

Annotating and Recognizing the Palm Up Open Hand Gesture in Dutch

Fenna Berger, Alan Cienki, Lucia Donatelli

Vrije Universiteit Amsterdam / Amsterdam, Netherlands

f.berger@student.vu.nl, l.e.donatelli@vu.nl

Abstract

Recurrent gestures are conventionalized gestures with a stable relationship between form and meaning. One such gesture is the palm up open hand (PUOH) gesture, typically produced by extending the fingers and rotating the lower arm so that the palm faces upward. The gesture is attested across many languages and cultures and is often used to present a point, yet little is known about the forms and functions it takes in Dutch. Automatic gesture recognition could reduce the manual labor that quantitative gesture studies require, but recognizing this gesture is difficult, partly because of the large amount of variation present in its form. This paper compares two gesture recognition approaches on a newly collected corpus of naturalistic Dutch dyadic conversation: a static model that takes single images as input and a dynamic model that takes video clips as input. Both models are trained and tested with varying amounts of gesture-form variation included in the data. The static model is found to be less affected by this variation and to yield overall better results than the dynamic model. Our results illustrate some of the central challenges that gesture recognition systems face in recognizing the PUOH gesture and offer insights that may improve performance in future studies of multimodal, face-to-face behavior.

1 Introduction

Gestures are an important part of face-to-face communication. They carry pragmatic and semantic information that adds a further layer of meaning to spoken language, and are attested across cultures (Kendon, 2004; Kita, 2009), ages (Feyereisen and De Lannoy, 1991), and communicative contexts (McNeill, 1992; Chu et al., 2014), including in children at the one-word stage (Iverson and Goldin-Meadow, 2005) and in congenitally blind speakers talking to blind listeners (Iverson and Goldin-Meadow, 1998; Chu et al., 2014).

Recurrent gestures are one gesture type that has only recently attracted sustained attention in gesture studies: a gesture is recurrent if it “is used repeatedly in different contexts and its formational and semantic core remains stable across different contexts and speakers” (Müller, 2010, p. 41). Qualitative work on recurrent gestures exists (Ladewig, 2014; Bressem and Müller, 2014), but quantitative work is rare, as it typically requires large amounts of costly, time-consuming manual annotation.

A possible solution is to use automatic gesture recognition models to support annotation, for instance by flagging candidate gestures for human annotators to label. However, such models are difficult to train, in part because of the high degree of variation present within and across individuals producing the same gesture; a recognition model must generalize across this variation to classify gestures correctly.

This paper investigates the broad question of how gesture recognition models can recognize recurrent gestures in Dutch, focusing on the palm up open hand (PUOH) gesture, in which the speaker rotates the lower arm to present the palm to an interlocutor, often to make a point. A preliminary analysis of our newly collected corpus of Dutch dyadic conversation found this gesture to be highly frequent, consistent with prior qualitative work showing it is used across spoken and signed languages (Müller, 2004; Cooperrider et al., 2018), but no existing work was found on the forms this gesture takes specifically in Dutch. This motivates two more specific research questions:

RQ1: What forms does the palm up open hand recurrent gesture take in Dutch?

RQ2: How can gesture recognition models recognize these gestures, in particular with respect to (a) individual variation and (b) limited training data?

This work is exploratory: although the PUOH gesture is among the most extensively studied recurrent gestures qualitatively, no prior study was found that attempts to detect it automatically with machine learning. Given this, and given the small size of the available data set, the goal is not to build the best-performing model, but to compare two model types with a focus on understanding how they work, which factors affect their performance, and which challenges they face.

One such factor that may impact performance is variation in gesture form. The different forms of the PUOH gesture lie on a continuum, ranging from a small finger lift to an extension of the entire arm (Cienki, 2021). A model must generalize across these variants to recognize the gesture reliably. A second factor is the limited availability of training data, which typically induces class imbalance, since speakers spend far more time not producing a PUOH gesture than producing one. Restricting the positive class to fewer form variants reduces this imbalance but further limits the amount of positive training data, creating a trade-off that this work investigates directly.

The contributions of this paper are threefold: (1) a small annotated corpus of naturalistic Dutch dyadic conversation, part of which will be made available to other researchers upon request, in line with participant consent; (2) an empirical characterization of the forms the PUOH gesture takes in Dutch (RQ1); and (3) two gesture recognition models, trained and evaluated across conditions that vary in form-variation and data-imbalance, together with an analysis of the trade-offs between them (RQ2).

