Lang Resources & Evaluation
https://doi.org/10.1007/s10579-022-09586-4

ORIGINAL PAPER

Semi-automation of gesture annotation by machine learning and human collaboration

Naoto Ienaga[1] · **Alice Cravotta**[2] · **Kei Terayama**[3] · **Bryan W. Scotney**[4] · **Hideo Saito**[1] · **M. Grazia Busà**[2]

Accepted: 7 February 2022

Abstract Gesture and multimodal communication researchers typically annotate video data manually, even though this can be a very time-consuming task. In the present work, a method to detect gestures is proposed as a fundamental step towards a semi-automatic gesture annotation tool. The proposed method can be applied to RGB videos and requires annotations of part of a video as input. The technique deploys a pose estimation method and active learning. In the experiment, it is shown that if about 27% of the video is annotated, the remaining parts of the video can be annotated automatically with an F-score of at least 0.85. Users can run this tool with a small number of annotations first. If the predicted annotations for the remainder of the video are not satisfactory, users can add further annotations and run the tool again. The code has been released so that other researchers and practitioners can use the results of this research. This tool has been confirmed to work in conjunction with ELAN.

Keywords Gesture detection · Machine learning · Active learning · Video annotation

✉ Naoto Ienaga
naotoienaga@keio.jp

[1] Keio University, Yokohama, Japan

[2] University of Padova, Padova, Italy

[3] Yokohama City University, Yokohama, Japan

[4] Ulster University, Jordanstown, Northern Ireland

Published online: 25 February 2022

1 Introduction

In this section, we provide some general background information about gesture studies and the challenges we aim to address in this work (Sect. 1.1). After that, we briefly describe the machine learning techniques used (Sect. 1.2).

2 Gesture research

Gesture research investigates how gestures are integrated with speech and how they convey meaning in communication. Over the past decades scholars from different disciplines have studied, for example, the relationship between gestures and speech in terms of semantics, pragmatics, syntax, phonology, temporal alignment; the role of gestures in social interaction and human cognition; the development of gestures and language in children; the decay of gestures in language impairments; the creation of codified/shared gestural forms from spontaneous gestures; the relationship between gestures and signs; and the role of gestures in language origins and evolution. The interest in these topics has led to the flourishing of experimental studies that have had an impact on different fields such as cognitive science, psychology, psycholinguistics, cognitive linguistics, developmental psychology and linguistics, speech therapy, neuroscience, primatology, human communication, and computational multimodal research (Church et al., 2017). More specifically, gesture studies can have a wide range of applications, for example, in clinical settings with regard to the possibility of assessing speech disorders or language development impairments such as apraxia, Parkinson's disease, autism spectrum disorders, aphasia (among others, Goldenberg et al., 2003; Humphries et al., 2016; Özçalışkan and Goldin-Meadow, 2005; Özçalışkan et al., 2016).

In gesture research, qualitative observations have been widely used, but in the last decades, scholars have started to integrate more tools in their studies such as the annotation software ANVIL (Kipp, 2001) or ELAN (Wittenburg et al., 2006) that enable them to analyze their data quantitatively. For this purpose, technology such as motion capture (e.g., Kinect) can be employed as a first step in the analysis of visual data: this technology can be used to identify individual gestures and the kinematic features on which manual annotators can perform further classifications and analyses (Trujillo et al., 2019). The techniques that enable the temporal properties of gesture kinematics to be quantified in combination with speech can be both device-based motion capture techniques as well as video-based motion tracking methods (i.e., pose estimation methods) (Pouw et al., 2020). However, gesture annotation is generally done entirely manually through the annotation software. This means that, in most cases, gestures need to be detected by the human eye of a trained annotator who needs to manually mark their beginning and end points. Moreover, to ensure objectivity, the process often involves more than one annotator for a single dataset. This whole process is extremely time-consuming and labor-intensive. Although different researchers annotate gestures in different ways depending on their field, research purpose, and aims, in many cases, as a primary

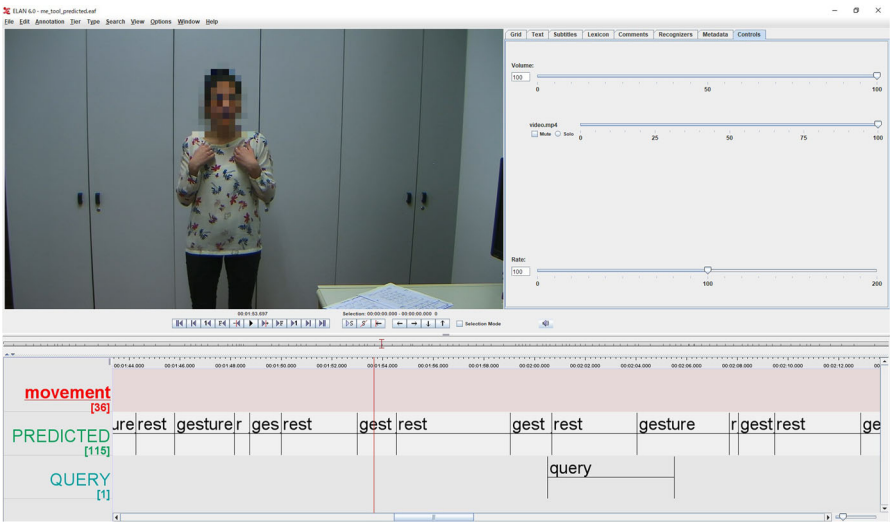

Fig. 1 Output file of the released tool as read by ELAN. The "movement" tier includes manual annotation; the "PREDICTED" tier shows predicted (automatic) annotation. The tool recommends the user to annotate "query" in the "QUERY" tier if the automatic annotation is not satisfactory. This way, the system will repeat predictions

step, it is important to assess if and when the gesture occurs. More finely grained types of analysis (e.g., alignment with speech) and gesture classification (e.g., gesture types, functions, etc.) can be conducted as a second step.

Developing a tool that provides researchers with semi-automatic gesture annotation on RGB videos could be highly advantageous in accelerating at least the primary steps of a wide range of studies and applications. This would significantly contribute to advance multimodal communication research in general, as well as to develop human–computer interaction applications. Also, a method that enables movement tracking and gesture detection on simple RGB videos as input could be applicable much more widely than using less accessible three-dimensional (3D) visual data or device-based motion tracking techniques. It is often the case that data of interest might have been collected for other purposes as simple RGB video material, or it might be useful to collect data from the internet, depending on the different research aims. Capitalizing on pose estimation methods currently available via open-source tools can be a good and accessible way to improve research that draws upon videos without any 3D information.

The present work proposes a method to detect gesture occurrences automatically based on RGB videos that can be of use for researchers studying multimodal communication. This method uses machine learning in the Active Learning (AL) framework. It requires as input the manual annotation of only a small subset of the videos, and it provides an automatic annotation of the remaining set. Figure 1 shows how the resulting automatic annotation appears when imported in ELAN. Another unique aspect of our research is that our target is *gesticulations*, which have not been studied widely in the fields of machine learning and computer vision compared

Table 1 Kendon's continuum (McNeill, 1992, 2005)

Gesticulation (co-speech gestures)	Pantomimes	Emblems (quotable gestures)	Sign languages
Speech present	Speech absent	Speech present or absent	Speech absent
No linguistic properties	No linguistic properties	Some linguistic properties	Linguistic properties
Not conventionalized	Not conventionalized	Partly conventionalized	Fully conventionalized

Table 2 The percentage of frames annotated as gesture (frames where gestures occurred) against the entire video frames (gesture ratio) and the length of the video for the six videos in the first dataset

	Gesture ratio (%)	Video length (min s)
Negation	57.82	4′24
Palm-up	41.92	4′26
Pointing	28.14	4′29
Me	38.12	4′26
Precision grip	60.93	5′25
Combination	45.85	8′31

with other gestures (see Sect. 2). In fact, the term "gesture" has been used in various ways in such fields, often as a synonym of hand movement, action, or sign. In the next section, we clarify what it is meant by "gesture" in the field of gesture research.

