (2014) study show that 21.4% of the identified beat gestures guided and controlled the flow of speech, whereas the remaining 78.6% served to reinforce "intonation or prosody of speech" [16].

Thirdly, many interpret McNeill's functional types to be mutually exclusive, that is, a gesture can only be annotated under a single label. This is not representative of the meaning that gestures carry in real-life scenarios. The MultiModel MultiDimensional (M3D) labeling system manual provides a clear example [22]. Figure 1 shows a speaker shows a speaker making a series of counting gestures while pointing to a tree. The manual interprets this instance as having a dimension of iconicity (represented by the counting) and of deixis (due to the pointing motion)[1].

Figure 1: Speaker producing a counting gesture directed towards a tree

[1] Retrieved from https://www.youtube.com/watch?v=fbGfe78_2jc

The next Section shows that McNeill's types constitute the basis for emerging theories and conceptions of gesture meaning. Nonetheless, for a coding scheme to genuinely capture meaning, it must consider the abovementioned points.

2.2 Gesture Annotation Schemes

Crowder [5] expands on McNeill's [18] and Crowder and Newman's [6] work by introducing a coding scheme that combines both of their divisions of gesture types—one based on the "degree of representational features" [18] and another based on functional value. The functional value of a gesture represents the degree to which it extends upon or adds meaning to co-occurring speech. A gesture is classified as redundant if it adds no new information to what is being verbally expressed, as enhancing if it extends the meaning of the language in some significant way[2], and as content-carrying if it provides new information not contained in the speech. Although Crowder's proposal has not been realised as an actual coding scheme for video annotation, the idea of a dual classification of gestures based on their representational and functional meanings has influenced emerging coding schemes. A schematic overview of this scheme can be found in appendix A.2.

Kongb et al.'s study [16], which utilizes the DoSaGE database, examines how age and linguistic performance relate to the frequency of gesture employment [16]. The annotations are based on gesture form and function, coupled with linguistic information from speech. These three factors comprise the meaning of a gesture. The categories of form are the following: iconic, metaphoric, deictic, emblem, beat and non-identifiable. Once again, this shows a clear influence of McNeill's work [18]. The functions of gestures are listed as 1) providing additional information to the message conveyed, 2) enhancing the speech content, 3) providing alternative means of communication, 4) guiding and controlling the flow of speech, 5) reinforcing the intonation or prosody of speech, 6) assisting lexical retrieval, 7) assisting sentence re-construction and 8) no specific function deduced. This coding scheme provides an elaborate yet easy-to-implement account of meaning. However, relying on speech transcripts to obtain linguistic information poses a challenge for this project, given that only low-frequency audio–from which no transcripts can be made–is available in the ConfLab dataset.

Trujillo et al. [26] suggest that 1) regarding speech and gesture as equally important aspects of communication and 2) studying their alignment will result in a better understanding of how humans adapt to contextual interaction requirements. Each gesture is labelled as either representational (iconic or metaphoric [18]), abstract deictic, pragmatic (beats, emphatics and stance modifiers [15]), emblem (generally recognised gestures such as a 'thumbs-up') or interactive (e.g. a deictic that directly addresses the other speaker). The main pitfall of this scheme is that the categories above are mutually exclusive, that is, it is assumed that each gesture pertains to a single category.

In contrast, the MultiModal MultiDimensional (M3D) labelling scheme [22] rejects the idea that gesture types are mutually exclusive. It provides a multidimensional coding annotation framework where gesture meaning is based on two independent tiers: semantic and pragmatic. See appendix A.1 for definitions of these terms. These tiers parallel both Crowder's [5] and Kong et al.'s [16] approaches. The semantic meaning of a gesture concerns its referentiality[3] and its categories (iconicity metaphoricity, deixis, emblem and beat/non-referential) are once again based on McNeill's functional dimensions [18]. The pragmatic dimension refers to the function of a gesture with respect to co-occurring speech. This tier contains five function types: speech act marking, operational marking,

[2]The definition of this term remains vague throughout Crowder's paper. It can be understood as an instance where the gesture supports the spoken content by visually representing some complex concept. For instance, Crowder provides the example of a child making an *o* shape with their hand to represent the concentration of rays in the sun [5]

[3]Referred to as the 'degree of representational features' by Crowder and as 'form' by Kong et al.

stance-taking marking, discourse organisation, and interactional marking.

The main strength of this scheme is that it is available as a template that can be loaded into the ELAN annotation tool (see Figure 2). Moreover, there are online tools available such as an in-depth manual on how the system works [22] and a series of tutorial videos[4]. Therefore, I believe the M3D system is the most accessible and comprehensive coding scheme available so far.

2.3 Transformers for Video Classification

Two types of architectures were considered for this project: deep convolutional networks ([10], [4]) and transformer-based architectures ([2], [25]). While convolutional networks have traditionally been the standard for computer vision tasks, recent research indicates that transformer architectures can be highly effective for video classification tasks, often surpassing the performance of state-of-the-art CNN-based methods.

Dosovitskiy et al. show that Vision Transformer (ViT) attains more favourable results than state-of-the-art CNNs at a lower pre-training cost [8]. They argue that these results can only be achieved on large-scale datasets because ViT has fewer image-specific inductive biases than convolutional models. This poses a problem for our video classification task, as it involves an extremely small dataset of 1.1k clips (see Section 4.2). However, Arnab et al. show that video vision transformers can be effectively trained on small datasets by leveraging pre-trained image modes and applying regularisation methods during training [2]. ViViT is a pure-transformer architecture that extracts a sequence of spatiotemporal tokens from the input video and computes multi-headed self-attention. Arnab et al. propose multiple model variants with differing approaches for tokenisation and factorisation of the input video's spatiotemporal dimensions. As such, ViViT provides an adaptable framework that can be trained on a wide variety of datasets, including smaller-scale ones.

Tong et al. propose another transformer-based model inspired by ImageMAE [13]. VideoMAE uses a masked autoencoder (MAE) architecture with ViT [8] as a backbone to perform self-supervised video pre-training [25]. Tong et al. show that VideoMAE yields favourable results with an extremely high masking ratio[5] of around 90% to 95%. Furthermore, VideoMAE achieves relatively good results when trained on small datasets. For instance, it achieved a Top-1 Accuracy score of 62.6% with 3.5k clips from the HMDB51 dataset [17] and 91.3% with 9.5k clips from the UCF101 dataset [24]. Most importantly, VideoMAE outperforms other state-of-the-art methods—including ViViT—when trained on Something-Something V2 [12] and Kinetics 400 [14]. Comparison tables extracted from the original paper can be found in appendix A.3.

