Noname manuscript No.
(will be inserted by the editor)

A semi-automatic annotation tool for unobtrusive gesture analysis

Stijn De Beugher · Geert Brône · Toon Goedemé

Received: date / Accepted: date

Abstract In a variety of research fields, including linguistics, human-computer interaction research, psychology, sociology and behavioral studies, there is a growing interest in the role of gestural behavior related to speech and other modalities. The analysis of multimodal communication requires high-quality video data and detailed annotation of the different semiotic resources under scrutiny. In the majority of cases, the annotation of hand position, hand motion, gesture type, etc. is done manually, which is a time-consuming enterprise requiring multiple annotators and substantial resources. In this paper we present a semi-automatic alternative, in which the focus lies on minimizing the manual workload while guaranteeing highly accurate annotations. First, we discuss our approach, which consists of several processing steps such as identifying the hands in images, calculating motion of the hands, segmenting the recording in gesture and non-gesture events, etc. Second, we validate our approach against two existing corpora and one own-recorded corpus in terms of accuracy and usefulness. The proposed approach is designed to provide annotations according to the McNeill (1992) gesture space and the output is compatible with annotation tools such as ELAN or ANVIL.

Keywords (semi)-automatic annotation · gesture analysis · video analysis · hand annotation · gesture space · motion analysis

1 Introduction

Research on human communication and interaction increasingly focuses on the embodied nature of these processes and more specifically on the interplay between speech and gesture in construing and coordinating meaning. A variety of disciplines, including linguistics, psychology, sociology, human-computer interaction research, and behavioral studies inquire into the specifics of multimodal com-

Stijn De Beugher · Toon Goedemé
EAVISE KU Leuven Belgium,
E-mail: stijn.debeugher@kuleuven.be E-mail: toon.goedeme@kuleuven.be

Geert Brône
MIDI KU Leuven Belgium, E-mail: geert.brone@kuleuven.be

munication (see e.g. Müller et al (2013, 2014); Jewitt (2009) for multidisciplinary overviews). One of the main challenges for this type of analysis, independent of the specific research question or approach, is gaining access to qualitatively and quantitatively rich data. Experimental and corpus-based studies rely on high-quality video data of language in (inter)action, with detailed annotations of relevant parameters related to speech (transcriptions including verbal and para-verbal information such as hesitation markers, pauses, intonational contours, etc.) and bodily action (including hand gestures, body posture, eye gaze data, etc.). For the majority of studies and availably resources (corpora & databases), the annotation work was done manually, based on existing multimodal coding schemes (see Kipp et al (2009); Bressem (2013); Abuczki and Esfandiari (2013) for overviews of such schemes). This is a labor-intensive process, requiring multiple annotators and thus substantial resources. As a general guideline, it is assumed that the annotation of a recording has a time-ratio of at least 10:1, thus one minute of video material takes up a minimum of 10 minutes of annotation, depending on the level of detail required. When a detailed transcription is needed, in combination with the annotation of several multimodal layers, this ratio can easily amount to 50:1, which makes the compilation of large-scale annotated data sets practically unfeasible. Due to this issue of labor intensiveness, some large scale databases designed for multimodal analysis have not been annotated yet, as is the case for the Distributed Little Red Hen Lab (http://www.redhenlab.org/). For projects such as these, a (semi-)automatic approach to at least some steps in the annotation process is essential. Only by introducing such an approach, the wealth of available data can be made accessible for multimodal analysis.

In this paper, we present a semi-automatic tool for the segmentation and annotation of specific gestural features during language production. The purpose of this tool is to provide a reliable basic annotation that can be easily enriched by further manual analysis. The first step in the development of an automated gesture detection algorithm logically consists of measuring the position of the hands in each frame of a video. This information can be obtained in several ways, including depth sensors such as the Kinect or motion capturing systems such as OptiTrack. Since the majority of (existing) recordings are made using standard 2D cameras, we need to rely on other methods. Based on a comparison of existing methods, we selected several concepts as basis of our approach. On top of that, we present a novel integration of a minimal amount of manual interventions to obtain highly accurate estimations of hand positions. Once the exact position of each hand in each frame is determined, the next step consists of a basic gesture annotation. The gesture analysis algorithm produces a segmentation of gesture and non-gesture sequences, a location in gesture space based on McNeill's model (McNeill 1992), and an indication of the directionality of the gesture. For each of these dimensions, an XML file is generated, which makes the output compatible with multimodal annotation tools such as ELAN or ANVIL, and integratable with existing transcriptions and annotations.