2 Related Work

Recurrent gestures. Interest in recurrent gestures is a relatively recent development within linguistics, although related phenomena were long studied under different names in rhetoric and acting (Loeb, 1912; Mosher, 1916; Ott, 1902). Early linguistic treatments used terms such as *pragmatic gestures* (Kendon, 1995; Streeck, 2005, 2009) or *interactive gestures* (Bavelas et al., 1992, 1995); Ladewig (2014) argued these terms fail to capture the language-like properties of such gestures, and the term *recurrent gesture* was later coined within the “Towards a Grammar of Gesture” project (Müller et al., 2013). Recurrent gestures resemble emblematic gestures in showing a stable form–

meaning relationship, but unlike emblems they always act on or accompany a verbal utterance rather than replacing it (Bressem and Müller, 2014). Recurrent gestures can serve either a referential function (depicting objects, actions, or events, typically co-expressed with speech) or a meta-communicative function as a speech performative, acting on the discourse structure or expressing an attitude (Ladewig, 2014; Teßendorf, 2014).

The palm up open hand (PUOH) gesture. The PUOH gesture combines a rotation of the lower arm, so the palm faces upward, with an extension of the fingers, and occurs frequently across spoken and signed languages (Cooperrider et al., 2018). Müller (2004) described it as metaphoric, rooted in the practical actions of giving and receiving, with giving-related senses (e.g., presenting an object for joint inspection) and receiving-related senses (e.g., expressing not knowing). Kendon (2004) termed it the “open hand supine” and distinguished palm presentation, palm addressed, and palm lateral variants, associated respectively with explaining or summarizing, acknowledging or offering, and expressing unwillingness, obviousness, or a rhetorical stance. Quantitative studies remain rare: Chu et al. (2014) found that palm-revealing and conduit gestures (subtypes of PUOH) made up about 24% of all gestures produced while participants defined phrases and solved social dilemmas, and that their frequency correlated with participants’ empathy scores. Jehoul et al. (2017), analyzing roughly eight hours of Flemish Dutch conversation, found the palm lateral variant to be the most frequent marker of obviousness, supporting Kendon’s account. No existing work was found on which PUOH forms occur in Dutch specifically, which the present corpus addresses.

Gesture recognition. Automatic recognition is one way to reduce the annotation cost that limits quantitative gesture research (Xu, 2017; Zou et al., 2018; Tateno et al., 2019; Ojeda-Castelo et al., 2022). Within hand-gesture recognition, vision-based approaches (Sinha et al., 2019) are more common than sensor-based (glove) approaches (Kim et al., 2019), since gloves are more invasive and may alter how participants gesture. A further distinction is between *static* recognition, which classifies poses (commonly using SVM, k-NN, or clustering methods; Cheok et al., 2019; Tharwat et al., 2015; Kurdyumov et al., 2011), and *dynamic* recognition, which classifies movements over time

(commonly using dynamic time warping or Hidden Markov Models; [Al-Shamayleh et al., 2018](#); [Darrell and Pentland, 1993](#); [Chen et al., 2003](#)). Recognized challenges in this field include temporal variance (the same gesture produced at different speeds), spatial complexity (the same gesture produced at different sizes, locations, or orientations), sensitivity to camera angle, the cost of annotated training data, and the difficulty of segmenting gestures that occur in continuous, sometimes cyclic sequences ([Nam and Wohn, 1996](#)).

Variation as signal, not noise. Both linguistics and NLP have tended to treat data variation as noise to be eliminated, despite variation being a core feature of language use through which speakers express identity and social affiliation ([Curry et al., 2024](#)). [Schmidt \(2010\)](#) distinguishes micro- (individual), meso- (group), and macro- (standard-language) levels of linguistic “variation space,” and observes that NLP typically models only the macro-level, treating other variation as disruptive ([Plank, 2016](#)). [Lutgen et al. \(2026\)](#) argue that data selection should be guided by which varieties should be represented rather than by data volume alone, illustrated by [Ramponi \(2024\)](#)’s speaker-centric critique of NLP for non-standard Italian. In the specific case of PUOH gestures, [Cienki \(2021\)](#) argues that existing gesture-description systems based on external features (hand shape, palm orientation, direction and manner of motion; [Stokoe Jr, 1960/2005](#); [McNeill, 1992](#); [Bressem et al., 2013](#)) fail to capture the continuum linking different PUOH forms, and proposes instead the internal, joint-based kinesiological system of [Boutet et al. \(2018\)](#), which describes gestures in terms of flexion/extension, abduction/adduction, rotation, and supination/pronation at each joint. This system is adopted for the annotation scheme used in the present study (Section 3).