3 Approach

This section summarises the approach taken for this study based on three main pillars: selecting the annotation framework, annotating video data, and fine-tuning a video classification model on the annotated data.

Selection of annotation framework. Based on the information collected in Section 2.1, gesture meaning should be multi-dimensional[6] and account for both referentiality and pragmatic meaning. The selected coding scheme should 1) enable us to annotate a gesture's meaning regardless of the physicality of its motion, and 2) be reproducible in annotation tools such as Covfee [20] and ELAN [28]. Ultimately, the M3D coding scheme was chosen as the best fit. Other coding schemes (namely

[4] Visit https://m3d.upf.edu/home for further information.

[5] A high masking ratio indicated that a large portion of the input video is masked during training, increasing the difficulty of video reconstruction.

[6] In the sense of allowing annotation of a gesture under multiple labels.

[5] and [16]) also fulfill the requirements mentioned above. However, the M3D system provides additional benefits—namely a publicly available manual on how to use M3D for annotation and a template that can be directly loaded into ELAN—that make it the most optimal choice.

Data annotation. The selected annotation tool (software) and scheme are used to annotate video data from the Conference Living Lab (ConfLab) dataset [28]. A detailed explanation of this dataset can be found in Section 4.1. The EUDICO Linguistic Annotator (ELAN) tool was chosen to perform the annotations. Not only is it the tool most frequently used among the surveyed studies (see for instance [16] and [22]), but it also contains functionality to load pre-made templates of coding schemes. This provides a more time-efficient solution than manually creating tiers with their required internal dependencies.

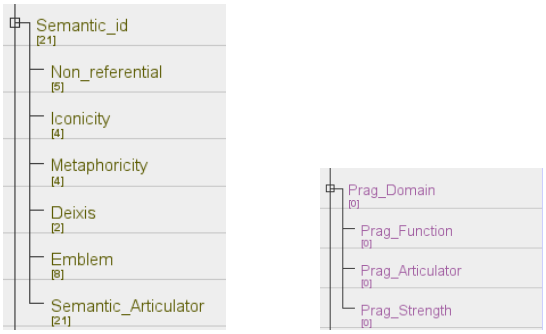
Model fine-tuning. Section 2.3 shows that transformer-based architectures can be effective when trained on small datasets. Due to time constraints, it was decided to fine-tune an existing machine-learning model rather than develop a new one. As it seemingly outperforms its predecessor ViViT [2], VideoMAE [25] was chosen.

4 Experimental Setup

This section explains how the research was conducted. Section 4.1 focuses on the data annotation process that was carried out using ELAN [28]. It explains the main challenges encountered in this process. Section 4.2 shows how the annotated data was pre-processed before fine-tuning the model.

4.1 Data Annotation

ELAN was chosen as the annotation tool, as mentioned in Section 3, allowing the pre-existing M3D template[7] to be loaded. The hierarchical structure of the semantic and pragmatic sub-tiers as they are shown in the application's interface can be seen in Figure 2.



(a) Semantic tier (b) Pragmatic tier

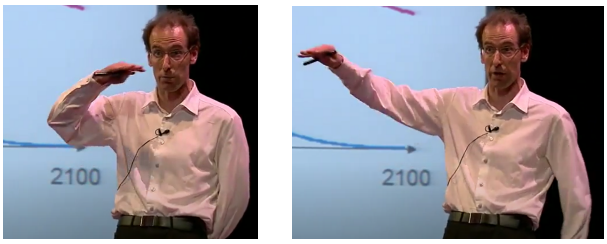
Figure 2: Hierarchical structure of the Meaning Dimension in ELAN. The semantic and pragmatic meanings are annotated independently of each other.

The ConfLab dataset contains 1 hour, 18 minutes and 10 seconds of recorded footage of a conference from an overhead view[8] (see Figure 15 from appendix A.5) [28]. Coupled with the fact that the conference had 48 attendees, annotating the entire dataset was not viable due to the project's

[7] This template can be downloaded from https://osf.io/ankdx/.

[8] Skeleton data of the participants and low-frequency recordings of their conversations are also included.

time constraints. As such, approximately 40 minutes of footage have been annotated for 23 out of the 48 attendees. Each annotation file corresponds to a single attendee—if 5 of the chosen attendees are shown in a video segment, 5 separate annotation files are created.



(a) (b)

Figure 3: The speaker utters "It's saying that you do some geoengineering for *a little while*" as he draws a vertical line with his arm. This line represents the abstract concept of time. Example retrieved from the M3D manual [22].

The main bottleneck of the annotation process is the lack of clear audio from which to retrieve the utterances of the speakers. As explained in the M3D manual [22], the meaning of a gesture is not independent of its co-occurring speech—it is partially constituted by it. For instance, a gesture is labelled as metaphoric if it illustrates an abstract concept used in speech (see Figure 3). Without knowing whether a gesture represents an abstract or concrete concept, one must assume it is either iconic or metaphoric.

Three key factors shaped the annotation process. Firstly, the semantic meaning of gestures is annotated based on strong assumptions based on their observed form and the overall trajectory of the associated articulators. Gestures that involve an index finger pointing toward some concrete or imaginary entity in space are automatically classified as deictic. Moreover, any gesture representing some form of counting is assumed to be iconic. Similarly, all rotational hand and arm motions are labelled as metaphoric. This is based upon McNeill's [18] claim that "rotation is a frequent gestural metaphor for trying". Secondly, due to time constraints and the lack of high-quality audio, pragmatic meanings were not annotated in this project. The M3D manual lists a total of 23 pragmatic functions [22], which can be found in appendix A.4. These functions are inherently dependant on speech content and, unlike semantic meaning, cannot be partially inferred from a gesture's form and trajectory. In fact, the M3D manual highlights the importance of annotating pragmatic meaning in parallel with other tiers[9] that solely account for verbal content [22]. Further discussion on the ambiguity of gestures without co-occurring speech and the implications of omitting pragmatic meaning in this project can be found in Section 5.3. Thirdly, since a single gesture can have multiple dimensions of semantic meaning [22], this is a multi-class, multi-label classification problem. However, due to the challenge of annotating gestures without co-occurring speech, it was decided only to annotate one dimension per gesture. Although this approach does not comply with the criteria established in Section 2.1, it eased the annotation and model training processes. The VideoMAE model used in this study is a multi-class classification model that predicts one label per input.