To maximize the applicability of our approach, the system must deal with the following challenges: (i) the system should be maximally unobtrusive, i.e. it should work on existing video data without information collected with markers or other sensors (ii) the system has to function even if the position of the camera changes during the recording (e.g. recordings made with head-mounted cameras or mobile eye-trackers). The participant must have the ability to move freely (iii)

the accuracy of the resulting annotated gesture sequence should be very high, requiring virtually no manual correction afterwards.

To summarize, our goal is to develop a highly accurate, semi-automatic gesture analysis tool that is applicable on recordings in which a minimum of restrictions is imposed on the participants. Using such an approach will benefit the analysis of this type of data. In general, the analysis of gesture in human interaction is based on either experimental data or on relatively small-scale corpora. The first method has the advantage that the researcher can control some of the variables, making it easier to elicit rich data (and thus to manually process the collected data). The drawback is (i) that researchers need to design new experiments for each new research question, and (ii) that it is notoriously difficult to elicit naturally occurring interaction in a controlled lab setting. The second method, using video corpus data of spontaneous interactions, obviously reduces the latter risk, but comes with a cost as well. Naturally occurring data may confront the researcher with relative data scarcity (low density of the phenomenon under scrutiny in the data), thus requiring the collection of a large corpus to be able to make well-founded claims. Given the above-mentioned challenges of lab our-intensiveness, multimodal corpus studies based on *big data* are scarce, especially in comparison to the strong quantitative corpus movement that can be observed for written language. A similar development for multimodal corpora would, however, open up a massive new area of research on *multimodal patterning*, i.e. the study of recurrent co-occurrences between verbal and nonverbal resources (e.g. markers of obviousness co-occurring with shoulder shrugs, hesitation markers such as *uhm* co-occurring with gaze aversion, etc., see Feyaerts et al (2016) for an overview). A semi-automatic annotation procedure like the one presented in this paper can pave the way for a more quantitative approach.

We present our results on two subsets of existing corpora: the NeuroPeirce corpus (Brenger and Mittelberg 2015) and The Bielefeld Speech and Gesture Alignment Corpus (SaGA) (Lücking et al 2010) and on one own recorded corpus.

The remainder of this paper is structured as follows: in section 2 we give a detailed overview of existing methods for the detection of human hands in images as well as the automatic detection of gestures. In section 3, an introduction to our hand annotation approach is given, while in section 4, we give a thorough explanation of our gesture detection approach including the characteristics we are able to extract. Finally, in section 5 we validate our approach on the three corpora and in section 6, we formulate some concluding remarks.

2 Related work

This section presents an overview of existing algorithms and approaches relevant to the field of gesture and interaction analysis. Since the concept of automated gesture analysis is a general term, we narrow down the focus of this paper as the detection of gestures in standard 2D color camera images. It is important to highlight the difference between gesture *detection* and gesture *recognition*. Gesture detection essentially involves segmenting gesture sequences from non-gesture sequences. Gesture recognition, on the other hand, requires the re-identification of specific gestures. In this paper, the focus is primarily on gesture detection as a first but essential step in any gesture annotation process. Moreover, there is also a

difference between fine-grained finger movements with specific semantics, as is the case in sign languages, and the larger movement of the entire hand as pointing or waving. In this paper, we focus on the larger movements rather than the detection of individual finger poses, again as a first but necessary step towards even more fine-grained systems. In what follows, we present an overview of existing techniques regarding various aspects of gesture analysis. The following subsections are organized as follows: in subsection 2.1 some state-of-the art approaches for fine grained finger movements are described. In subsection 2.2 a short overview of approaches that combine several modalities is given. Finally subsection 2.3 gives an overview of techniques that work on traditional 2D images.