2.1 Gesture categorizations

Gestures have been defined and classified in many ways. In general, they can be intended as body movements produced in communication exchanges (Kendon, 2004). This broad definition can be further specified in more detailed categorizations. Kendon's continuum (McNeill, 1992, 2005) classifies gesture as in Table 1, depending on different dimensions. Moving from left to right in the continuum, the obligatory presence of speech decreases, and the stability of the meaning, standardization and linguistic properties of the hand movement increase. "Gesticulation" is the most pervasive type of gesture in that it is any motion produced spontaneously together with speech in everyday communication (Kendon, 2004). It mainly involves hand and arm movements, but it is not limited to these body parts (shoulders, head, face, legs can be part of it). Gesticulations can also be referred to as co-speech gestures, or often, for brevity, simply as gestures. Along the continuum, "pantomimes" are (sequences of) gestures produced without speech that can convey a whole narrative without being conventionalized or having any linguistic properties. Next, "emblems" (Kendon, 1992) are those conventionalized hand movements whose meaning is established and shared in a (linguistic) community (e.g., the thumb up gesture that often stands for "OK" in American English). What distinguishes emblems from gesticulations is that the former can be understood in the absence of speech and have standards of well-formedness not found in other types of gestures (McNeill, 2005). However, observations of how people gesture spontaneously whilst speaking have highlighted that gesticulations show some regularities in expressing a set of meanings through some specific hand shapes and movement patterns (Müller, 2017) (see five types of gestures in Sect. 3.1) but their form-meaning mapping is less rigid than in emblems. On the rightmost side of the continuum, there are "sign languages", in which elements such as handshape, movement, location, orientation, and non-manual elements are the

Fig. 2 An example of each gesture (stroke phases). From left to right: negation; palm-up; pointing; me; precision grip

main building blocks of the language phonological and morphosyntactic structures. Importantly, Kendon's continuum excludes movements such as self-adaptors (like scratching one's nose or touching one's hair). Gesticulations or co-speech gestures are the object of the present work.

2.2 Gesture temporal structure

It has been observed that gesticulations unfold by passing through a series of phases (Kendon, 1980, 2004; McNeill, 2005). These phases are organized around the *stroke* phase. The stroke is the "nucleus" of the gesture, it takes on the gesture's communicative role. It is the phase of the excursion in which the hand shape and movement dynamics are manifested with greatest clarity (Kendon, 1980). It is by observing the stroke phase that a gesture can be described and classified in terms of types (Fig. 2). Any prototypical gesture, as described by Kendon (1980), starts with a *preparation* phase. In the preparation phase, the hands start departing from a *rest position* to reach the stroke phase. The hands can then return to a rest position again (*retraction* phase). Together these phases constitute a gesture phrase. There might not be a retraction phase if the speaker moves directly from a stroke to a new stroke. When combined in sequence, different gesture phrases between two rest positions are defined as a single *Gesture unit* (G-Unit). Gestures can be annotated differently depending on the research purpose. However, this temporal structure is likely to be the basis of most of the studies in gesture research.

2.3 Light gradient boosting machine and active learning

As discussed in the next section, machine learning techniques have been frequently used for gesture detection and gesture recognition. The technique as the basis of the semi-automatic annotation method proposed here is Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017). LightGBM is a machine learning framework for gradient boosting based on the decision tree algorithm. In gradient boosting, a decision tree is trained using the error between the actual values and the estimates provided by the previous decision tree. There are several other methods of

gradient boosting, such as XGBoost (Chen & Guestrin, 2016) and CatBoost (Dorogush et al., 2018). LightGBM was chosen because it can be trained quickly and with less memory than many other machine learning methods (this is why LightGBM was named "Light"). In fact, since the computational resources of the user are unpredictable, it is preferable to use a light method. LightGBM reduces the calculation cost by the following contrivances: firstly, training the decision tree leaf-wise rather than level-wise; secondly, treating the continuous values as a histogram.

Amongst the many training methods available, semi-supervised learning is often used for gesture detection. In semi-supervised learning, manual annotations of part of a video are used as training data, and the rest of the video is then annotated automatically, based on a model developed by using the training data as inputs to a machine learning algorithm. To reduce the amount of manual annotation work, AL can be used as an approach that is expected to require fewer annotations than semi-supervised learning whilst adding informative data to the training data little by little. The cycle of AL is as follows: (i) training the model with a few manually created annotations (training data) and then estimating the remaining annotations; (ii) asking an "oracle" (an instrument that tells the correct answer, a human annotator in our case) for a true annotation of the "query" (frames that are most likely to contribute to performance improvement); (iii) adding the annotation of the query to the training data; (iv) training the model again (repeat steps (ii) to (iv) afterwards); (v) finishing the cycle when a satisfactory result in terms of classification accuracy is obtained. By using AL and a cycle of human–machine collaboration, our method reduces the amount of manual annotation work required.

The following sections are structured as follows: in Sect. 2, related works are briefly summarized; in Sect. 3, the datasets used to test our approach are described; in Sect. 4, our approach to automatically detect gestures is explained in detail; in Sect. 5, the approach is tested on different types of gestures to verify its accuracy and robustness; in Sect. 6, a brief tutorial for using the published code of the proposed method is provided; Sect. 7 summarizes the proposed method and the key results, with a discussion of some future directions.

3 Related works

For automatic detection and recognition, research has been focused much more on signs within given sign languages than on other types of gestures (Koller et al., 2015; Ong & Ranganath, 2005; Sagawa & Takeuchi, 2000). In fact, sign language research is likely to contribute directly to innovations that can be of use to deaf people. A sign language wiki to check sign language lexicon in multiple languages along with the movement of avatars (Efthimiou et al., 2012), and an interpreter system for deaf people in a specific situation (hotel reception) have been developed (López-Ludeña et al., 2014). Signs result from the combination of features (i.e., hand shape, orientation, location, and movement) that follow the linguistic rules (i.e., grammar, syntax, morphology) of a given sign language. Therefore, it is possible to build datasets with a set of predetermined signs that can be annotated (Forster et al., 2014; Neidle & Vogler, 2012; Neidle et al., 2012; Von Agris et al.,

2008). The construction of a dataset is essential for supervised learning and has allowed Convolutional Neural Networks (CNN) to recognize signs accurately (Camgoz et al., 2017; Cui et al., 2017; He et al., 2017).

The recognition of a set of predetermined hand shapes and movements designed specifically for human–computer interaction has also been actively researched (Jacob & Wachs, 2014; Park & Lee, 2011; Rautaray, 2012; Rautaray & Agrawal, 2015; Vardy et al., 1999). In this field, wrist-worn devices using sensors other than cameras have been developed (Fukui et al., 2011; Kim et al., 2012; Rekimoto, 2001). Some specific gestures have been developed for human–robot control (Park et al., 2005; Waldherr et al., 2000) or human–robot interaction (Droeschel et al., 2011).

ChaLearn Looking at People Dataset is a recent RGB-D video dataset (Wan et al., 2020). It is one of the largest and best-known datasets for gesture detection and recognition. The dataset contains 249 gestures, including some specific signals (e.g., helicopter/aviation signals, diving signals), pantomimes, and Italian emblems, performed by 21 people (Wan et al., 2016). ChaLearn dataset does not include spontaneous gesticulations. Previous works on this dataset showed that CNN was able to recognize the 249 gestures contained in the dataset (Camgoz et al., 2016; Pigou et al., 2017), or another machine learning method was able to recognize the Italian emblems (Chen & Koskela, 2013). As in the case of ChaLearn dataset, other existing datasets have focused on predetermined hand configurations and movements, such as emblem-like gestures or pantomimes (i.e., imitation of actions) (Negin et al., 2018) or as reviewed (Ruffieux et al., 2014).