[9]These kind of tiers, such as Information Structure (IS) marking, are not discussed in this paper. However, they can be found in the original M3D template and the manual.

4.2 Data Preprocessing

The final dataset contains 1119 videos of unique gesture units (see definition in appendix A.1). The original videos from the ConfLab dataset were manually cropped using the video editing program PowerDirector (see Figure 16 in appendix A.5). Each cropped video was trimmed to generate clips for each gesture unit. The start and end times for each gesture unit were retrieved from the ELAN annotation files. The filenames of these clips follow a specific format, defined as `{annotation}_{clipID}_video_{attendeeID}_{annotation order}`. For example, the filename `abstract-deixis_21_video_37_1` can be interpreted as follows:

- `abstract-deixis`: The semantic meaning annotation of the gesture.
- `21`: The unique clip identifier.
- `video_37`: The video related to attendee 37.
- `1`: The first video annotated for attendee 37.

The dataset was split based on the 70-15-15 rule: 70% for the training set, 15% for the test set, and 15% for the validation set. The rest of the pre-processing was done by following this video classification task guide. VideoMAE uses an asymmetric encoder-decoder architecture that takes downsampled video frames as input [25]. The original clips were resized using the **`image_processor`** associated with the pre-trained model. The duration of each clip is calculated using the following formula:

$$\text{clip\_duration} = \frac{\text{num\_frames\_to\_sample} \times \text{sample\_rate}}{\text{frames\_per\_second}}$$

where 'num_frames_to_sample' is a constant value defined in the model's configuration file, 'sample_rate' is a constant set to 4, and 'frames_per_second' is a constant set to 30. This formula is then applied using the built-in **`make_clip_sampler`** function from the PyTorchVideo library.

The training set was transformed using uniform temporal subsampling, pixel normalisation, random cropping and random horizontal flipping. These transformation functions are provided by the PytorchVideo library. Only uniform temporal subsampling and pixel normalisation were applied to the test and validation sets. This led to a slight augmentation of the test and validation sets, which increased by 1 and 5 samples respectively (see Figure 4). The model was fine-tuned for 10 epochs, using a batch size of 8 and a learning rate of 1e-4.

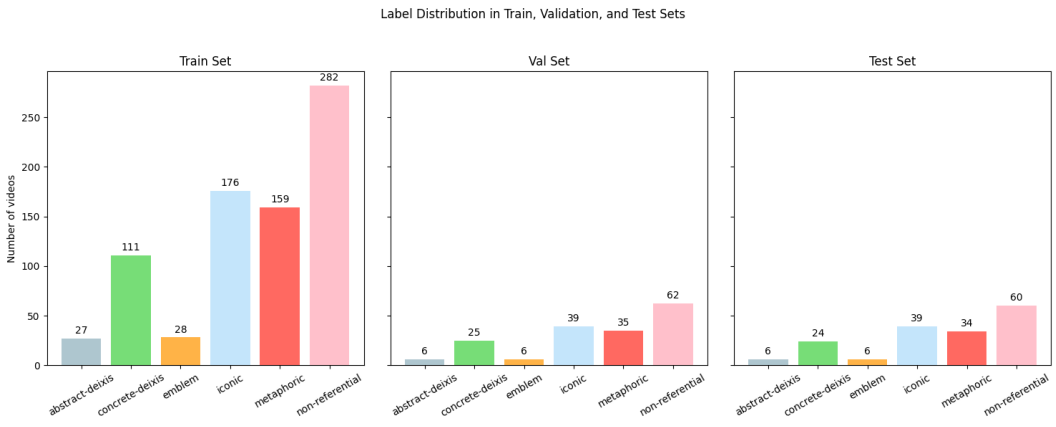


Figure 4: Dataset distribution across the train, validation and test sets.

5 Results and Discussion

This section presents and analyses the results obtained when fine-tuning the pre-trained VideoMAE [25] model with the created dataset (see Section 4.2). Section 5.1 includes Figures and tables showing relevant metrics to evaluate the model's performance, which is analysed in Section 5.2. Section 5.3 explains the limitations of the project that influenced the results. The code for pre-processing data and fine-tuning the model can be found in the GitLab repository assigned to this project [1].

5.1 Results

The model achieved an overall accuracy of 49% on the test set and 48% on the validation set. Figure 5 shows the accuracy of the model per label on the test set. Different parameter configurations than those mentioned in Section 4.2 were tested, but they mostly led to similar or slightly worse results.

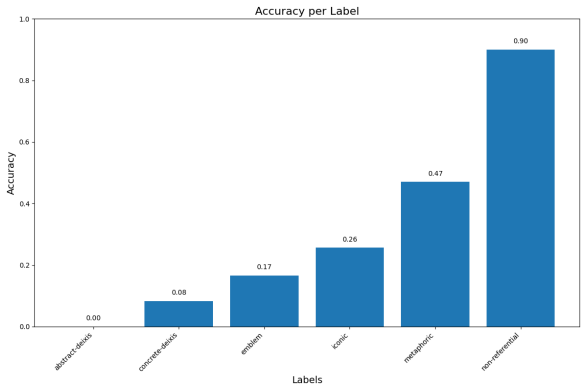


Figure 5: Accuracy per label on the test set.

The dataset was tested with three baseline classifier model strategies: stratified, most frequent and uniform (see Figure 6). Even with an accuracy score of only 49%, the fine-tuned model outperformed the other baselines.

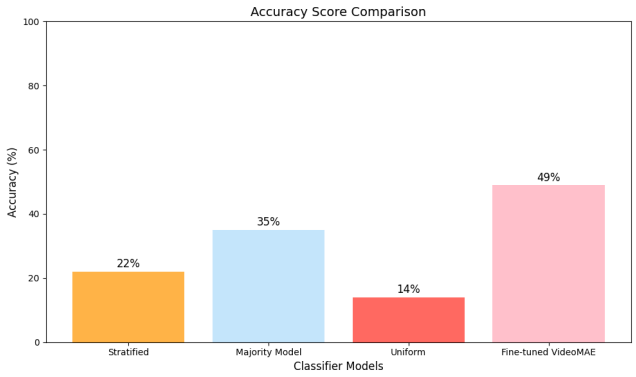


Figure 6: Comparison of accuracy scores among different baselines.