2.1 Fine grained finger movement

Rautaray and Agrawal (2012) present a thorough overview of existing approaches to the analysis of fine-grained finger gestures. In another recent contribution, Badi (2016) proposes a novel method for the recognition of six specific hand poses in the context of human-computer interaction (HCI): open, close, cut, paste, maximize and minimize. Input images are standard RGB images, but severely conditioned: containing only a single hand on a black background. Two features are extracted from these images: hand contours and complex hand moments. In a final step, these features are used in an Artificial Neural Network (ANN) classifier to identify the different hand poses. Another approach to the recognition of specific finger poses can be found in the work of Kapuscinski et al (2015). Here, two types of depth cameras, viz. ToF (time-of-flight) and Kinect, are used to recognize hand gestures. Next to the identification of dynamic gestures such as "to feel" or "to ache", they also developed a technique for the recognition of the specific signs for the Polish finger alphabet. In this system two classification approaches are compared: a Hidden Markov models (HMM) classifier and a nearest neighbors technique with dynamic time warping (DTW), allowing a non-linear mapping of one pose to another by minimizing the distance between them. A similar approach is found in Kuznetsova et al (2013), where a highly precise method for the recognition of static hand gestures is proposed using data from a consumer depth camera. In addition to the approach of Kapuscinski et al (2015), a multi-layered random forest (MLRF) classifier is used to identify different signs such as the 24 letters of American Sign Language (ASL). It is important to note that in the techniques described above, the input data contain information of a single hand, either in a standard image or in depth information. In the approach we present in this paper, we are interested in larger movements of the entire upper body. Here, the input data traditionally contain footage of an entire person in a much more natural setting and thus, an additional challenge consists in the segmentation of relevant body parts from the background. In the next subsection, several existing approaches to this type of gesture analysis are discussed.

2.2 Hand detection using a combination of multiple modalities

A recent challenge that addresses gesture recognition for larger movements is the *Chalearn looking at people challenge* (Escalera et al 2015). In the model presented

here, several modalities can be utilized to automatically recognize a vocabulary of 20 Italian cultural/anthropological signs in image sequences. These modalities include RGB images, depth images, skeleton representation and binary masks. As expected the top-competitors (Monnier et al 2015; Neverova et al 2014; Chang 2015) of this challenge combine several modalities to achieve top accuracy. Another approach in which several modalities are used was developed by Yin and Davis (2013). Here RGB, depth and data from a motion capturing system (Xsens) are combined to locate the position of both hands. An off-the-shelf skin segmentation is used to mask the Kinect depth data and combined with a motion mask averaged over three frames. Once the position of the hands is found in each frame, a HMM is trained for each phase for each gesture. This means that a separate model was trained for pre-stroke, nucleus and post-stroke. In the end, a Viterbi decoding was used to optimally segment the gesture sequences. The above-mentioned approaches are capable of detecting gesture sequences and they can identify specific gestures. However, since not all existing recordings are captured using multiple and/or specific camera's, our goal is to develop a system that is able to detect gestures in normal RGB images in natural settings. In the next subsection we discuss existing approaches that only rely on RGB data.

2.3 Hand and gesture detection in RGB images

In case a recording was made without a depth sensor nor motion tracking system, the level of complexity of gesture detection increases significantly. Due to the lack of this additional depth and skeleton information, one needs to detect the relevant body parts (i.e. hands, face, torso) in advance and thereafter one could start detecting the gesture sequences. In recent years, some promising approaches for the detection of human hands in images were developed. In this subsection, we thoroughly discuss several methods and approaches for accomplishing this complex task. In general, gesture analysis in standard RGB images consists of two main phases: first the retrieval of the hand positions, and secondly, the segmentation of the recording in gesture and non-gesture sequences using extracted information from the hands.

The first step in an automatic gesture detection in RGB images is retrieving the hand positions in each frame. In general, this is done using skin segmentation in various ways. A first approach is using manual fine-tuning of a set of sliders selecting threshold on specific color channels (Gebre et al 2012; Schreer and Masneri 2014). Often the images are therefore converted to another color range such as HSV, since skin segmentation in RGB is known to be sensitive for slight illumination changes. Other approaches automatically detect a face (Viola and Jones 2001) in an image and uses this for the extraction of skin information (Mittal et al 2011). In Jones and Rehg (2002) a statistical color model was developed, allowing the calculation of skin-tone probability of each pixel.