As for previous studies that focused instead on gesticulations specifically, a Support Vector Machine (SVM) was used for gesture units/phases recognition in storytellers' speeches recorded via Kinect (Madeo et al., 2016). By using a logistic regression classifier with hand positions, orientations and velocities as inputs, gesture strokes were detected in spontaneous speech (Gebre et al., 2012). The possibility of using audio-visual features has been explored to improve gesture recognition. For example, the recognition of some specific gestures occurring in TV weather forecasts (i.e., pointing, area, and contour gestures) was improved by feeding a Hidden Markov Model (HMM) with both spoken keywords and gesture features (Sharma et al., 2000). An acoustic feature of speech, fundamental frequency (F0), improved gesture phase recognition with an HMM and Bayesian network (Kettebekov et al., 2005), and both F0 and intensity improved the recognition of different types of beat gestures that align with speech prosodically in an HMM-based formulation (Kettebekov, 2004). Spontaneous gestures produced in storytelling were used to predict a set of selected co-occurring words from spontaneous speech (Okada & Otsuka, 2017). The dataset used was the "Multimodal Storytelling Interaction Dataset" (Okada et al., 2013). In the dataset, participants were asked to narrate an animated cartoon story from memory to another participant. Motion data were acquired with an optical motion capture system. After manual annotation of the dataset, machine learning methods were used to detect gesture features (e.g., the total length of the gesture segments, gesture phases, etc.). Then, spoken words were classified by using an HMM and SVM trained on the gesture features. Alternatively, there are two possible approaches to

detect gestures without machine learning methods: (1) detecting gestures depending on whether the hand is moving (Ienaga et al., 2018; Ripperda et al., 2020; Schreer & Masneri, 2014); (2) thresholding the distance between the hand and the rest position (De Beugher et al., 2018; Peng et al., 2014). However, there are drawbacks to this approach, such as the inability to respond flexibly to various situations (for example, when there are multiple rest positions) and the requirement for many thresholds that need to be tuned.

Finally, some studies have used AL in the classification of hand gestures, actions and the detection of hands. Nine emblems (circle, come here, down, go away, point, stop, up, horizontal wave, and vertical wave) were classified by ensemble learning of a support vector regression, a multi-layer perceptron, and a polynomial classifier (Schumacher et al., 2012). Hand positions, velocity values, and trajectory curvatures in 3D space acquired by a multi-camera system were used as features. The study also examined how the classification accuracy changed with the amount of training data. The twenty kinds of daily human activities were classified by using an acceleration obtained from wearable sensors attached to the hip and wrist (Liu et al., 2010). The dataset included walking, sitting, working on a computer, standing, eating or drinking, running, bicycling, vacuuming, folding laundry and so on. AL was compared with a method of increasing the training data at random, and it was shown that AL was more accurate. AL and a boosting algorithm were used effectively for hand detection (Francke et al., 2007). A skin model to detect hands was generated from the pixel values of the detected face area. Hands were detected by cascade classifiers with the features of rectangular features and modified local binary pattern. Finally, four handshapes (fist, palm, pointing, five) were classified by the decision tree algorithm.

4 Datasets

Two different datasets are used to test the proposed method. The first dataset (Sect. 3.1) was built for this study in a controlled setting (a single speaker gesturing, gestures performed in a controlled manner, and a fixed camera angle facing the speaker). To further validate the method, we also tested it on a second dataset (Sect. 3.2) that was collected as part of a previous study (Cravotta et al. 2019). This dataset is better representative of typical data used in gesture studies, consisting of videos of different speakers speaking spontaneously during a storytelling task.

5 First dataset

To design and preliminarily test our method we built a dataset in a fairly controlled setting. We decided to have a single speaker speaking and gesturing in a controlled manner and to focus on a closed group of five possible gestures (or, rather, gesture families) appearing in the videos. We focused on the following five gestures: *negation*, *palm-up*, *pointing*, *me* and *precision grip*. These wide gesture categories were selected because they occur frequently in a variety of discourse types and

contexts (Müller, 2017), including public speeches (Streeck, 2008), and their use is observed in many cultures. Figure 2 shows examples of the five gestures. Gestures accompanying the wide semantic field of *negation* and negativity (Calbris 2003; Inbar & Shor, 2019; Kendon, 2004; Bressem & Müller, 2014) have been observed to be often performed with the palm held downwards or towards the interlocutor, moving laterally. Such gestures show a common pattern of lateral movements and are believed to derive from actions like sweeping or knocking aside unwanted objects (Bressem & Müller, 2014). The *palm-up* gesture (Cooperrider et al., 2018; Müller, 2004) is characterized by palms open upwards, and fingers extended more or less loosely. Its meaning includes many nuances depending on motion patterns, trajectories and other visual cues (shoulders, facial expressions, etc.). One possible communicative function is to express an obvious perspective on a topic/entity or to present/offer an idea for the interlocutor to share (presentational palm up, conduit gesture) (Chu et al., 2014); but it can also express uncertainty (epistemic palm up, lateral palm, palm revealing) (Kendon, 2004; Chu et al., 2014). The *pointing* gesture directs the recipient's attention to an object or a location in the space. This gesture can also refer metaphorically to a point in time or an absent object (Kita, 2003). What we refer to as *me* gesture appears as a special kind of pointing gesture that the speaker directs towards themself. It can consist of one or two hands over the heart [appearing also in absence of speech (Parzuchowski et al. 2014)] or a hand or index finger-pointing to oneself. It may be used when sharing one's beliefs and ideas, or when talking about something one really cares about. The *precision grip* (Lempert, 2011; Streeck, 2008) is generally performed with the tip of the index finger and the thumb touching one another and can have various hand configurations along these lines. It is claimed to be used to convey specificity or precision in everyday communication (Kendon, 2004), but it also highlights sharp argumentation, or, more generally, information structure (Lempert, 2011) (e.g., new information or focus). The dataset consists of six videos recorded at 25 fps. Five of them contained many occurrences of only one type of gesture (either the negation, palm-up, pointing, me, or precision grip gesture. Fig. 2), while the other video contained occurrences of all five gesture types (combination). Table 2 shows summary information about the videos.

The speaker appearing in the videos told improvised short stories in Italian freely inspired by a few comic strips. To make speech and gestures as natural as possible, the speaker associated gestures with the stories. For example, for the negation gesture, the speaker tells the story by turning all story events negative (e.g., "the cat did not climb up the tree"); for the me gesture, the speaker told a personal story about herself and cats. The gestures performed were to some extent controlled in terms of structure and dynamics but allowed a certain variability, coherent with the abovementioned well-established gesture phase descriptions and categorizations (Kendon, 2004): (i) the gestures start from a rest position and the hands return to the rest position at the end. The speaker's hands rest in different rest positions (e.g., hands clenched on the chest, lower to the hips, or hands hanging beside the legs); (ii) the gestures consist of only one stroke or multiple strokes; (iii) the gestures are performed with one or both hands; (iv) the amplitude, speed, and position of the

Fig. 3 Examples from the second dataset

gestures vary, and patterns of gesture movements vary even for the same type of gesture.

The detection of specific moments within a single gesture (i.e., gesture phases such as "preparation" or "retraction" preceding or following the strokes) was out of the scope of this work. Therefore, the gestures were annotated as movements occurring between two rest positions (that is, G-Units, see Sect. 1.1.2), without specific attention to where the strokes occurred or where the preparation or retraction phase started/ended. Also, when multiple strokes were performed within a single G-Unit, these were all of a single gesture type. For example, a single G-Unit annotated might consist of a preparation phase, three negation gesture strokes and a retraction phase. A linguist manually annotated the videos with ELAN. The annotation labels were *gesture* (gestures happening between two consecutive rest positions) and *rest* (the periods in which the hands and arms were still and held in a rest position).

6 Second dataset

As a second step, we decided to test our method on a second dataset. This dataset is better representative of a typical dataset used in gesture research, as it was built to investigate gesture from a linguistic perspective (Cravotta et al., 2019). From this dataset 195 videos were used, where 20 different speakers in total tell short stories based on comic strips. Example frames are shown in Figs. 3 and 8. All instances of gesture in the dataset were annotated to investigate the relationship between gestures and prosodic features of speech by Cravotta et al. (2019). Therefore, differently from the annotations made in the first dataset, where only the G-Units were marked, these annotations were more finely grained, and marked all gesture strokes singularly. Every single gesture stroke was also classified in terms of gesture types. In this study, the strokes were annotated as either *representational* (gestures that depict images of concrete (*iconic* gestures)—or abstract (*metaphoric* gestures) entities or actions via hand shape or manner of execution, e.g., trajectory, direction) or *non-representational* gestures (all other gestures, including pragmatic gestures and interactive gestures, gestures with speech parsing functions, and epistemic meaning etc.).