3 Data

3.1 Collection

Data collection followed the methodology of [Cienki et al. \(Forthcoming\)](#) (briefly described in [Cienki et al., 2025](#); [Prokofyeva et al., 2025](#)) and was approved by the host institution’s Research Ethics Review Committee. Recordings took place over three days in classrooms with a neutral background, with two participants seated at a 90 angle in front of a table holding a laptop. Three GoPro cameras (two HERO12 Black, one HERO9 Black)



Figure 1: Schematic representation of the setup used for data collection as described in Sec. 3.

were used (one overview camera and one individual camera per participant) together with a lapel microphone per participant (Fig. 1).

Ten pairs of native Dutch-speaking students who already knew each other (N=20, 13 female) took part, recruited on a university campus and compensated with a gift card. Each pair first watched a short video about artificial intelligence (AI)(with a Dutch voiceover) as a shared conversational prompt, then discussed a series of questions for up to 20 minutes. Gestures were not mentioned to participants at any point, to avoid drawing attention to them. Minor data-collection issues occurred, including one camera losing roughly half of one session’s recording due to an SD-card fault, a camera tilting out of position on the first day, and a brief interruption to charge equipment; two conversations were briefly interrupted by the experimenter for unrelated reasons (a participant consulting a phone, and a participant becoming distracted by suspicion about the study’s purpose).

3.2 Annotation

Recordings were synchronized in Adobe Premiere Pro and annotated in ELAN ([Max Planck Institute for Psycholinguistics, The Language Archive, 2026](#)), marking the full gesture (stroke, hold, retraction) where possible. Annotations were divided into a positive class (PUOH gestures) and a negative class (all other (non-)gestures), each with four subclasses. Following the joint-based description in [Cienki et al. \(2025\)](#), positive-class instances were labeled by which parts of the arm and hand were involved: thumb lift (TL; thumb only), hand turnout (HT; wrist rotation), open hand diagonal (OHD; elbow rotation/extension), and palm up open hand (PUOH; shoulder rotation/extension), treated as increasingly larger points on Cienki’s continuum. Negative-class instances were labeled as other gesture (OG), self-adaptor (SA; e.g., fidgeting, adjusting hair), no gesture (NG), or unsure (US; gestures resembling PUOH in form but



Figure 2: Example image of a participant making a positive-class gesture (affirmative PUOH; Cienki et al. (2025)) as used in the static model data.



Figure 3: Example image of a participant making a negative-class gesture (non-PUOH) as used in the static model data.

judged, by native-speaker intuition, to serve a different function, e.g. iconic). For the static model, one image was manually captured at the hold of each gesture (or, where no clear hold existed, at the point where the hands were most in focus); for the dynamic model, video clips were extracted automatically from the annotation boundaries, with individual per-participant videos trimmed from the overview and side cameras and mirrored for right-seated participants to control for handedness effects related to seating position. Figure 2 shows an example of a screenshot of a positive-class instance used for the static model. Figure 3 shows the same for a negative-class instance.

The full corpus comprises 3h04m40s of recording, 1,812 annotated instances (567 positive, 1,245 negative; Tables 1 and 2), with positive clips averaging 1.58s and negative clips 1.90s in duration. Within the positive class, open hand diagonal was the most frequent subclass (298 instances) and thumb lift the least frequent (12); within the negative class, other gesture was most frequent (580) and unsure least frequent (139). The number of ges-

Class	Subclass	Total
Positive	Thumb lift	12
	Hand turnout	114
	Open hand diagonal	298
	Palm up open hand	143
	Total	567
Negative	Other gesture	580
	Self-adaptor	247
	No gesture	279
	Unsure	139
	Total	1245
Overall	Total	1812

Table 1: Total number of occurrences of classes and subclasses of palm up open hand (PUOH) gestures. Positive indicates affirmative PUOH gesture; negative indicates another gesture type.

	Min	Max	Avg	Total
Recording length	10:20	20:19	18:28	3:04:40
Positive instances	4	135	57	567
Negative instances	41	190	125	1245

Table 2: Corpus statistics across ten recording sessions (min/max/avg per session, and totals). Instances refer to PUOH gestures.

tures produced varied considerably across participants (22–197 total gestures, excluding the session with the camera fault; average 98.17), as did the proportion that were positive-class gestures (0.05–0.50, average 0.29), consistent with known individual variation in gesture use (Chu et al., 2014); no instances were excluded from the data set on this basis, since such variation reflects genuine differences among language users.

4 Models

Two gesture recognition approaches were compared: a *static* model, which classifies single images, and a *dynamic* model, which classifies video clips. Because this project was conducted within the scope of a master’s thesis, the model selection was heavily influenced by the limited time and resources available. While this means that the results, do not reflect the abilities of state-of-the-art gesture detection models, they do shine light on some differences between gesture detection models, as well as challenges that need to be tackled in the future. All code will be made publicly available upon acceptance.