Skin segmentation is often combined with monitoring the velocity of the skin regions. Obtaining this motion is done in several ways: a basis approach is to calculate the displacement in subsequent frames, but more sophisticated methods such as Mixtures of Gaussian (MoGs) are also widely applied (Zhang et al 2014). In Marcos-Ramiro et al (2013) an approach for modeling non-verbal communication was presented and here a 2D hand likelihood map was developed. This map

follows the assumption that in an image, the hands are skin colored and that they show more movement than the face, which is obviously also skin colored. In Alon et al (2009), the same assumption was followed and here the motion of the skin regions was applied to further refine the hand detections.

Other methods for the detection of human hands exist as well, whether or not combined with the skin segmentation. An example is the approach of Mittal et al (2011), in which a Deformable Part Model (DPM) (Felzenszwalb et al 2010) of a human hand was developed. By applying this model to an image, they are able to recognize human hands in various poses. In Bennewitz et al (2008) a similar methodology was used, but here a Haar feature classifier (Viola and Jones 2001) was used to detects the hands. Another model based approach is found in Karlinsky et al (2010) and is capable to locate parts of interest in a robust and precise manner, even when the surrounding context is highly variable and deformable. Applied to hand detection, the chains model generates a feature chain between an easily detectable object, such as a face, and the object of interest (i.e the hands). Although model based approaches allow for accurate hand detections, they are often extremely slow, making them inapplicable for real-life use. For example, the approach of Mittal et al (2011) requires more than 200 seconds for the processing of a single image frame of 1280×720 pixels.

Many hand detection approaches work fine for still images, however when applying them to recordings where persons move naturally a problem arises. Since these algorithms obtain no information regarding the human pose, it is impossible to discriminate left and right hand. Nevertheless such a distinction is indispensable in gesture analysis. In Marcos-Ramiro et al (2013), next to the hands, a face is also detected, making the hand positions relative w.r.t the position of the person. This is combined with a synthetic 3D polygonal torso model, resulting is an approximated 3D pose of upper body of the person. Another approach for distinguishing left and right hand is found in Bennewitz et al (2008) here, next to applying a model for the detection of a human hand, they also use specific models for left and right hand allowing them to differentiate both hands.

Once the location of both hands is found in each image of a recording, a gesture detection algorithm can be applied. The purpose of such an algorithm is to segment a recording in gesture and non-gesture sequences. In Gebre et al (2012) an approach for the automatic detection of gesture strokes was presented. Next to the skin segmentation, they also apply a corner tracking algorithm to the segmented image. Their approach is developed to cluster three sets of corners: one cluster for each hand and one for the head, assuming there is only one person presented in the video. Finally, values extracted from the clusters of corners are fed into a machine learning algorithm that is trained to predict whether or not a given frame is inside a stroke. Unfortunately, they achieve an average accuracy (F1-measure) of only 38.71%, which makes their approach insufficiently accurate for practical use. Another gesture spotting system is presented by Peng et al (2015) here, a simple yet effective approach to divide a video in short clips of gestures and non-gestures was applied. They assume that the hands of an actor are almost in the same position when he or she is not performing a gesture. Using this assumption, one could determine a static hand position for each hand. By performing a frame-based calculation of the distance between each hand and the static hand positions, one could easily distinguish the gestures from the non-gestures. However they achieve high accuracy in terms of recognizing different gestures, they do not provide any

measurements of the detection of gesture sequences. Since they define a static position for each hand, their approach can only be applied in a context with a fixed camera and a immobile subject. In Schreer and Masneri (2014) an automatic video analysis for the annotation of human body motion in humanities research was presented, which is highly similar to our goal. The first step in their approach consists of a skin color segmentation that is done manually using a set of sliders. Once the skin-segmentation was done, their software tracks the hands based on the motion of them. This motion information is used to segment the video in gesture and non-gesture parts. On top of the detection of gesture sequences, their tool also provides information regarding the type of movement in terms of: Phasic, Repetitive and Irregular. Next to the gesture sequence detection, they also provide automatic information on the position of the hands related to the body as defined in the McNeill gesture space (McNeill 1992). Despite their efforts, the accuracy of the gesture sequence detection on their datasets is limited to 75.3%. In our opinion, another drawback of this approach is the skin segmentation using a set of sliders, making the accuracy of their approach unpredictable.