Table 3 Comparison of datasets. STD stands for standard deviation

	First dataset	Second dataset
Data	One single speaker (standing) Controlled speech & gesture	20 speakers (sitting down) Spontaneous speech
Annotation	(1) G-Units (all strokes within a G-Unit are of one single type)	(1) G-Units (2) Gesture strokes (different stroke types within a single G-Unit)
Gesture type categorization	Classification done at the G-Unit level (1) Pragmatic gestures and non-representational gestures • Negation • Palm-up • Precision grip (2)Deictic • Pointing • Me	Classification done at the stroke level (1) Pragmatic gestures and non-representational gestures (2) Representational gestures • Iconics • Metaphorics (abstract pointing, placing)
Number of videos	6	195
Mean (STD) gesture ratio (%)	45.46 (11.24)	63.87 (17.43)
Mean (STD) video length (min s)	5′17 (1′30)	0′27 (0′10)
Total length [min s]	31′41	87′32

Table 4 Number of occurrences of each gesture type

Gesture type	First dataset			Second dataset		
	Pragmatic	Deictic	Other type	Non-repr	Representational	
					Iconics	Metaphorics
Number of occurrences	85	69	4	1185	874	337

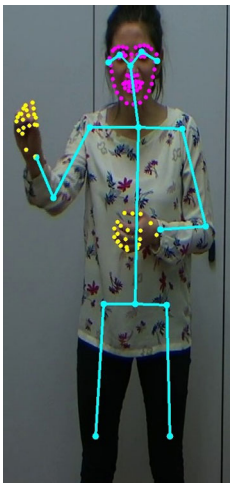

Fig. 4 An example of pose estimation by OpenPose (Cao et al., 2017). The keypoints of body, hands, and face are drawn in cyan, yellow, and magenta, respectively. Eyes, ears, and nose are also included in the body keypoints

7 Comparison of datasets

Table 3 shows the comparison of the datasets and how the datasets were annotated. There are some differences that should be noted. In the first dataset, gestures were annotated in terms of G-units only; in the second dataset gestures were annotated in terms of G-units as well as gesture strokes. In the first dataset there were no instances of representational gestures (iconics or metaphorics), that instead were pervasive in the second dataset, together with heterogeneous non-representational gestures (including many instances of palm up gestures).

Table 4 gives additional information about how the different gesture types are distributed in terms of occurrence in the two datasets. In the first dataset, four instances of hand movements occurred accidentally and were not classifiable within the five gesture types. These were annotated as "other type".

The following section describes the proposed automatic gesture annotation method and how it was tested using both the first and second datasets described above.

8 Methods

In the proposed method, the keypoints of the speaker's body are first detected in each frame of the video using a pose estimation method. An input feature vector is created for each frame based on the detected keypoints. LightGBM is then trained on the feature vectors to predict annotation.

8.1 Feature vector based on acquired keypoints

The keypoint positions of the speaker in each frame of the input video are detected by OpenPose (Cao et al., 2017). OpenPose is a CNN method for pose estimation for multiple people in real-time and is available as an open-source program. An example of pose estimation by OpenPose is shown in Fig. 4. OpenPose enables us to obtain the two-dimensional (2D) position of the body, face, and hands along with a confidence value (ranging from 0 to 1) in each frame. The keypoints for which the confidence value is less than 0.5 are interpolated linearly between frames. The number of keypoints used is 48: wrists (2); elbows (2); shoulders (2); hands (42). Each hand has 21 keypoints, denoting the wrist, knuckles, finger and thumb joints, and finger and thumb tips. In order to reduce the influence of the overall movement of the speaker and to accommodate changes in the distance between the camera and the speaker, keypoints are normalized in each frame by subtracting the neck position and dividing by the distance between the right and left shoulders after the interpolation.

For each frame the feature vector for LightGBM consists of position features and distance features. The position features are the 96-dimensional (2D positions of 48 keypoints) averaged keypoint positions over a window of temporally consecutive frames. The window size w is chosen to be an odd number. The distance features are the averaged 2D Euclidean distances between the keypoints in the center frame of the w frames and the corresponding keypoints in the neighbouring frames. Hence, the dimension of the feature vector is 144 (96 + 48).

9 Training LightGBM by AL

In this subsection, the way to train LightGBM by AL, based on the feature vector, is described in detail. It is assumed that part of the input video has been annotated. The annotated frames are divided into training data and validation data. When training stops, the annotations for the remaining part (that has not been annotated) of the video are predicted and the query (the next candidate frames to be annotated) are determined. To select the query, uncertainty sampling is adopted in the proposed method. Uncertainty sampling is a method of selecting the most uncertain data. The idea is that the accuracy can be improved most effectively if the most uncertain data are annotated (that is, the more uncertain, the more informative). The uncertainty is calculated by the predicted probability p. When an annotator annotates a video, it is very difficult to annotate only one frame. Hence a continuous sequence of frames

from the unannotated parts should be selected as the query. The total uncertainty u of a continuous sequence of n frames is calculated as follows:

$$u = \sum_{i=0}^{n-1} {p_i}'$$

$${p_i}' = \begin{cases} 1 - p_i (p_i \geq 0.5) \\ p_i (otherwise) \end{cases}$$

The consecutive frames with the highest u are selected as the query.

10 Experimental results and discussion

Three results are presented and discussed. The result of tenfold cross-validation (CV) (AL is not used) is described in order to verify the gesture detection capability of LightGBM. The method of increasing the training data by AL and the method of adding frames from the beginning to the training data are compared. Furthermore, we investigated whether the proposed method could be applied to gesture type recognition in both datasets tested.

The evaluation indices, accuracy, F-score (macro), and recall (G) used in the experiments are explained. Accuracy is the most intuitive evaluation index: it is the proportion of video frames overall that are classified correctly. In the case of two-class classification in Table 5, accuracy is formulated as (TG + TR)/(TG + FG + FR + TF), where T and F denote "True" and "False", respectively, and G and R denote "G-Unit" and "Rest", respectively. If the number of frames annotated as gesture and rest are not approximately balanced, then accuracy may not be an appropriate index. F-score is calculated from precision and recall. Precision (for each predicted class) is the proportion of frames predicted to be from the class that are actually from the class. Formulas (1) and (2) show how to calculate precision for the gesture and rest classes, respectively. Recall (for each actual class) is the proportion of frames actually from the class that are predicted to be from the class (formulas (3) and (4) for gesture and rest, respectively). F-score is a harmonic mean of precision and recall (formulas (5) and (6) for gesture and rest, respectively). F-score (macro) is a mean value of F-score (G) and F-score (R). F-score (macro) can be calculated as a mean value of F-score of each class even in the case of multiclass

Table 5 Confusion matrix of gesture vs rest. T and F denote "True" and "False", respectively, and G and R denote "G-Unit" and "Rest", respectively

		Predicted	
		Gesture	Rest
Actual	Gesture	TG	FG
	Rest	FR	TR

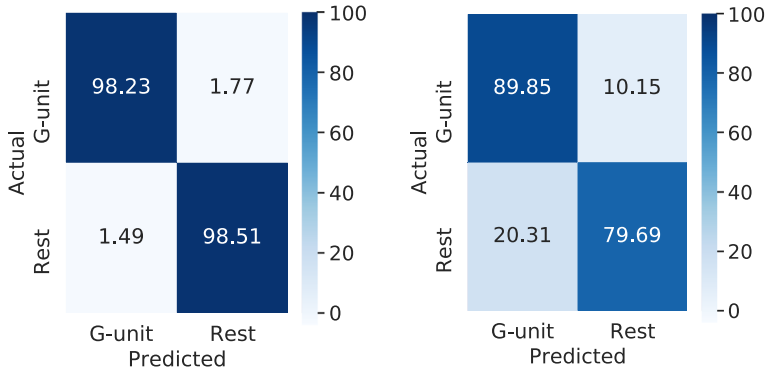

Fig. 5 Confusion matrix of gesture (G-Unit) detection (left: first dataset, right: second dataset). The values shown are for recall

classification (See Sect. 5.3). F-score (macro) better evaluates the performance of the proposed method than accuracy when the numbers of annotations of gesture and rest differ significantly. The reason that we focus on recall (G) is that if recall (G) is high, then the actual gestures are mainly annotated correctly, and so the user has only to check the annotated area, reducing the effort of checking the entire video.