Table 1 shows the precision, recall and F1-score metrics computed on the test set. This data was obtained using scikit-learn's built-in **`classification_report`** method.

Table 1: Classification Report (see Figure 18 in appendix A.6 for the ROC curve).

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

Figure 7 presents the confusion matrix computed on the results obtained from the test set. 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.

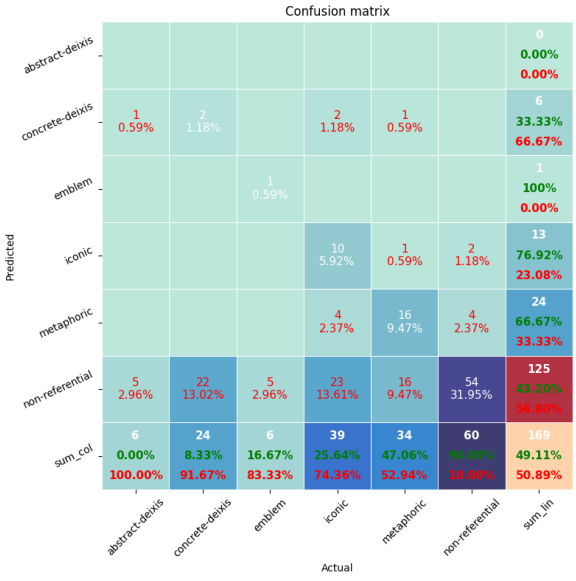


Figure 7: Confusion matrix on the test set. The numbers in the 'sum_col' row represent the number of samples per label in the test, and the percentages in green show the recall scores. The numbers in the 'sum_lin' column represent the number of predictions per label made by the model, and the percentages in green show the precision scores. Light blue cells indicate fewer number of predictions. The darker the tone of blue, the more predictions were made for a specific class. Cell colors become 'warmer'—going from dark blue to purple to red—as the number becomes high for the total number of samples in the test set. When that number corresponds to a significant portion of the dataset, the 'warm' tone becomes lighter (see the bottom right portion of the matrix).

5.2 Result Analysis

The overall accuracy, as well as the high accuracy of 90% for the non-referential class (see Figure 5), are misleading. The results indicate that the model fails to identify different gesture types. Approximately 74% of test samples were incorrectly classified as non-referential, as shown in Figure 7. The training data is imbalanced (see Figure 4), leading to a strong bias towards the non-referential category. The precision score for this class (see Table 1) highlights this bias, given that 56.8% of non-referential predictions are false positives (see Figure 17 in appendix A.6 for reference). Performing data augmentation on the training set could counter the challenge posed by this imbalance. Although various augmentation techniques such as `AutoAugment` and `RandAugment` were applied to the train set in multiple runs, none led to better results. Due to time constraints, issues concerning data augmentation could not be further investigated.

(a) Metaphoric gesture (b) Iconic gesture

Figure 8: The gesture in Figure 8a is assumed to represent the abstract concept of distance. In Figure 8b, attendee 21 gradually brings his right hand closer to him, representing the physical motion of coming closer.

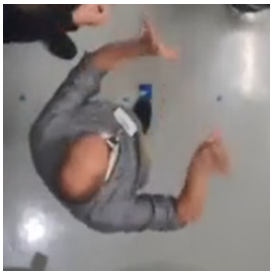
The model's inability to distinguish gesture types can thus be attributed to two factors: the bias towards non-referentiality and the similarity between different gestures (see Section 5.3 for a comparison between abstract and concrete deixis). Two examples of visually similar gestures are illustrated in Figure 8. One is labeled as metaphoric and the other as iconic, yet both were classified as non-referential by the model. Therefore, even if high-quality audio recordings or transcripts were provided for the annotation process, the model could achieve similar results to the ones provided in Section 5.1.

The VideoMAE model outperforms the three baseline models shown in Figure 6, suggesting that it generalises better than the baseline classifiers. This likely occurs because the model 1) has already been pre-trained on a similar dataset (UCF101 [24])[10], 2) has been fine-tuned on the gesture dataset as explained in 4.2, and 3) captures meaningful data patterns that the baseline models do not. The baseline models are implemented using a `DummyClassifier`, which ignores the input feature values of the training data.

[10]Tong et al. mention that VideoMAE can achieve a 90.8% accuracy score on the UCF101 dataset without using any extra data.

5.3 Limitations

(a) He brings his hands together in front of his chest.



(b) He expands the space between his hands.

Figure 9: Attendee 4 makes a motion with his hands that could be interpreted as either iconic or metaphoric, showcasing the ambiguity that gestures acquire with no access to co-occurring speech content. It was annotated as metaphoric and classified as non-referential.

The lack of high-quality audio or overall availability of verbal content from the recordings impacted the annotation process and thus performance of the model. In the examples provided by the M3D manual [22], the tiers corresponding to the meaning dimension are annotated in parallel with the speaker's utterances—primarily because intonation and emphasis are important elements for identifying beat gestures. This is not possible with the ConfLab dataset, leading to the 'Utterance' tier from the original M3D template being entirely removed.

Another issue related to the lack of high-quality audio is the ambiguity of gestures. Many strong assumptions had to be made in the annotation process concerning the types of deixis (concrete and abstract) and iconic and metaphoric gestures. For instance, if some hand motion were directed towards a point in space but did not seem to point towards some specific object or entity, it would be labelled as abstract deictic (see Figure 10a). In the absence of content from co-occurring speech, this gesture could be interpreted as concrete deictic. The key difference between these two types is that concrete deictic gestures point to some object in the physical space where the speaker stands, while abstract deictic gestures "point to imaginary objects in abstract space" [22]—which can only be inferred from speech. A pointing gesture is displayed in both Figure 10a, which shows an example of abstract deixis, and Figure 10b, which shows an example of concrete deixis. Without further contextual data[11], they will likely be classified under the same category. Due to the strong bias towards non-referentiality, the current model predicts both gestures as non-referential. However, with a larger set of deictic gestures, it may become a problem that abstract deictic gestures are misclassified as concrete deictic and vice versa.