4.1 Static Model

The static model adapts Tokhirov (2023), who trained several classifiers on MediaPipe hand landmarks (21 joint coordinates per hand; Fig. 4) to

distinguish four webcam-recorded hand shapes, one of which closely resembles the PUOH hand shape (though facing the camera rather than upward). Of the three classifiers Tokhirov compared (a dense neural network, XGBoost, and random forest), the random forest performed best while also being the fastest, and is adopted here, retrained as a binary classifier on the present, more naturalistic data (which, unlike Tokhirov’s, contains variation in gesture form, viewing angle, lighting, and 20 distinct speakers).

4.2 Dynamic Model

The dynamic model adapts the convolutional architecture of Pouw (2024), itself based on Yung (2022), which uses MediaPipe Holistic to extract 29 features per 25-frame window (16 hand-related, plus rotation, facial, and body-distance features, together with their time-derivatives and standard deviations). The network has three convolutional layers (32, 64, and 128 filters) with pooling between layers, and outputs a probability of motion for each window together with probabilities for each gesture type; following Pouw (2024), a motion threshold of 0.5 and a gesture threshold of 0.9 are used to assign labels. Two changes were made to the original design: the model was trained for 5 rather than 500 epochs, owing to computational constraints (a limitation discussed in Section 8), and, since one label per clip (rather than per window) was required, the highest per-clip motion and PUOH probabilities were used to assign the clip-level label.

5 Experimental Setup

Both models were trained and tested with a 70/30 split, balanced across the eight subclasses. Conditions varied the subclasses included in the positive class (six combinations for the static model: PUOH-only, OHD-only, HT+TL-only, PUOH+OHD, OHD+HT+TL, and All-Variants, merging TL with HT owing to its low frequency) and in the negative class (with or without the Unsure subclass), to test how increasing variation and class imbalance affect performance. Owing to its longer training time, the dynamic model was evaluated on a subset of four conditions (OHD-only and All-Variants, each with and without Unsure). A follow-up experiment repeated the static-model conditions with a fixed sample size (250 images per class) to disentangle the effect of variation from

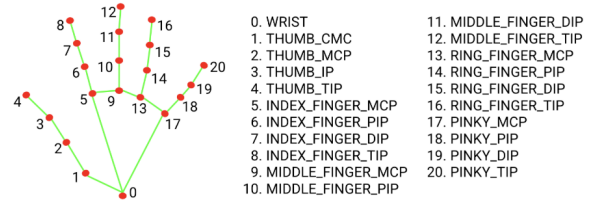


Figure 4: Overview of the 21 landmarks that are tracked by MediaPipe handtracking (source).

that of sample size.

Models were evaluated using accuracy, precision, recall, and macro F1 (the unweighted mean of per-class F1, giving equal weight to each class regardless of size), since accuracy alone can be misleadingly high under class imbalance.

6 Results

6.1 Static Model

Table 3 shows the main experiment results. The best recall and F1 (0.60, 0.75) were obtained in the All-Variants/Without-US condition; the best accuracy (0.91) in the HT+TL-only conditions and PUOH-only/Without-US; and the best precision (0.83) in HT+TL-only/Without-US. Of the single-variant conditions, OHD-only performed best. Excluding the Unsure subclass improved results in almost every condition; the one partial exception was PUOH+OHD, where With-US gave marginally higher precision and accuracy. Confusion matrices (not shown for space) indicate that in the best-performing condition, true negatives were most common (290/494), followed by true positives (108), false negatives (61), and false positives (35); in the harder HT+TL-only/Without-US condition, the model predicted a positive label for only 16 of 372 instances (7 correct), indicating a strong majority-class bias.

The follow-up experiment, which fixed the sample size at 250 images per class, yielded more homogeneous results: OHD-only (both US conditions) and All-Variants/Without-US all reached F1 and accuracy of 0.76, while All-Variants/With-US performed worse (0.70). Relative to the main experiment, precision, recall, and F1 improved in nearly all conditions once sample-size was controlled for, while accuracy decreased slightly, consistent with accuracy’s sensitivity to class imbalance; this indicates that the class-imbalance effect, not variation in form itself, was largely responsible for the poorer scores observed for smaller-subclass conditions in

Condition	With-US		Without-US	
	F1	Acc	F1	Acc
PUOH-only	0.63	0.90	0.71	0.91
OHD-only	0.68	0.84	0.74	0.85
HT+TL-only	0.50	0.91	0.59	0.91
PUOH+OHD	0.74	0.82	0.75	0.81
OHD+HT+TL	0.69	0.80	0.74	0.81
All-Variants	0.73	0.78	0.75	0.79

Table 3: Macro F1 and accuracy of the static model (main experiment). Precision and recall are omitted for space; the model achieved its highest recall (0.60) and precision (0.83) in All-Variants/Without-US and HT+TL-only/Without-US respectively.