$$\text{Precision (G)} = \frac{\text{TG}}{\text{TG} + \text{FR}} \tag{1}$$

$$\text{Precision (R)} = \frac{\text{TR}}{\text{TR} + \text{FG}} \tag{2}$$

$$\text{Recall (G)} = \frac{\text{TG}}{\text{TG} + \text{FG}} \tag{3}$$

$$\text{Recall (R)} = \frac{\text{TR}}{\text{TR} + \text{FR}} \tag{4}$$

$$\text{F} - \text{score (G)} = \frac{2 \cdot \text{Precision(G)} \cdot \text{Recall(G)}}{\text{Precision(G)} + \text{Recall(G)}} \tag{5}$$

$$\text{F} - \text{score (R)} = \frac{2 \cdot \text{Precision(R)} \cdot \text{Recall(R)}}{\text{Precision(R)} + \text{Recall(R)}} \tag{6}$$

10.1 Result of tenfold CV

Before considering the results obtained by using AL, we consider how accurately gesture detection is possible with LightGBM. The parameters of LightGBM were determined by a two-phase grid search. After a rough grid search over a wide area, a fine grid search was done around the best parameters of the first grid search. The

Table 6 Mean (standard deviation) of accuracy/F-score/recall of gesture detection for each dataset and method

Dataset/method	First/proposed	First/simple	Second/proposed	Second/simple
Accuracy	0.984 (0.001)	0.817 (0.050)	0.862 (0.082)	0.769 (0.112)
F-score (macro)	0.984 (0.001)	0.808 (0.054)	0.814 (0.103)	0.707 (0.123)
Recall (G)	0.982 (0.002)	0.805 (0.119)	0.871 (0.171)	0.788 (0.201)

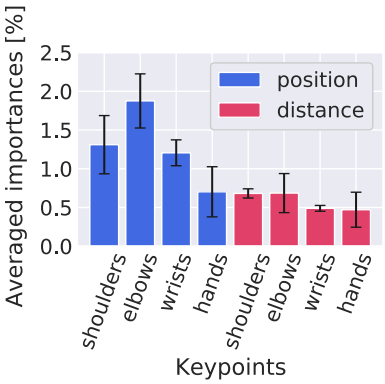

Fig. 6 Averaged importance for each keypoint

optimal parameters that give the highest F-score was searched in 225 combinations of four parameters. In the grid search, the first and second datasets were combined. In all experiments, the learning rate was 0.1, w was 13, and early stopping was used. When training, weights were assigned according to the number of annotations of each class.

Each of the first and second datasets was tested separately, using the optimal parameters found by the grid search. For the second dataset, tenfold CV was carried out with 10% of the videos as test data, 10% of the videos as validation data, and 80% of the videos as training data. For the first dataset, the frames of all videos were shuffled and divided into test (10%), validation (10%), and training (80%) data since each video contains only one type of gesture (other than the combination). Figure 5 is confusion matrices showing the result of tenfold CV. They represent recall of each class with a total of 10 folds. Table 6 shows the mean and standard deviation of 10 folds of the evaluation indices. It can be seen that G-Units were detected with very high accuracy in the first dataset, and with an accuracy of more than 86% for the second dataset, which includes more natural scenes.

To verify that the proposed method is superior to a purely movement-based approach, the proposed method was compared with a simpler method. In this simpler method, only the distance features described in Sect. 4.1 are used. The mean value of the 48-dimensional distance feature is thresholded (that is, gesture detection was based only on whether the mean value exceeded the threshold). The optimal threshold was calculated from each dataset. Table 6 shows the results of

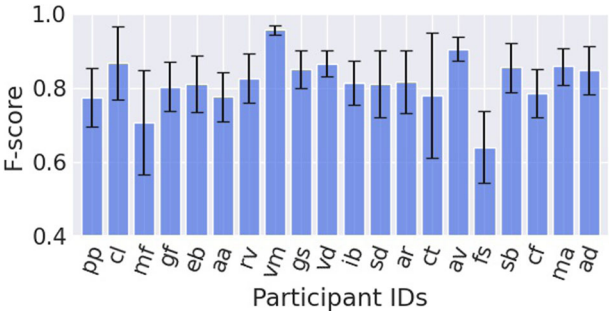

Fig. 7 Mean and standard deviation of F-scores for each participant in the second dataset

Fig. 8 Examples of non-gesture frames of participants with particularly high ("vm", left side) and low ("fs", right side) F-scores

comparing the proposed method and the simpler method for both the first and second datasets. In any case, the proposed method is better, in both accuracy and F-score, by approximately 10%.

LightGBM can also calculate the importance of each dimension of the input feature. The higher the importance, the more the feature contributed to gesture detection. 20 sets of importance values (10 folds for the two datasets) were averaged, and Fig. 6 shows the averaged importance for each type of keypoint (i.e., the averaged importance values further averaged over each type of keypoint). In particular, we see that the features related to the position of shoulders, elbows and wrists contributed most significantly to the accuracy of the gesture (G-Unit) detection.

Figure 7 is a bar plot showing the averaged F-score for each participant in the second dataset. Overall, the F-scores are similar across the participants, but the F-score for participant "vm" was particularly high, with low variance, and for participant "fs" was particularly low, with high variance. These differences in the F-scores can be attributed to the observations that "vm" almost always had her hands in the same rest position, and her hands hardly moved in rest, whereas "fs"

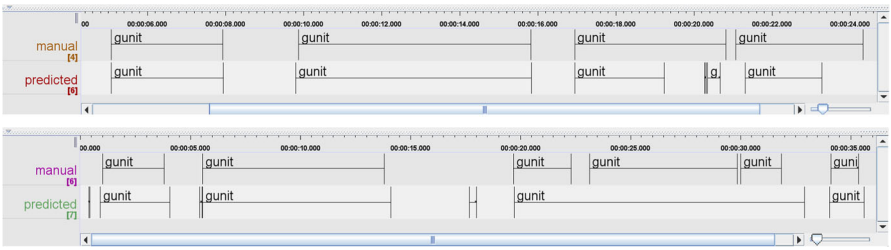

Fig. 9 Examples of predicted annotations. The top is a video of "vm", and the bottom is a video of "av". The LGBM model trained in each of the other three videos predicted. The "manual" tier is a manual annotation and the "predicted" tier is a predicted annotation

often moved her hand even in rest (Fig. 8). This suggests the vulnerability of the proposed method to movements other than gestures (e.g., self-adaptors (scratching one's leg)).

Furthermore, Fig. 9 shows the results of importing the predicted annotations with ELAN. These are the annotations for a video of two participants ("vm" and "av"), predicted by the LGBM model trained on the other three videos for each participant. Some predictions were almost consistent with manual annotations (the first half on the top). On the other hand, there was a tendency for a few seconds' gap between manual and predicted annotations in the start and the end of G-Units, and holds in G-Units may be predicted as rest (not G-Unit) (the second half on the top). Short G-Units were erroneously predicted in some places where there were no G-Units, and multiple G-Units that occurred consecutively at short intervals were predicted as one G-Unit (bottom). A more desirable result could be obtained by disabling short annotations (Sect. 6). As a whole, the proposed method could roughly predict G-Unit.

10.2 Results of AL

First of all, all videos were concatenated for each dataset, and then the video was sectioned into units of equal length, with each unit containing 0.5% of the total number of frames in the dataset. Two strategies were adopted, and the results are compared here. In the first strategy, AL, one of the units was selected as the query from the test data at each cycle by the method based on the measure of uncertainty as described in Sect. 4.2. The training/validation data would thus increase by 0.5%. In the second strategy, denoted *outset*, units from the test data were simply added to the training/validation data one by one from the beginning. Both of the initial training/validation data was the first unit. If either rest or G-Unit was included in the first unit, the next unit was added to the initial training/validation data. We continued to add units until both rest and G-Unit were included in the initial training/validation data. The F-score and the computational times required to train the LightGBM and to estimate the test data annotation for the AL method were calculated at each cycle.