Attendee 4 makes an 'open and close' motion with his hands in Figure 9. This gesture could be interpreted as iconic; for instance, it could represent the width or broadness of some entity. However, it could also represent the abstract concept of "expansion" or "growth", and thus be labelled as a metaphoric gesture.

Moreover, the annotation process did not fully adhere to the procedure outlined in the M3D manual, which explains that the semantic and pragmatic meanings of a gesture correspond only to its stroke. Instead, they were directly associated with the entirety of the gesture unit. Considering the limited time allocated for this project, it was unfeasible to annotate gesture phases in addition to their meaning. This problem could be solved by combining the annotation files used in this project

[11] In the sense of speech content data or cropping the video such that the target the attendee is pointing towards is shown.

with annotations of gesture phases. By identifying overlapping gesture units, one can trim each clip based on the timestamps of the gesture stroke and, by adhering to the format described in Section 4.2, name the file after the semantic meaning associated with the gesture unit.

Additionally, there were no annotations of the pragmatic meaning of gestures. It is crucial to annotate the pragmatic meaning to avoid the assumption that non-referential gestures are meaningless (see Section 2.1). In fact, most of the gestures that accompany our speech are non-referential, as seen in the distribution from Figure 4.

Finally, the size of the dataset used for training is likely too small to achieve a favourable performance of the model. VideoMAE obtains fairly impressive results on small datasets of around 3k-4k videos [25], but the current gesture-meaning dataset only contains 1124 videos. Doubling its size, optimally by adding more instances of referential and emblematic gestures, would lead to better results.

(a) Abstract deixis (b) Concrete deixis

Figure 10: Attendee 4 makes a pointing gesture that is not directed towards any particular element in his nearby space in Figure 10a. It is assumed that he is pointing to some imaginary object. He makes a pointing gesture directed toward attendee 33 in Figure 10b. Since attendee 33 is a physical entity present in the space where the conversation occurs, this gesture is annotated as concrete deictic. Both were classified as non-referential by the model.

6 Responsible Research

This section reflects on the ethical aspects of this study and the reproducibility of its methods. Section 6.1 discusses how data privacy is handled in this project. Section 6.2 describes how Large Language Models were used for and throughout the research. Section 6.3 discusses the ethical implications of this research by showing how factors such as gender negatively impact the generalisability of the results. Section 6.4 analyses how reproducible the methods used in this study are.

6.1 Data Privacy

The ConfLab dataset, owned by the Technical University of Delft, is not publicly available. To ensure that the data and privacy of the participants are protected, the model was trained using resources provided by the university, specifically the DelftBlue supercomputer.

The overhead camera perspective and low-frequency audio recordings from the ConfLab dataset mitigate the re-identification of attendees [21]. However, the format of the filenames used in the dataset (see Section 4.2) could pose a threat. Each video filename includes a unique identifier of the attendee shown in that clip. A folder containing face pictures of the participants along with their corresponding IDs was provided at the start of the study. If an unauthorized individual were to gain

access to this folder through malicious means, the filename IDs could be used to trace back and identify the faces of the attendees.

As for the limitations of the project, it is clear that high-frequency audio or speech transcripts should be made available for more accurate gesture-meaning annotation. Since access to high-frequency audio would facilitate re-identification of the participants, generating written transcripts is a safer alternative. Nonetheless, this approach still requires audio recordings of the conference that would subsequently be processed into text, raising concerns about the secure storage of this audio data.

6.2 Use of Large Language Models (LLMs)

Large Language Models, namely ChatGPT, were occasionally used throughout the project to fix code errors often caused by mismatched versions of Python packages. None of the data from the ConfLab dataset or the 1119 clips used for fine-tuning were included in the prompts. The information provided in Section 2 is entirely based on the cited sources. ChatGPT was sometimes used to paraphrase sentences from the original source and thus avoid plagiarism.

6.3 Ethical Implications

The ConfLab paper explains that "there is an implicit selection bias in the population represented in the data" because the recordings occurred in a scientific conference and participation was voluntary [21]. Furthermore, the subset of clips used in this study involves a selected number of participants. This means the dataset used for model training contains several clips of the same person. Consequently, the results may not generalise well to the rest of the original dataset and, most importantly, the general population.

Men represent 82% of the total participants in the original ConfLab dataset [21]. Approximately 78% of the attendees in the generated dataset (see Section 4.2) are male. Au et al. argue that gender differences only affect the hand-grasp speed of manual gestures and do not influence the amplitude and rhythm of such gestures [3]. Conversely, Skomroch et al. suggest that women perform "gestures with picturesque content more often" while men perform "movements in which the hands act on each other significantly more often than women" [23]. The former study was conducted with 39 men and 41 women, whereas the latter involved 49 women and 42 men. Although additional factors such as age, ethnicity and height might influence hand movement [3], Skomroch et al.'s results do suggest that an equal balance of male and female attendees might have resulted in a greater number of iconic and metaphoric gesture samples. As such, the imbalance between non-referential and pictorial (iconic and metaphoric) gestures in the dataset (see Figure 4) might cause incorrect assumptions about how women gesticulate.

6.4 Reproducibility of Methods

The ELAN annotation tool and the M3D scheme template are both publicly available and free to use. The code for pre-processing the data and fine-tuning the model can be found in the GitLab repository created for this research [1]. No data from the ConfLab dataset is available in this repository due to privacy concerns. It includes two empty folders: **'conflab-videos'**, which originally included the cropped and trimmed videos (see Section 4.2), and **'conflab-dataset'**, which included the train-validation-test division. The file named **'fine-tuned-model.pt'** is the fine-tuned model that obtained the results shown in Section 5.1. To preprocess the data and train the model, one must run the jupyter notebook named **'conflab_videomae.ipynb'**.

Most of the code is directly retrieved from the previously mentioned video classification task guide. In order to access the VideoMAE model, one must have an account in Hugging Face and create a private access token with a 'write' role. The login-in function where this token must be entered is provided in the abovementioned jupyter notebook.

7 Conclusions and Future Work

This section presents the main conclusions drawn from this study (Section 7.1) and recommendations for further research (Section 7.2).