Condition	With-US		Without-US	
	F1	Acc	F1	Acc
OHD-only	0.45	0.55	0.53	0.59
All-Variants	0.26	0.30	0.27	0.33

Table 4: Macro F1 and accuracy of the dynamic model. Precision and recall track F1 closely in all four conditions (within 0.10).

the main experiment.

6.2 Dynamic Model

Table 4 shows the dynamic model’s results across its four conditions. OHD-only/Without-US performed best on every metric (precision, recall, F1 = 0.53; accuracy = 0.59); the All-Variants/With-US condition performed worst on recall, F1, and accuracy (0.34, 0.26, 0.30) despite a comparatively high precision (0.48). Confusion matrices show that in the two OHD-only conditions the negative class was most frequently (and largely correctly) predicted, whereas in the All-Variants conditions the no-gesture label dominated predictions for all classes: in All-Variants/With-US the positive label was predicted for only 16 of 167 true positive clips (10 correct), and in All-Variants/Without-US it was never predicted at all. This indicates that increased variation in the positive class pushed the model’s PUGH-probability below the 0.9 classification threshold even for genuinely positive clips.

6.3 Individual Variation

Per-participant macro F1 was computed for the best static model (All-Variants/Without-US) and best dynamic model (OHD-only/Without-US) (Table 5). Scores varied substantially across participants: 0.36–1.00 for the static model and 0.32–0.84 for the dynamic model. For the static model, F1 correlated significantly and positively with a participant’s number of test instances (Pearson’s $r = 0.46$, $p = 0.04$), indicating sensitivity to per-

Participant	Static		Dynamic	
	F1	Sample	F1	Sample
1L	0.44	5	0.44	5
1R	0.71	38	0.52	27
2L	0.38	5	0.84	7
2R	1.00	4	0.46	9
3L	0.60	41	0.34	45
3R	0.75	33	0.32	19
4L	0.46	13	0.49	13
4R	0.42	11	0.76	14
5L	0.63	28	0.54	28
5R	0.42	18	0.66	15
6L	0.69	23	0.35	20
6R	0.79	40	0.57	31
7L	0.85	61	0.50	45
7R	0.75	24	0.50	16
8L	0.80	15	0.43	9
8R	0.81	34	0.58	32
9L	0.60	30	0.44	26
9R	0.80	33	0.32	34
10L	0.78	24	0.33	12
10R	0.36	14	0.61	10

Table 5: Macro F1 score and sample size in test set per participant for the All-Variants/Without-US condition of the static model and the OHD-only/Without-US condition of the dynamic model in the main experiment.

participant sample size; the one clear exception (a perfect F1 from only four test instances) is likely attributable to chance given the small sample. No significant correlation was found for the dynamic model ($r = -0.28$, $p = 0.23$).

7 Discussion

7.1 Interpretation

The static model’s results correlate strongly with the relative size of the positive class: conditions with more positive instances (e.g., All-Variants) yield the best recall, precision, and F1, while conditions with fewer positive instances (e.g., HT+TL-only, PUGH-only) yield the highest raw accuracy, an artifact of accuracy’s sensitivity to class imbalance rather than evidence of better classification. The follow-up experiment supports this: once sample size was controlled for, the model performed worst specifically in the condition combining maximal positive-class variation with the Unsure subclass (All-Variants/With-US), suggesting that variation in the positive class and the presence of near-miss negative instances compound one another, either because Unsure gestures resemble specific positive subclasses that only co-occur in this condition, or because greater positive-class variation broadens the model’s implicit criterion for a positive label enough to admit Unsure instances.

For the dynamic model, the reverse pattern held:

performance was better with less positive-class variation (OHD-only) despite greater class imbalance, and markedly worse under All-Variants, where the model rarely predicted the positive label at all. A plausible explanation is architectural: the static model’s features describe only hand shape, and the dynamic model’s features span the full body (arms, shoulders, face), so that arm and shoulder movements distinguishing larger PUOH variants may be informative for the static model but a source of noise for the dynamic one.

Both models showed comparable sensitivity to individual variation in the participant-level analysis, though only the static model showed a significant relationship with sample size, suggesting the two models may be sensitive to different sources of inter-participant variation.