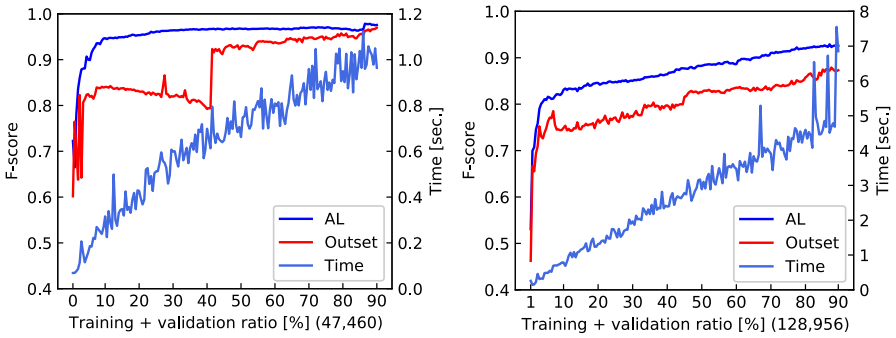

Fig. 10 F-scores for annotation prediction for both AL and *outset* strategies for first dataset (left) and second dataset (right). The graphs show how the F-score and the computational time required for training/prediction vary as the percentage of data used for training/validation increases. The total number of frames used is shown in parentheses for each dataset

Figure 10 shows the results for the AL and outset strategies. In most cases the F-score for AL is above that for outset, demonstrating that annotation according to AL is more efficient than simple annotation based on units of data from the beginning of the video. For example, for the first dataset, if approximately 2% of the video were annotated, the remaining parts of the video were annotated automatically with an F-score of approximately 0.84; increasing the amount of annotation to 4.5% would yield an improved F-score of 0.91. For the second dataset, it was possible to predict annotations with F-score 0.80 using 3.5% annotation; to increase the F-score to 0.85 required approximately 27% annotation. Even if the predicted annotations were to be corrected afterwards manually, it can be expected that annotation with AL would require much less effort than fully manual annotation. Depending on the amount of training data and computational resources, not more than 10 s were required for training and prediction in any case with the parameters used this time (the CPU used this time was Intel Xeon CPU E5-2690 v3 2.60 GHz 12 cores. There were 24 CPUs, though we did not use any parameters related to the number of threads or parallelization. Also, no GPU was used).

10.3 Gesture type classification

To further test our method, we conducted an experiment to verify whether the proposed framework can be applied to gesture type classification. Parameter tuning, and training and testing methods are the same as the experiment in Sect. 5.1. However, the parameter tuning was done separately for the first and second datasets, because the gesture types included in them were different (see Table 3). In the first dataset: (1) Pragmatic gesture (includes negation, palm up and precision); (2) Deictic (includes pointing and "me" gestures). In the second dataset: (1) Non-representational (includes a broad variety of pragmatic gestures); (2) Iconics and (3) Metaphorics (includes a variety of gestures depicting images of concrete or abstract entities or actions via the shape of the gesture, manner of movement).

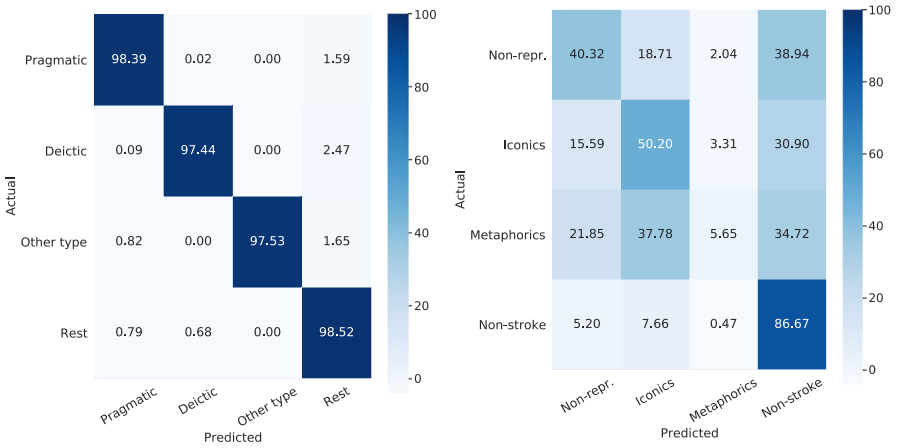

Fig. 11 Confusion matrix of gesture type classification (top: first dataset, bottom: second dataset). The values shown are for recall

Table 7 Mean (standard deviation) of accuracy and F-score of gesture type classification for each dataset

Dataset	First	Second
Accuracy	0.983 (0.002)	0.650 (0.115)
F-score	0.983 (0.006)	0.417 (0.109)

Figure 11 presents confusion matrices and Table 7 shows the results of gesture type classification. For a simple dataset (where the difference between gesture and rest and the differences in movement for each gesture type are clear), such as the first dataset, it is clear that the proposed method can classify gesture types with high accuracy. However, the results for the second dataset suggest that the proposed method, in its current form, cannot accurately classify gesture types when applied to more complicated datasets. In the right-hand confusion matrix in Fig. 11, "Non-stroke" denotes frames other than the frames annotated as stroke. As explained in Table 3, gesture types in the second dataset were classified at the stroke level, unlike in the first dataset, where only G-Units were considered. The non-stroke frames in the second dataset accounted for more than 52% of the total. In addition, preparation, retraction, and non-gesture movements (Fig. 8 right side) were included in the non-stroke frames, so it is conceivable that the model often confused parts of strokes with the preparation and retraction phases of the gesture, leading to a high proportion of non-stroke frames being predicted.

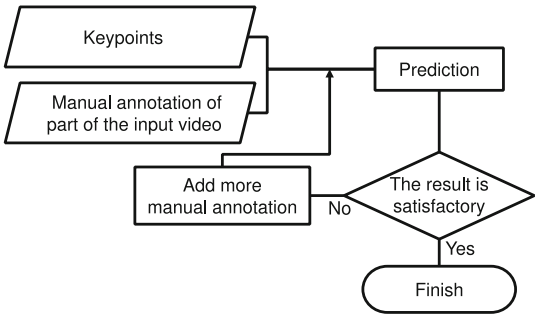

Fig. 12 Flowchart of the use of the released tool

11 Annotation tool: brief tutorial

The code has been released so that anyone can use the proposed method (https://github.com/naotoienaga/annotation-tool). With this tool, an ELAN file with predicted annotations is automatically output as shown in Fig. 1. This tool has been confirmed to work with the latest version of ELAN 6.0 (as of February 21, 2020). Users can use this tool with the following steps (refer to Fig. 12 also):

1. Run OpenPose to detect keypoints of the speaker, and annotate part of the input video. The annotation should include at least one rest and one gesture (any tier name is acceptable other than PREDICTED and QUERY). Specify paths to the zip file of json files generated by OpenPose and the ELAN file in the code (also some parameters such as w can be changed as an advanced setting).
2. Run the code.
3. An ELAN file with predicted annotations will be generated, like Fig. 1. If the result is not satisfactory, add more annotations and run the code again (delete PREDICTED tier and QUERY tier).

Please refer to the URL above for further details and the implementation. In Fig. 1, the manual annotations are included in the "movement" tier. The annotations in the "PREDICTED" tier are predicted (the manual annotations will be copied to "PREDICTED" tier). The annotation in the "QUERY" tier is query selected by uncertainty sampling (as described in Sect. 4.2).

This tool can also modify short annotations. Annotations with a length within the number of seconds specified by the user are modified taking into account surrounding annotations. At the top of Fig. 13, it can be seen that the annotation switches at very short intervals, especially at the boundaries of the annotations. The short annotations were modified as shown at the bottom of Fig. 13. This is a function to reduce the work of manual modification of the predicted annotation rather than to improve accuracy.

Fig. 13 Results of disabling short annotation modification (top) and of modification of annotations within 0.5 s (bottom)

12 Conclusion

In this paper, we have proposed a method to semi-automatically annotate gestures in any type of RGB videos without requiring use of 3D videos or device-based motion capture techniques. The proposed system uses a state-of-the-art pose estimation method which detects body keypoints that are used as features to train the machine learning model. The method is based on machine learning and human collaboration; active learning requires the manual annotation of a small subset of the data in order to semi-automatically annotate gestures of the remaining subset to be annotated. Also, if users are not satisfied with the accuracy of the automatic annotation obtained, they can decide to increase the amount of manually annotated data following the instruction of AL to improve accuracy efficiently. Since the program is publicly available, anyone who is interested can try our gesture annotation system on their own data. The program can run on the service that can execute Python from a web browser without the need to build an environment, so the program can be easily used. The output of the annotation tool can be read in ELAN for double-checking, accuracy improvement, further analyses and making its use more accessible to researchers across different fields.