7.1 Conclusion

This study shows the feasibility of classifying the meanings of gestures produced in densely crowded social settings using VideoMAE [25] and the MultiModal MultiDimensional (M3D) labelling scheme [22]. Previous research on gesture meaning and types demonstrates that McNeill's division of functional types [18] can be used to define a gesture's referential or representational value, but that the definition of gesture meaning should be expanded to account for pragmatic meaning. Both referential and pragmatic values of gesture are tightly coupled with the verbal content of its co-occurring speech. However, the ConfLab dataset only provides low-frequency audio recordings to preserve privacy. The lack of high-quality audio or transcripts and the ambiguous nature of gestures complicated the annotation process, leading to many strong and potentially incorrect assumptions (see Section 5.3). Therefore, it is not possible to determine whether gestures are correctly annotated.

Due to time constraints and the abovementioned lack of verbal information, gestures were annotated as having a single dimension and solely based on referentiality (see Section 4.1). This violates the selection criteria specified in Sections 2.1 and 3 since it disregards the multidimensional nature of gestures and their pragmatic meaning. As such, this study only partially analyses the classification of gesture meaning.

Most of the gestures retrieved from the ConfLab dataset are annotated as being non-referential (see Figure 4). The classifier has a strong bias towards non-referentiality due to this imbalance. As a result, more samples of scarce gesture types (mainly deictic and emblematic gestures) are required to prevent this bias and better assess the model's ability to identify gesture types. However, even if gesture meaning was annotated based on co-occurring speech, the fine-tuned model could still achieve poor results due to visual similarities between gestures (see Section 5.2). Although VideoMAE can only be trained on video data, re-annotating the dataset with access to speech content could change the distribution of the current dataset. Consequently, using high-quality audio or transcripts for annotation could influence the model's performance by potentially removing, maintaining or furthering the bias toward non-referentiality.

7.2 Future Work

Future work should explore privacy-preserving approaches to recording high-frequency co-occurring speech. Low-frequency audio helps prevent re-identification; however, gestures cannot be effectively annotated and thus classified without access to co-occurring speech. Furthermore, the pragmatic meaning of gestures should be annotated. Solely considering semantic meaning can lead to the assumption that non-referential gestures are essentially meaningless, as explained in Section 5.3. Referentiality only partially constitutes gesture meaning. To accurately evaluate a model's ability to classify gesture meanings in various social settings, it is essential to ensure that the annotations properly account for the full meaning of gestures. Since the semantic and pragmatic tiers of the

M3D coding scheme are independent of one another [22], a possible solution would be to train two separate classifiers: one that predicts semantic meaning and one that predicts pragmatic meaning. Finally, this task should be approached as a multi-class, multi-label classification problem. The M3D system has already enabled multi-dimensional annotation of gestures, but further research on multi-class, multi-label video classifiers is required.

References

[1] Irene Aldabaldetrecu. Deciphering the secret language of gesture in soc. https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Hung_Agarwal_Kondyurin_Li/shared-deciphering-the-secret-language-of-gesture-in-soc, 2024.

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

[3] Wing Lok Au, Irene Soo Hoon Seah, Wei Li, and Louis Chew Seng Tan. Effects of age and gender on hand motion tasks. *Parkinsonâs Disease*, 2015:1â5, 1 2015.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2018.

[5] Elaine M. Crowder. Gestures at work in sense-making science talk. *Journal of the Learning Sciences*, 5(3):173–208, Jul 1996.

[6] Elaine M. Crowder and Denis Newman. Telling what they know. *Pragmatics and Cognition*, 1(2):341â376, Jan 1993.

[7] Diana Dimitrova, Mingyuan Chu, Lin Wang, Asli Ozyurek, and Peter Hagoort. Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, 28(9):1255â1269, Sep 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Merouane Ertel, Azeddine Sadqui, Amali Said, Intissar Mahmoudi, Younes Bouferma, and Nour-eddine El Faddouli. Predicting the severity of new sars-cov-2 variants in vaccinated patients using machine learning. *Journal of Theoretical and Applied Information Technology*, 05 2023.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.

[11] Valerie Freeman, Richard Wright, Gina-Anne Levow, Yi Luan, Julian Chan, Trang Tran, Victoria Zayats, Maria Antoniak, and Mari Ostendorf. Phonetic correlates of stance-taking, 10 2014.

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[15] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[16] Anthony Pak-Hin Kong, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam. A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage. *Journal of Nonverbal Behavior*, 39(1):93–111, Sep 2014.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

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

[19] Francis Quek, David Mcneill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl-Erik Mccullough, and Rashid Ansari. Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9:171–193, 10 2002.

[20] Jose Vargas Quiros. Covfee: continuous video feedback tool. https://josedvq.github.io/covfee/, 2024.

[21] 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.

[22] 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.

[23] Harald Skomroch, Kerstin Petermann, Ingo Helmich, Daniela Dvoretska, Robert Rein, Zi-Hyun Kim, Uta Sassenberg, and Hedda Lausberg. *Gender Differences in Hand Movement Behavior*. 2013. TiGeR 2013: Tilburg Gesture Research Meeting: International Gesture Workshop (GW) and 3rd Gesture and Speech in Interaction (GESPIN ; Conference date: 19-06-2013 Through 21-06-2013.

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

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

[26] James P. Trujillo, Stephen C. Levinson, and Judith Holler. A multi-scale investigation of the human communication system's response to visual disruption. *Royal Society Open Science*, 9(4), Apr 2022.

[27] Ulya Tutuncubasi, Ada Ren-Mitchell, Stefanie Shattuck-Hufnagel, Patrick Louis Rohrer, and Pilar Prieto. M3d training.

[28] Peter Wittenburg, 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.

A Appendix

A.1 Definitions and notations

The following concepts are essential for understanding the annotation of meaning in gesture:

- **Coding scheme.** A coding scheme is a framework that allows for the labeling or annotation of co-speech gestures.
- **Gesture unit.** The M3D manual defines a gesture unit as "the span of time from when the articulators leave a position of rest or relaxation to their return to a state of rest or relaxation" [22].
- **Semantic meaning.** The semantic meaning of a gesture is related to its referentiality, that is, the object or entity that the speaker refers to when enacting the gesture.
- **Pragmatic meaning.** The pragmatic meaning of a gesture is its function with regard to co-occurring speech. It is independent of its referential value.
- **Iconicity.** A gesture is iconic when it represents a physical object or entity described in speech (see Figure 11).
- **Metaphoricity.** A gesture is metaphoric when it represents an abstract concept used in speech.
- **Deixis.** Pointing gestures or gestures that signal some location in the speaker's (current or abstract) space are deictic. There are two types of deixis: concrete, when the speaker points to something located in their immediate space, and abstract when they point to some imaginary space referred to in the speech.
- **Emblem.** Emblems are gestures whose meaning has been agreed upon by a specific culture. For instance, a 'thumbs-up' gesture generally signals approval from the speaker's side in western culture. However, gestures that are emblematic for one culture may mean something entirely different for others.
- **Non-referentiality.** According to the M3D labeling manual, "movements that do not clearly visually reference the propositional content in speech are said to be non-referential" [22].