7.2 Error Analysis

A qualitative error analysis (samples of 10 images/videos per confusion-matrix category) surfaced three recurring patterns for both models. First, both models appear more sensitive to hand *shape* (open vs. closed) than to hand *orientation* (palm up vs. down): open hands with the palm facing down were repeatedly misclassified as positive PUOH instances, and this was the single most consistent source of false positives in both models. Second, incomplete visibility of the hand due to occlusion, camera angle, or video trimming to remove the other participant, degraded performance more for the dynamic model than the static model, and for the static model specifically, false negatives disproportionately involved gestures made with only one hand, consistent with the model receiving only single-hand landmarks as input. Third, for the dynamic model, a short hold phase (where retraction followed the stroke almost immediately) was associated with missed positive instances, suggesting the model relies substantially on the hold rather than the full gesture trajectory. Several instances in the negative-class samples that a human annotator would likely mislabel as positive without full video context were nonetheless correctly rejected by both models, indicating the models pick up on some cues that are not obvious to annotators working from static images alone, though this could also partly reflect the negative class’s status as the training majority.

	Static	Dynamic
Better overall performance	X	
Requires little manual labor		X
Mainly limited by class imbalance	X	
Mainly limited by form variation		X
Input features	hand	full body
Max. hands used as input	1	2
Output classes	2	3

Table 6: Comparison of the static and dynamic models.

7.3 Comparison of Approaches

Table 6 summarizes the trade-offs between the two approaches. The static model achieved better overall performance and required under a minute to train and evaluate, versus roughly ten hours for the dynamic model, but required substantially more manual labor, since gesture holds had to be located and screenshotted by hand, whereas the dynamic model could use automatically extracted clips directly. The static model’s single binary output is also simpler than the dynamic model’s two-stage motion-then-gesture-type classification, though the latter’s added complexity could in principle support classifying additional gesture types. Based on these results, the static model seems to be the more practical choice for naturalistic data (typically imbalanced and highly variable in form, as here) while the dynamic model may be better suited to more controlled, artificial recording settings. However, as is further discussed in Section 8, the observed results could also be due to a lack of training time and data for the dynamic model.

7.4 Connection to the Literature

That both models can recognize the PUOH gesture despite substantial within-class variation supports Cienki (2021)’s claim that its different forms share underlying features across the continuum. Consistent with this, both models’ most reliable feature appears to be hand shape (an external-system feature; Bressem et al., 2013), both also struggle most with lower-arm rotation and supination, a feature central to Boutet et al. (2018)’s internal, joint-based system, lending some support to Cienki’s argument that the internal system better captures what unifies these gesture forms. The results also speak to Lutgen et al. (2026)’s broader argument for treating variation as signal rather than noise in NLP data: multiple PUOH forms were found to co-occur naturally in Dutch conversation, gesture frequency varied substantially by participant (consistent with Chu et al., 2014), and, in the follow-up experiment,

choosing what to represent in the data (a fixed, balanced sample) outperformed simply maximizing data volume, directly supporting [Lutgen et al.](#)'s recommendation to base data-selection choices on desired representation rather than availability.

8 Conclusion

This paper investigated how gesture recognition models can recognize recurrent gestures in Dutch, focusing on the palm up open hand gesture. A small corpus of naturalistic Dutch dyadic conversation was collected and annotated, in which the PUOH gesture, previously documented in many spoken and signed languages but not, to this knowledge, in Dutch specifically, was found to occur frequently (567 instances) and in the range of forms described by [Cienki \(2021\)](#), addressing RQ1. Two recognition models, a static image-based model and a dynamic video-based model, were trained and compared across conditions varying in positive-class variation and class balance to address RQ2. The static model was found to perform better overall and to be less affected by increased form variation, and the dynamic model performed better only when positive-class variation was minimized; both models struggled most with distinguishing palm orientation from hand shape. These results indicate that automatic recognition of the PUOH gesture is possible, though a great deal of further work is needed to build higher-performing, more robust models, and that real-world gesture variation, an essential part of how individuals and groups use language to express themselves, should remain a central focus of that work.

Future Work. Both models could benefit from more training data, and from data recorded from multiple or varied camera angles, to determine how sensitive current results are to viewpoint. Comparing additional recognition architectures beyond the two considered here, and specifically studying how models handle palm orientation (e.g., by constructing data that isolates this variable), would help address the orientation-confusion errors observed in both models. For the static model specifically, incorporating arm and shoulder landmarks alongside hand landmarks, automating the currently manual hold-detection step, and controlling for per-participant sample size during training are promising directions. For the dynamic model, testing performance on artificially elicited (rather than naturalistic) PUOH productions would help isolate the

effect of positive-class variation from other confounds. More broadly, extending this approach to other recurrent gestures, and to other languages, could make quantitative analysis of recurrent gestures considerably more efficient.