It is clear that automatic gesture detection, although it has been studied for several years, remains a hot topic in several research fields. Many approaches rely on multiple modalities to detect the gesture sequence. In the above-mentioned papers, some novel methods for the detection of human hands as well as the detection of gestures are presented. Unfortunately none of the above-mentioned approaches meets our imposed requirements in terms of applicability and accuracy. We conclude that a fully automatic approach will never reach the necessary accuracy. Nevertheless, we used some of the previously described concepts in our implementation to develop the gesture detection algorithm. For example: taking into account the distance between a hand and its static position as proposed in Peng et al (2015) and expressing the usage of the gesture space according to the McNeill definition as proposed in Schreer and Masneri (2014). In the next section, we discuss the methodology that was used for retrieving the position of the hands.

3 (Semi-) automatic hand annotation

As mentioned above, each gesture analysis algorithm starts with retrieving the positions of the hands in the recording. Here, we discuss our approach for the accurate extraction of the hand positions in images. To accomplish that goal, we argue that the only way to ensure highly accurate annotations in all circumstances is to integrate a minimal amount of manual intervention in an automatic detection approach. Moreover, this offers the gesture analysis researcher a minimal but important amount of control and quality assurance over the annotation process. In the next subsections, we give a more detailed description of our approach, illustrated in 1.

3.1 Semi-automatic approach

One of the most important contributions of our hand annotation approach is the novel integration of manual interventions. For each frame we analyze, our tool automatically calculates a confidence score of the hands. This confidence score

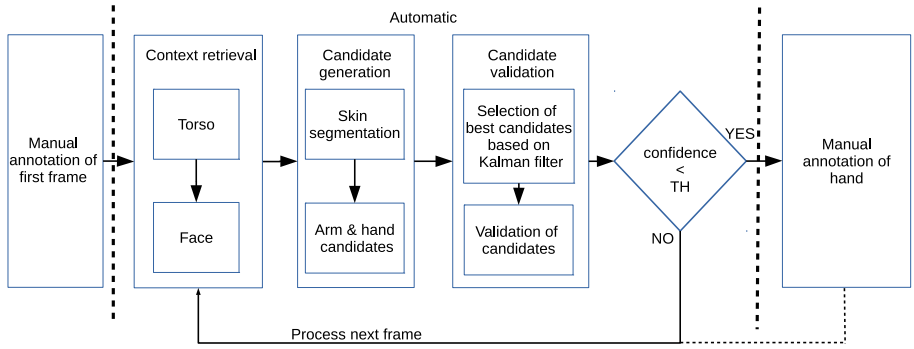


Fig. 1: Workflow of our semi-automatic hand annotation approach.

Fig. 2: Generation of hand candidates: a) original image, b) skin segmentation, c) contour detection, d) fit ellipse, e) final hand candidates.

indicates how confident our algorithm is that the correct hands are detected. In case the confidence of a hand drops below a predefined threshold, the automatic analysis is halted. Then and only then manual intervention is asked from the user using a graphical user interface in which one can easily manually annotate the position of the respective hand. After this intervention, our automatic analysis continues. As shown in figure 1, the first frame of each recording is also annotated manually ensuring a good starting point.

3.2 Retrieve hand candidates

Retrieving the hands in an image is done automatically using a combination of several image processing techniques as shown in figure 2. First, we start by applying a human torso detector to find out whether there is a person present in the image. Once we have that location information, we define a region around that person in which we search for his hands as indicated by the blue rectangle in figure 2a. Next, an automatic skin segmentation is applied to this cropped region (figure 2b), using a set of predefined color ranges (Rahim et al 2006) as well as color information that is automatically obtained from an automatic face detection (Viola and Jones 2001). This is different from some of the other approaches discussed above, such as (Schreer and Masneri 2014; Gebre et al 2012) where the skin segmentation relies on manual fine-tuning. In a next step, we fit a contour around each group of skin pixels using the hypothesis that some contours coincide with the actual hands. This result is shown in figure 2c.