The proposed method was tested on two datasets and performed gesture detection with higher accuracy in both datasets than a method that simply thresholds the distance moved by the keypoints. The first dataset contained a set of gesticulations produced in a controlled setting. For the first dataset, both gesture (G-Unit) detection and gesture type classification achieved an accuracy higher than 98%. The second dataset contained 20 speakers, who gestured spontaneously. For this dataset, the proposed method achieved 86% accuracy in gesture (G-Unit) detection but was less accurate in gesture type classification.

As for gesture type classification, in the second dataset, we believe our system was challenged for different reasons: speakers gestured in a spontaneous manner, and variability was not controlled in any way. This led to an infinite variability in terms of gestures performed; also, the number of non-meaningful (i.e., non-gesture movements) is certainly higher in the second dataset compared to the first dataset, in fact, as mentioned, the accuracy was lower for videos that included many non-gesture movements. Also, it should be noticed that researchers evaluate gesture types when having access to the accompanying speech as well, which can be a fundamental cue to distinguish, for example, iconic gestures from metaphorics. For these main reasons, we expected a decrease in the accuracy of gesture type classification in the second dataset however, we believe that these results provide a more realistic idea of how well our method performs on experimental (i.e., highly

controlled) videos compared to corpus data. The results of gesture type classification suggest that an improvement is needed in more challenging/naturalistic datasets and further testing is a fundamental step to achieve improvement. In principle, we believe that our method can be tested with any dataset. Gesture detection/recognition can be done in videos of multiple people co-present in the scene or other speaker types (e.g., children, infants, apes) as long as 2D body keypoints can be extracted from the image. A possible evolution of our method could be to use 3D information about the speaker, when available.

Despite some current limitations, we believe that our method can be a fundamental step towards providing a semi-automatic gesture annotation tool. This tool would significantly reduce the time-consuming work of annotators, by assisting them in the first steps of the gesture annotation process. Also, we believe that our method can boost significantly the capacity for gesture research in that it does not require motion capture techniques (e.g., Kinect) that might not always be available; this means that our method can be used on any videos that one wishes to annotate. To conclude, our method contributes to making gesture annotation faster, easier and more accessible. In the long term, this could increase the amount of annotated gesture data (contributing to building larger video corpora) and would benefit the study of multimodal communication and the development of human–computer interaction applications.

Funding Funding was provided by Japan Society for the Promotion of Science (Grant No. 17J05489).

Data availability Not applicable.

Code availability https://github.com/naotoienaga/annotation-tool.

Declarations

Conflict of interest There are no conflicts of interest to declare.

References

Bressem, J., & Müller, C. (2014). The family of away gestures: Negation, refusal, and negative assessment. *Body–language–communication: An International Handbook on Multimodality in Human Interaction*, *2*, 1592–1604. https://doi.org/10.1515/9783110302028.1592

Calbris, G. (2003). From cutting an object to a clear cut analysis: Gesture as the representation of a preconceptual schema linking concrete actions to abstract notions. *Gesture, 3*(1), 19–46. https://doi.org/10.1075/gest.3.1.03cal

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2016). Using convolutional 3d neural networks for user-independent continuous gesture recognition. *In 2016 23rd international conference on pattern recognition*, pp. 49–54. https://doi.org/10.1109/ICPR.2016.7899606

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE international conference on computer vision*, pp. 3075–3084. https://doi.org/10.1109/ICCV.2017.332

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, X. & Koskela, M. (2013). Online RGB-D gesture recognition with extreme learning machines. In *Proceedings of the 15th ACM on international conference on multimodal interaction*, 467–474. https://doi.org/10.1145/2522848.2532591

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (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. https://doi.org/10.1037/a0033861

Church, R. B., Alibali, M. W., & Kelly, S. D. (2017). *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins Publishing Company.

Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication, 3*, 23. https://doi.org/10.3389/fcomm.2018.00023

Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of encouraging the use of gestures on speech. *Journal of Speech, Language, and Hearing Research, 62*(9), 3204–3219. https://doi.org/10.1044/2019_JSLHR-S-18-0493

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2017.175

De Beugher, S., Brône, G., & Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation, 52*(2), 433–460. https://doi.org/10.1007/s10579-017-9404-9

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Droeschel, D., Stückler, J., Holz, D., & Behnke, S. (2011). Towards joint attention for a domestic service robot-person awareness and gesture recognition using time-of-flight cameras. In *2011 IEEE international conference on robotics and automation* (pp. 1205–1210). https://doi.org/10.1109/ICRA.2011.5980067

Efthimiou, E., Fotinea, S. E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., & Lefebvre-Albaret, F. (2012). The dicta-sign wiki: Enabling web communication for the deaf. *International Conference on Computers for Handicapped Persons*. https://doi.org/10.1007/978-3-642-31534-3_32

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., & Ney, H. (2014). Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *International conference on language resources and evaluation* (pp. 1911–1916).

Francke, H., Ruiz-del-Solar, J., & Verschae, R. (2007). Real-time hand gesture detection and recognition using boosted classifiers and active learning. *Pacific-Rim Symposium on Image and Video Technology*. https://doi.org/10.1007/978-3-540-77129-6_47

Fukui, R., Watanabe, M., Gyota, T., Shimosaka, M., & Sato, T. (2011). Hand shape classification with a wrist contour sensor: Development of a prototype device. In *Proceedings of the 13th international conference on ubiquitous computing* (pp. 311–314). https://doi.org/10.1145/2030112.2030154

Gebre, B. G., Wittenburg, P., & Lenkiewicz, P. (2012). Towards automatic gesture stroke detection. In *LREC 2012: 8th international conference on language resources and evaluation* (pp. 231–235). http://hdl.handle.net/11858/00-001M-0000-000F-8479-7

Goldenberg, G., Hartmann, K., & Schlott, I. (2003). Defective pantomime of object use in left brain damage: Apraxia or asymbolia? *Neuropsychologia, 41*(12), 1565–1573. https://doi.org/10.1016/S0028-3932(03)00120-9

He, T., Mao, H., & Yi, Z. (2017). Moving object recognition using multi-view three-dimensional convolutional neural networks. *Neural Computing and Applications, 28*(12), 3827–3835. https://doi.org/10.1007/s00521-016-2277-9

Humphries, S., Holler, J., Crawford, T. J., Herrera, E., & Poliakoff, E. (2016). A third-person perspective on co-speech action gestures in Parkinson's disease. *Cortex, 78*, 44–54. https://doi.org/10.1016/j.cortex.2016.02.009

Ienaga, N., Scotney, B. W., Saito, H., Cravotta, A., & Busà, M. G. (2018). Natural gesture extraction based on hand trajectory. In *Irish machine vision and image processing conference* (pp. 81–88).

Inbar, A., & Shor, L. (2019). Covert negation in Israeli Hebrew: Evidence from co-speech gestures. *Journal of Pragmatics, 143*, 85–95. https://doi.org/10.1016/j.pragma.2019.02.011

Jacob, M. G., & Wachs, J. P. (2014). Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters, 36*, 196–203. https://doi.org/10.1016/j.patrec.2013.05.024

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication, 207–227*,. https://doi.org/10.1515/9783110813098.207

Kendon, A. (1992). Some recent work from Italy on quotable gestures (emblems). *Journal of Linguistic Anthropology, 2*(1), 92–108. https://doi.org/10.1525/jlin.1992.2.1.92

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kettebekov, S. (2004). Exploiting prosodic structuring of coverbal gesticulation. In *Proceedings of the 6th international conference on multimodal interfaces* (pp. 105–112). https://doi.org/10.1145/1027933.1027953

Kettebekov, S., Yeasin, M., & Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE Transactions on Multimedia, 7*(2), 234–242. https://doi.org/10.1109/TMM.2004.840590

Kim, D., Hilliges, O., Izadi, S., Butler, A. D., Chen, J., Oikonomidis, I., & Olivier, P. (2012). Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on user interface software and technology* (pp. 167–176). https://doi.org/10.1145/2380116.2380139

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In: *Seventh European conference on speech communication and technology*.