Limitations

The corpus is small (20 speakers, one regional variety of Dutch, a narrow age and education range) and drawn from a single elicitation task, limiting how far these findings generalize. The static model uses only one hand's landmarks as input, which likely contributes to false negatives on gestures produced with the non-tracked hand or with both hands together. The dynamic model was trained for only 5 of the 500 epochs used in the architecture it was adapted from, owing to time and compute constraints; it was still improving when training was stopped, so its reported performance likely understates what the architecture could achieve with full training. Video resolution was in some cases too low to clearly resolve hand landmarks, which may have affected the static model's input quality. Finally, because gesture subclass boundaries lie on a continuum, subclass labels (and the Unsure category in particular) involved some annotator judgment, which may not generalize across annotators.

Ethics Statement

This study was reviewed and approved by the Research Ethics Review Committee and Research Integrity of the host institution prior to data collection. All participants provided informed consent, including separate consent regarding the potential future use of anonymized data excerpts in publications or presentations. Recordings and derived materials are stored on institutional secure cloud storage; only the subset of the corpus covered by participants' consent will be shared with other researchers upon request.

Acknowledgments

This research was funded by the Department of Language, Literature and Communication (LLC) at VU Amsterdam School of Humanities. The authors would like to extend special thanks to the lab managers of the Social Robotics lab, who provided the necessary equipment, as well as the students who participated in this study.

References

- Ahmad Sami Al-Shamayleh, Rodina Ahmad, Mohammad AM Abushariah, Khubaib Amjad Alam, and Nazeen Jomhari. 2018. A systematic literature review on vision based gesture recognition techniques. *Multimedia Tools and Applications*, 77(21):28121–28184.
- Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and social psychology bulletin*, 21(4):394–405.
- Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse processes*, 15(4):469–489.
- Dominique Boutet, Aliyah Morgenstern, and Alan Cienki. 2018. Looking ahead: Kinesiological analysis. *Aspectuality cross Languages Event construal in speech and gesture*, pages 143–159.
- Jana Bressem and Cornelia Müller. 2014. A repertoire of german recurrent gestures with pragmatic functions. *Body–language–communication: An international handbook on multimodality in human interaction*, 2:1558–1574.
- Jana Bressem et al. 2013. A linguistic perspective on the notation of form features in gestures.
- Feng-Sheng Chen, Chih-Ming Fu, and Chung-Lin Huang. 2003. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and vision computing*, 21(8):745–758.
- Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1):131–153.
- Mingyuan Chu, Antje Meyer, Lucy Foulkes, and Sotaro Kita. 2014. Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, 143(2):694.
- A Cienki, O Iriskhanova, A Agafonova, M Kiose, A Leonteva, A Petrov, O Prokofeva, and E Smirnova. Forthcoming. Recurrent gestures and verbal stance markers in russian: Viewing each in the light of the other. *Gesture*.
- Alan Cienki. 2021. From the finger lift to the palm-up open hand when presenting a point: A methodological exploration of forms and functions. *Languages and Modalities*, 1(1):17–30.
- Alan Cienki, Maíra Avelar, and Olga Iriskhanova. 2025. Action schemas can account for prototypical and non-prototypical palm-up open hand gestures. *Languages*, 10(11):282.
- Kensy Cooperrider, Natasha Abner, and Susan Goldin-Meadow. 2018. The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, 3:23.
- Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished language technology: The lack of (social) class in nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8675–8682.
- Trevor Darrell and Alex Pentland. 1993. Space-time gestures. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340. IEEE.
- Pierre Feyereisen and Jacques-Dominique De Lannoy. 1991. *Gestures and speech: Psychological investigations*. Cambridge University Press.
- Jana M Iverson and Susan Goldin-Meadow. 1998. Why people gesture when they speak. *Nature*, 396(6708):228–228.
- Jana M Iverson and Susan Goldin-Meadow. 2005. Gesture paves the way for language development. *Psychological science*, 16(5):367–371.
- Annelies Jehoul, Geert Brône, and Kurt Feyaerts. 2017. The shrug as marker of obviousness: corpus evidence from dutch face-to-face conversations. *Linguistics Vanguard*, 3(s1):20160082.
- Adam Kendon. 1995. Gestures as illocutionary and discourse structure markers in southern italian conversation. *Journal of pragmatics*, 23(3):247–279.
- Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
- Minwoo Kim, Jaechan Cho, Seongjoo Lee, and Yunho Jung. 2019. Imu sensor-based hand gesture recognition for human-machine interfaces. *Sensors*, 19(18):3827.
- Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167.
- Ruslan Kurdyumov, Phillip Ho, and Justin Ng. 2011. Sign language classification using webcam images. *Comput. Therm. Sci*, 10:9029.
- Silva H Ladewig. 2014. 118. Recurrent gestures. In *Volume 2*, pages 1558–1574. De Gruyter Mouton.
- James Loeb. 1912. The Loeb classical library.
- Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in nlp. *arXiv preprint arXiv:2603.24222*.
- Max Planck Institute for Psycholinguistics, The Language Archive. 2026. ELAN (version 7.1) [computer software]. <https://archive.mpi.nl/tla/elan>. Accessed: 2026-07-06.