It is clear that when the person is wearing short sleeves as in figure 2, the entire arm is selected using our skin segmentation technique. Since our goal is to retrieve the location of the hand, we need to apply an additional analysis. Therefore, we fit an ellipse around each contour and we identify the endpoints of its major axis (see figure 2d and e). Here we assume that when this axis coincides with an arm, one end point overlaps with the hand, while the other endpoint overlaps with the elbow. In case the subject wears long sleeves the same approach is applied: one endpoint corresponds to the fingertips while the other endpoint corresponds to the wrist. To summarize our approach: we fit an ellipse around each contour and we assume that one endpoint of the major axis will correspond to the hand, while the other endpoint overlaps with a joint (either wrist or elbow). The assignment of the endpoints is described in the next subsection.

3.3 Filter hand candidates

Figure 2d shows four ellipses, which illustrates that a procedure must be defined to remove wrong candidates. First, we remove the contours that coincide with the face, which is straightforward since we use a face detection algorithm (Viola and Jones 2001) as shown in figure 2e. In order to find which endpoint corresponds to an actual hand, we apply two methodologies: tracking and pose validation.

Our annotation tool was developed to detect human hands in a sequence of images (such as a video) rather than in isolated frames. In fact, we employ the sequential character of a video to further enhance the recognition rate. Once we know the exact position of a hand in a particular frame (obtained using manual intervention), we are able to track the position of that hand over time using a mathematical filter (i.e. a Kalman filter (Kalman 1960) with a constant velocity motion model). By using such a filter we can predict the position of the hand in the next frame(s). This allows us to easily assign a hand-label to the endpoint that is closest to this prediction. Once a hand-label was assigned to one endpoint of an axis, we automatically assign a joint-label to the remaining endpoint. Of course, this approach is used for both left and right hand.

To verify the correctness of the chosen endpoints, we developed a validation method based on knowledge of the human pose. From a fully annotated video corpus, (Ferrari et al 2009) we built up a set of probability maps of human arm poses, representing all possible poses a human arm may take w.r.t. the torso. We compare our obtained poses against these maps and calculate the likelihood that this particular pose is a valid one. Finally we take the tracking information and the likelihood into account in the confidence calculation. In case the confidence in a particular frame drops below a predefined threshold, we interrupt the automatic analysis and ask for manual intervention. On the other hand, in case the confidence lies above the threshold, the position of both hands and joints are stored in a temporary text file. On top of that, we also store both face and torso detection in each frame, since it might be useful for further analysis. In figure 3 we show some examples of our technique on four image frames. The green circles represent the final hand detections, while the yellow circles indicate the joints. In the last image of this figure, we see that even when only the hand is visible, our approach still manages to detect the hand. In this case the joint corresponds to the wrist

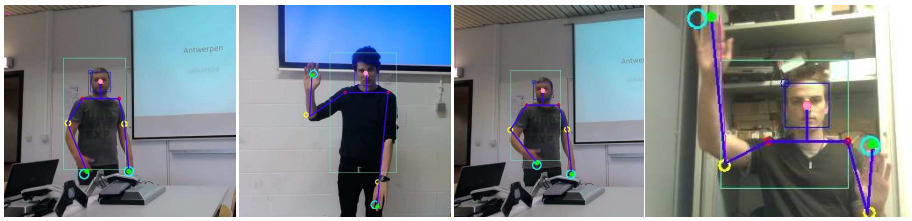


Fig. 3: Examples of our detections on four image frames. Green circles are the hand detections, yellow circles are the corresponding joints. Green rectangles represent the torso detections and the blue rectangles represent the face detections.

instead of the elbow as shown in the other frames. A video illustrating our system is available online [1].

An experimental validation of this hand annotation approach is given in section 5, where we discuss the accuracy of the proposed method and the degree of manual intervention needed. For more information regarding this approach, we refer to De Beugher et al (2016). The above-mentioned approach is used as basis of our gesture analysis tool as discussed in the next section.