(a) (b)

Figure 11: The speaker utters "We actually descended the camera and cameraman through" as he brings his left hand downward. This motion represents the action of descending the camera and cameraman. Example retrieved from the M3D training website [27]. The original video can be found here.

A.2 Overview of Crowder's coding scheme

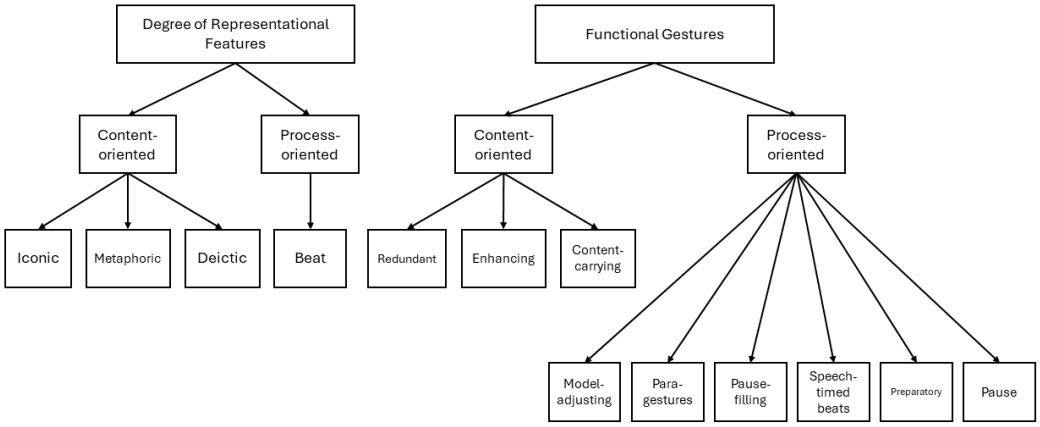


Figure 12: Overview of the coding scheme proposed by Crowder [5].

A.3 Performance comparison with state-of-the-art methods

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
TEINet$_{En}$ [40]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50	66.5	N/A
TANet$_{En}$ [41]	ResNet50$_{\times 2}$		✓	8+16	99×2×3	51	66.0	90.1
TDN$_{En}$ [75]	ResNet101$_{\times 2}$		✓	8+16	198×1×3	88	69.6	92.2
SlowFast [23]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53	63.1	87.6
MViTv1 [22]	MViTv1-B		✓	64	455×1×3	37	67.7	90.9
TimeSformer [6]	ViT-B	ImageNet-21K	✓	8	196×1×3	121	59.5	N/A
TimeSformer [6]	ViT-L		✓	64	5549×1×3	430	62.4	N/A
ViViT FE [3]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
Motionformer [51]	ViT-B		✓	16	370×1×3	109	66.5	90.1
Motionformer [51]	ViT-L		✓	32	1185×1×3	382	68.1	91.2
Video Swin [39]	Swin-B		✓	32	321×1×3	88	69.6	92.7
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	68.1	N/A
BEVT [77]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88	70.6	N/A
MaskFeat↑$_{312}$ [80]	MViT-L	Kinetics-600	✓	40	2828×1×3	218	75.0	95.0
VideoMAE	ViT-B	Kinetics-400	✗	16	180×2×3	87	69.7	92.3
VideoMAE	ViT-L	Kinetics-400	✗	16	597×2×3	305	74.0	94.6
VideoMAE	ViT-S	*no external data*	✗	16	57×2×3	22	66.8	90.3
VideoMAE	ViT-B		✗	16	180×2×3	87	70.8	92.4
VideoMAE	ViT-L		✗	16	597×2×3	305	74.3	94.6
VideoMAE	ViT-L		✗	32	1436×1×3	305	**75.4**	**95.2**

Figure 13: VideoMAE outperforms ViViT and other state-of-the-art methods on Something-Something V2 [25].

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [78]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [41]	ResNet152		✓	16	242×4×3	59	79.3	94.1
TDN$_{En}$ [75]	ResNet101		✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L		✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [51]	ViT-L		✓	32	1185×10×3	382	80.2	94.8
Video Swin [39]	Swin-L		✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [77]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
MaskFeat$\uparrow_{352}$ [80]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [69]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [23]	R101+NL		✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [22]	MViTv1-B		✗	32	170×5×1	37	80.2	94.4
MaskFeat [80]	MViT-L		✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B		✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L		✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H		✗	16	1192×5×3	633	**86.6**	**97.1**
VideoMAE$\uparrow_{320}$	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
VideoMAE$\uparrow_{320}$	ViT-H		✗	32	7397×4×3	633	**87.4**	**97.6**

Figure 14: VideoMAE outperforms ViViT and other state-of-the-art methods on Kinetics 400 [25].

A.4 List of pragmatic functions in M3D

This section summarises the pragmatic functions in section 5.2.1 of the M3D manual [22]. The following definitions and examples are directly quoted from the manual. It distinguishes five independent dimensions of pragmatic meaning: speech act making, operational marking, stance-taking marking, discourse organisation and interactional marking.

Speech act marking

Gestures that express what the speaker intends to achieve.

- **Directives**: "commands, requests, challenges, invitations (e.g., 'Can we watch the game?')".
- **Representatives**: "assertions, statements, claims, suggestions (e.g., 'My team is the best.')".
- **Expressives**: "greetings, apologies, congratulations, condolences, giving thanks (e.g., 'Congrats on winning the game!')".
- **Commissives**: "promises, oaths, pledges, threats, vows (e.g., 'I bet you 5 dollars that my team will win.')".
- **Declarations**: "blessings, firings, baptisms, arrests (e.g., 'I now pronounce you the official winners of the competition.')".