Kita, S. (2003). *Pointing: Where language, culture, and cognition meet*. Psychology Press.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding, 141*, 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

Lempert, M. (2011). Barack Obama, being sharp: Indexical order in the pragmatics of precision-grip gesture. *Gesture, 11*(3), 241–270. https://doi.org/10.1075/gest.11.3.01lem

Liu, R., Chen, T., & Huang, L. (2010). Research on human activity recognition based on active learning. In *2010 international conference on machine learning and cybernetics* (pp. 285–290). https://doi.org/10.1109/ICMLC.2010.5581050

López-Ludeña, V., González-Morcillo, C., López, J. C., Ferreiro, E., Ferreiros, J., & San-Segundo, R. (2014). Methodology for developing an advanced communications system for the Deaf in a new domain. *Knowledge-Based Systems, 56*, 240–252. https://doi.org/10.1016/j.knosys.2013.11.017

Madeo, R. C. B., Peres, S. M., & de Moraes Lima, C. A. (2016). Gesture phase segmentation using support vector machines. *Expert Systems with Applications, 56*, 100–115. https://doi.org/10.1016/j.eswa.2016.02.021

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2005). Gesture and thought. *University of Chicago Press*. https://doi.org/10.7208/chicago/9780226514642.001.0001

Müller, C. (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.

Müller, C. (2017). How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. *Gesture, 16*(2), 277–304. https://doi.org/10.1075/gest.16.2.05mul

Negin, F., Rodriguez, P., Koperski, M., Kerboua, A., Gonzàlez, J., Bourgeois, J., & Bremond, F. (2018). PRAXIS: Towards automatic cognitive assessment using gesture recognition. *Expert Systems with Applications, 106*, 21–35. https://doi.org/10.1016/j.eswa.2018.03.063

Neidle, C., Thangali, A., & Sclaroff, S. (2012). Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th workshop on the representation and processing of sign languages: Interactions between corpus and lexicon, language resources and evaluation conference*.

Neidle, C. & Vogler, C. (2012). A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In *Proceedings of 5th workshop on the representation and processing of sign languages: Interactions between corpus and lexicon, language resources and evaluation conference*.

Okada, S., Bono, M., Takanashi, K., Sumi, Y., & Nitta, K. (2013). Context-based conversational hand gesture classification in narrative interaction. In *Proceedings of the 15th ACM on international conference on multimodal interaction* (pp. 303–310). https://doi.org/10.1145/2522848.2522898

Okada, S. & Otsuka, K. (2017). Recognizing words from gestures: Discovering gesture descriptors associated with spoken utterances. In *2017 12th IEEE international conference on automatic face & gesture recognition* (pp. 430–437). https://doi.org/10.1109/FG.2017.60.

Ong, S. C., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Computer Architecture Letters, 27*(06), 873–891.

Özçalışkan, Ş, Adamson, L. B., & Dimitrova, N. (2016). Early deictic but not other gestures predict later vocabulary in both typical development and autism. *Autism, 20*(6), 754–763. https://doi.org/10.1177/1362361315605921

Özçalışkan, Ş, & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition, 96*(3), B101–B113. https://doi.org/10.1016/j.cognition.2005.01.001

Park, H. S., Kim, E. Y., Jang, S. S., Park, S. H., Park, M. H., & Kim, H. J. (2005). HMM-based gesture recognition for robot control. *In Iberian Conference on Pattern Recognition and Image Analysis, 607–614*,. https://doi.org/10.1007/11492429_73

Park, S. Y., & Lee, E. J. (2011). Hand gesture recognition using optical flow field segmentation and boundary complexity comparison based on hidden Markov models. *Journal of Korea Multimedia Society, 14*(4), 504–516. https://doi.org/10.9717/KMMS.2011.14.4.504

Parzuchowski, M., Szymkow, A., Baryla, W., & Wojciszke, B. (2014). From the heart: Hand over heart as an embodiment of honesty. *Cognitive Processing, 15*, 237–244. https://doi.org/10.1007/s10339-014-0606-4

Peng, X., Wang, L., Cai, Z., & Qiao, Y. (2014). Action and gesture temporal spotting with super vector representation. *In European Conference on Computer Vision, 518–527*,. https://doi.org/10.1007/978-3-319-16178-5_36

Pigou, L., Van Herreweghe, M., & Dambre, J. (2017). Gesture and sign language recognition with temporal residual networks. *In Proceedings of the IEEE International Conference on Computer Vision Workshops, 3086–3093*,. https://doi.org/10.1109/ICCVW.2017.365

Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods, 52*, 723–740. https://doi.org/10.3758/s13428-019-01271-9

Rautaray, S. S. (2012). Real time hand gesture recognition system for dynamic applications. *International Journal of UbiComp*, 3(1). https://ssrn.com/abstract=3702844

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review, 43*(1), 1–54. https://doi.org/10.1007/s10462-012-9356-9

Rekimoto, J. (2001). Gesturewrist and gesturepad: Unobtrusive wearable interaction devices. *Proceedings Fifth International Symposium on Wearable Computers*. https://doi.org/10.1109/ISWC.2001.962092

Ripperda, J., Drijvers, L., & Holler, J. (2020). Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): A toolkit for the automatic detection of hand movements and gestures in video data. *Behavior Research Methods, 52*(4), 1783–1794. https://doi.org/10.3758/s13428-020-01350-2

Ruffieux, S., Lalanne, D., Mugellini, E., & Abou Khaled, O. (2014). A survey of datasets for human gesture recognition. *International Conference on Human-Computer Interaction*. https://doi.org/10.1007/978-3-319-07230-2_33

Sagawa, H., & Takeuchi, M. (2000). A method for recognizing a sequence of sign language words represented in a Japanese sign language sentence. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*. https://doi.org/10.1109/AFGR.2000.840671

Schreer, O. & Masneri, S. (2014). Automatic video analysis for annotation of human body motion in humanities research. In *Workshop on multimodal corpora in conjunction with language resources and evaluation conference* (pp. 29–32).

Schumacher, J., Sakič, D., Grumpe, A., Fink, G. A., & Wöhler, C. (2012). Active learning of ensemble classifiers for gesture recognition. In *Joint DAGM (German Association for Pattern Recognition) and OAGM symposium* (pp. 498–507). https://doi.org/10.1007/978-3-642-32717-9_50

Sharma, R., Cai, J., Chakravarthy, S., Poddar, I., & Sethi, Y. (2000). Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*. https://doi.org/10.1109/AFGR.2000.840669

Streeck, J. (2008). Gesture in political communication: A case study of the democratic presidential candidates during the 2004 primary campaign. *Research on Language and Social Interaction, 41*(2), 154–186. https://doi.org/10.1080/08351810802028662

Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods, 51*(2), 769–777. https://doi.org/10.3758/s13428-018-1086-8

Vardy, A., Robinson, J., & Cheng, L. T. (1999). The wristcam as input device. In *Digest of papers. Third international symposium on wearable computers* (pp. 199–202). https://doi.org/10.1109/ISWC.1999.806928

Von Agris, U., Knorr, M., & Kraiss, K. F. (2008). The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition* (pp. 1–6). https://doi.org/10.1109/AFGR.2008.4813472

Waldherr, S., Romero, R., & Thrun, S. (2000). A gesture based interface for human-robot interaction. *Autonomous Robots, 9*(2), 151–173. https://doi.org/10.1023/A:1008918401478

Wan, J., Lin, C., Wen, L., Li, Y., Miao, Q., Escalera, S., & Li, S. Z. (2020). ChaLearn looking at people: IsoGD and ConGD Large-Scale RGB-D gesture recognition. *IEEE Transactions on Cybernetics (early Access)*. https://doi.org/10.1109/TCYB.2020.3012092

Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., & Li, S. Z. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2016.100

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *Proceedings of the fifth international conference on language resources and evaluation* (pp. 1556–1559).