4 Gesture detection

Once the positions of the hands in an entire recording are retrieved, automatic gesture analysis can be initiated. As discussed in section 2, there are several approaches for this type of analysis. Despite the large variety in approaches, we propose the development of a gesture analysis tool in which we impose a minimum of restrictions to the participants in order to enlarge the applicability. For example, our gesture analysis tool should be able to handle participants in a sitting as well as standing position. On top of that, we allow a participant to walk during the recording and we even allow a moving camera position, which is particularly useful for recordings made with wearable cameras (e.g. GoPro) or head-mounted eye-tracking systems (e.g. Tobii, SMI). Finally it is important to notice that our gesture analysis tool relies only on the semi-automatic annotations as described in the previous section (i.e. annotations of the hands and the position of the human torso). No additional information such as depth information or motion sensors is required.

Our gesture analysis tool consists of several blocks as shown in figure 4. Each individual block is described in the following sub-sections, starting with the calculation of the rest position in subsection 4.1. Once the rest positions are known, we segment a recording in gesture segments and non-gesture segments based on the displacement between each hand and its respective rest position (subsection 4.2). Subsection 4.3 discusses our algorithm to automatically provide information concerning the gesture space. A final analysis is applied on the directionality of the gestures is given in subsection 4.4.

[1] *http://youtu.be/DsxdBc4gGjg*

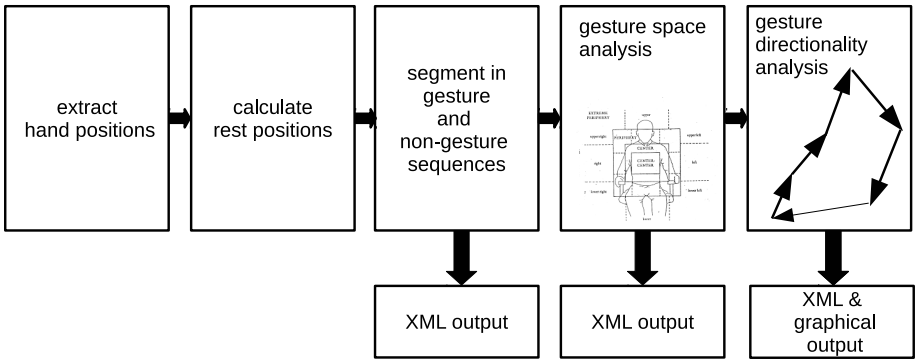


Fig. 4: Workflow of our automatic gesture analysis tool. This figure also reveals which type of analysis results in XML compatible output.

4.1 Calculation of the rest positions

Starting from the hand positions, there are two main approaches to segment a recording into gesture and non-gesture sequences. The first method monitors the velocity of the hands with the assumption that a hand does not move when one is not gesturing, as is proposed in Schreer and Masneri (2014). A second approach measures the distance between a rest position (i.e. the position of the hands when one is not gesturing) and the hands as proposed by Peng et al (2015). Despite the fact that both approaches are widely applied in the literature, we argue that the second one is more efficient. Indeed, during our experiments we noticed that, for particular large gestures, the velocity of the hands stalls within the gesture. An obvious example is the hold phase of a pointing gesture. Here, the above-mentioned approach would not recognize the hold phase as (part of a) gesture, since there is little or no velocity of the hand.

Essential for the chosen approach is the determination of the rest positions. We can define this position as: *the position were the hand is positioned most frequently during an entire recording.* An obvious approach is simply plotting the positions of both left and right hand into a map and calculating a local maximum for each hand. In our application on the other hand, we allow moving participants as well as a moving camera viewpoint, which means we need to transform the coordinates of the hands relative to the position of the participant. Since we use the annotations that are retrieved using our semi-automatic approach, we have access to the coordinates of the human torso detection in each frame to accomplish this task. The transformation of the hand coordinates is given in equation 1 where x_H^{rel} represents the relative x-coordinate of the hand, x_H stands for the original x-coordinate of the hand, x_T the center of the human torso detection and w_T width of the torso detection. The same methodology is used for the y-coordinates:

$$\begin{aligned} x_H^{rel} &= \frac{x_H - x_T}{w_T} \\ y_H^{rel} &= \frac{y_H - y_T}{w_T} \end{aligned} \tag{1}$$