- David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Joseph Albert Mosher. 1916. *The essentials of effective gesture, for students of public speaking*. Macmillan.
- Cornelia Müller. 2004. Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256.
- Cornelia Müller. 2010. Wie gesten bedeuten eine kognitiv-linguistische und sequenzanalytische perspektive. *Sprache und Literatur*, 41(1):37–68.
- Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessendorf. 2013. *Body-Language-Communication. Volume 1*. Walter de Gruyter.
- Yanghee Nam and KwangYun Wohn. 1996. Recognition of space-time hand-gestures using hidden markov model. In *Proceedings of the ACM symposium on virtual reality software and technology*, pages 51–58.
- Juan Jesus Ojeda-Castelo, Maria de Las Mercedes Capobianco-Uriarte, Jose Antonio Piedra-Fernandez, and Rosa Ayala. 2022. A survey on intelligent gesture recognition techniques. *IEEE Access*, 10:87135–87156.
- Edward Amherst Ott. 1902. *How to gesture*. Hinds & Noble.
- Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv preprint arXiv:1608.07836*.
- Wim Pouw. 2024. Envisionbox modules for social signal processing. https://github.com/WimPouw/envisionBOX_modulesWP. [Computer software]. Accessed: 2026-07-06.
- Olga Prokofyeva, Maria Kiose, Anna Leonteva, and Evgeniya Smirnova. 2025. Intensity and its manifestations in speech and recurrent gestures in spontaneous dialogue. In *Proceedings of the International Conference “Dialogue*, volume 2025.
- Alan Ramponi. 2024. Language varieties of italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.
- Jürgen Erich Schmidt. 2010. Language and space: The linguistic dynamics approach. *Language and space. An international handbook of linguistic variation*, 1:201–225.
- Keshav Sinha, Rashmi Kumari, Annu Priya, and Partha Paul. 2019. A computer vision-based gesture recognition using hidden markov model. In *Innovations in Soft Computing and Information Technology: Proceedings of ICEMIT 2017, Volume 3*, pages 55–67. Springer.
- William C Stokoe Jr. 1960/2005. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37.
- Jürgen Streeck. 2005. Pragmatic aspects of gesture. *International encyclopedia of languages and linguistics*, 5:71–76.
- Jürgen Streeck. 2009. *Gesturecraft*.
- Shigeyuki Tateno, Yiwei Zhu, and Fanxing Meng. 2019. Hand gesture recognition system for in-car device control based on infrared array sensor. In *2019 58th Annual conference of the society of instrument and control engineers of Japan (SICE)*, pages 701–706. IEEE.
- Sedinha Teßendorf. 2014. Pragmatic and metaphoric—combining functional with cognitive approaches in the analysis of the “brushing aside gesture”. *Body—language—communication: An international handbook on multimodality in human interaction*, 2:1540–1557.
- Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, Mohamed K Shahin, and Basma Refaat. 2015. Sift-based arabic sign language recognition system. In *Afro-European Conference for Industrial Advancement: Proceedings of the First International Afro-European Conference for Industrial Advancement AECIA 2014*, pages 359–370. Springer.
- Odlibek Tokhirov. 2023. How i built a hand gesture recognition model in python. [Link](#). Accessed: 06-07-2026.
- Pei Xu. 2017. A real-time hand gesture recognition and human-computer interaction system. *arXiv preprint arXiv:1704.07296*.
- Bosco Yung. 2022. Nodding pigeon. <https://github.com/bhky/nodding-pigeon>. [Computer software]. Accessed: 2026-07-06.
- Han Zou, Yuxun Zhou, Jianfei Yang, Hao Jiang, Lihua Xie, and Costas J Spanos. 2018. Wifi-enabled device-free gesture recognition for smart home automation. In *2018 IEEE 14th international conference on control and automation (ICCA)*, pages 476–481. IEEE.