Operational marking

Gestures that express affirmation (e.g. shaking the head up and down) or negation (e.g. wagging the index finger side-to-side).

Stance-taking marking

Gestures that express the speaker's stance in terms of their "personal feelings, attitudes, value judgements, or assessments" [11].

- **Affective stance**: "positive, negative, neutral".
- **Epistemic stance**: "certainty, uncertainty, ignorance, approximation, evidentiality".
- **Politeness stance**: "polite, non-polite, impolite".
- **Agreement**: "agreement, disagreement, confirmation, incredulity, obviousness".
- **Cooperation**: "checks for understanding, opinion".

Discourse organisation

Gestures that mark the structure of discourse.

- **Anaphoric marking**: "gestures produced without the referent in speech".
- **Abstract temporal deixis**: "pointing to a segment of speech uttered before or to be uttered in the future".
- **Linking**: "connecting a sentence with a previous sentence, e.g. 'As I was saying', 'However'".
- **New sequence**: "opening of a discourse sequence".
- **End sequence**: "end of a discourse sequence".
- **Parenthetical**: "parenthetical digressions, e.g. when giving examples".
- **Listing**: "punctuating items in a list".
- **Sequencing**: "description of the order of events ('first', 'second', 'third', etc.)".

Interactional marking

Gestures used to regulate discourse between speakers, particularly in terms of turn-taking.

- **Turn-concession**: "pointing to the listener, nodding to indicate their ability to take the turn".
- **Turn-hold**: "holding up a hand to stop the listener from interrupting, asking them to wait".
- **Turn-demand**: "raising a hand to express the desire to take a turn".

A.5 Preprocessing data from the ConfLab dataset

Figure 15: Image extracted from a video segment that showcases the top-down view used in ConfLab [21].

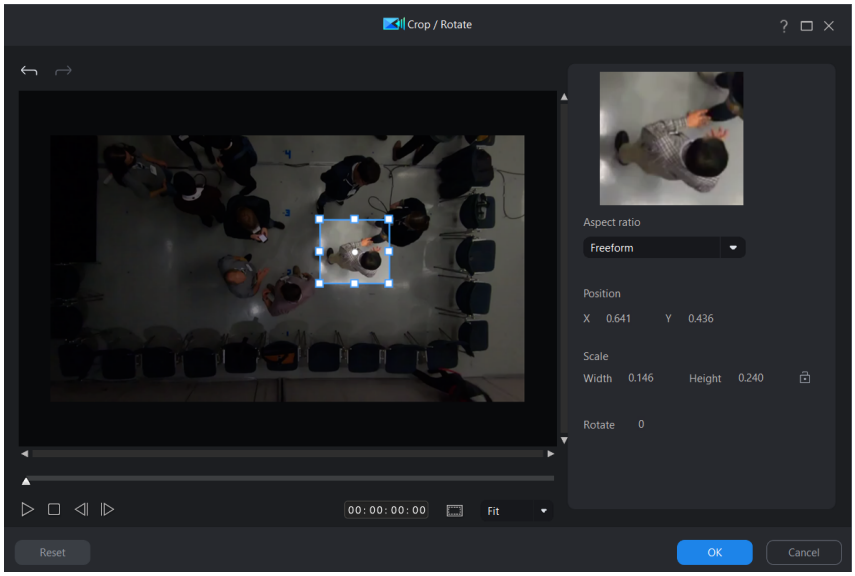


Figure 16: Image showcasing how the original ConfLab videos were cropped to create separate clips per attendee.

A.6 Results and Figures

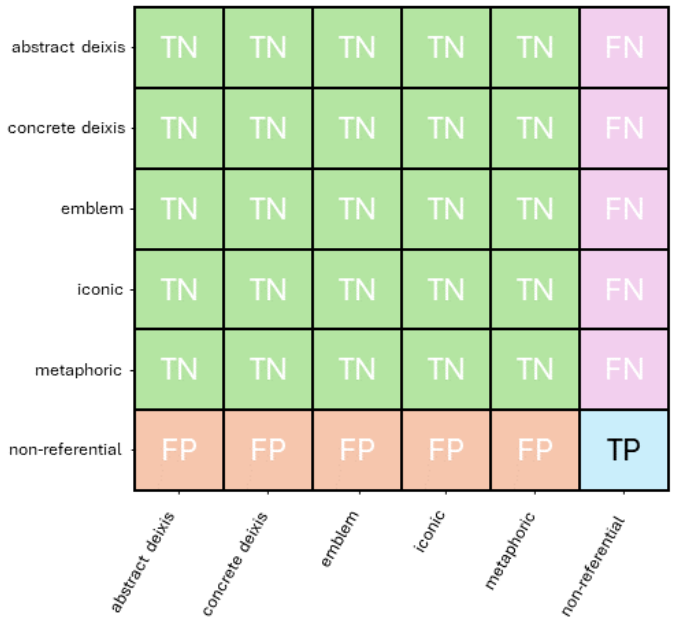


Figure 17: Confusion matrix for multi-class classification [9].

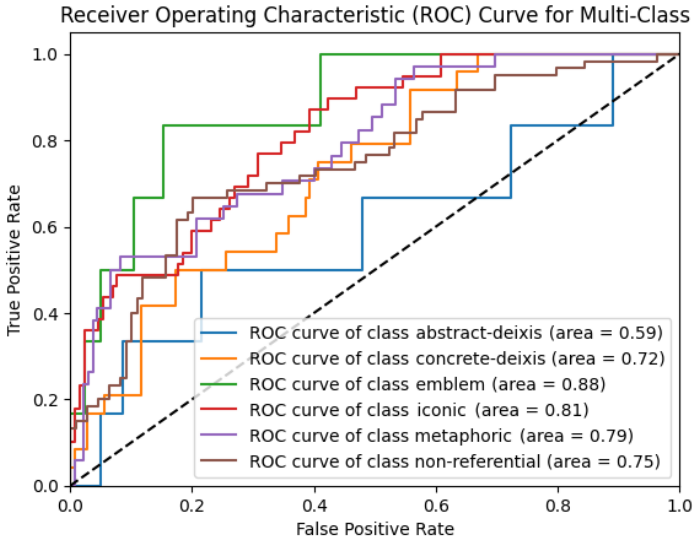


Figure 18: ROC curve for multi-class.