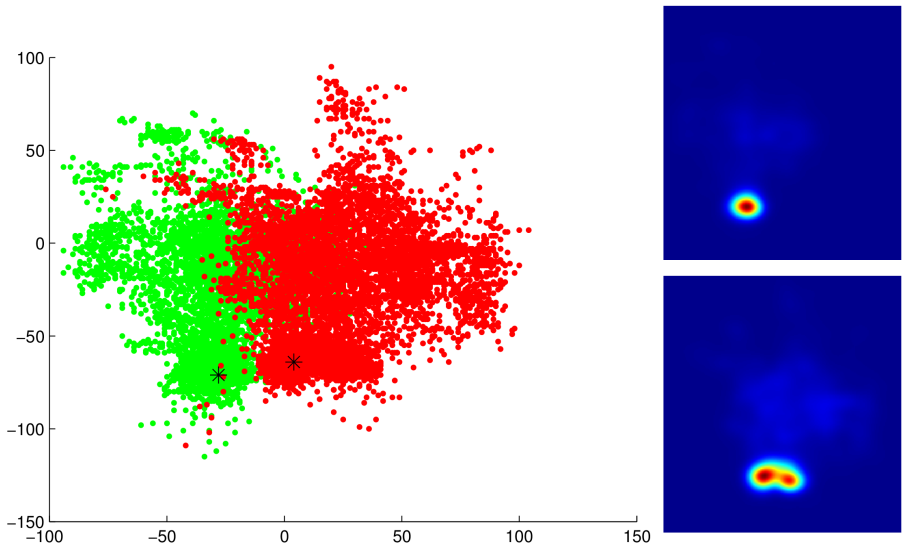


Fig. 5: Normalized hand positions of an entire recording. Green dots represent right hands, red dots represent left hands. Two asterisks indicate the respective rest positions. The upper smoothed map belongs to the right hand coordinates, the bottom smoothed map belongs to the left hand coordinates.

By applying this transformation to each frame of a recording we obtain a map as shown in the left part of figure 5. The two asterisks represent the local maxima for both left and right hand. These are indeed the rest positions for each hand for that particular recording. They are obtained by extracting the local maxima for each hand in a Gaussian smoothed map as shown in the right part of figure 5. Here the color represents the density of the hand coordinates: red means dense coordinates, whereas blue means sparse coordinates. Once the rest positions are known, we are able to segment an entire recording in terms of gesture sequences and non-gesture sequences as described in the next section.

4.2 Gesture phase segmentation

The segmentation of gesture and non-gesture sequences is done by calculating the displacement between each hand and its respective rest position. Once the displacement of at least one hand is beyond a set threshold we assume that a gesture sequence was initiated. When subsequently the displacement drops below this threshold, the gesture sequence ends since the corresponding hand is back in the vicinity of the rest position. Using this methodology we are able to segment an entire recording in gesture and non-gesture segments.

A decisive aspect in this approach is the calculation of the optimal threshold value. This value indicates how much deviation from the rest position is allowed before our software initiates a gesture sequence. A straightforward approach is to define a fixed value that is used for both hands. However, our experiments revealed that it is challenging to find a unique value that results in accurate segmentation

of multiple recordings each with its own characteristics. To overcome this problem we proposed a set of solutions: a) separate thresholds for both left and right hand, b) use separate threshold values in both x and y direction, and c) obtain threshold values from the data itself by extracting a sigma from the Gaussian smoothed hand maps, as shown in the right part of figure 5. An illustration of this technique for the right hand in a recording is given in figure 6. Here we plot, for the first 1000 frames of a recording, the distance between the rest position and the assumed right hand. The red line in the top part represents the applied threshold. In the bottom part, we illustrate the gesture segmentation based on this displacement.

$$\begin{aligned} GS = \{(D_{LX} > \alpha\sigma_{LX}) \vee (D_{LY} > \alpha\sigma_{LY})\} \vee \\ \{(D_{RX} > \alpha\sigma_{RX}) \vee (D_{RY} > \alpha\sigma_{RY})\} \end{aligned} \tag{2}$$

In equation 2 the condition to initiate a gesture sequence GS is given. D_{LX} represents the displacement of the left hand in x-direction, σ_{LX} is the sigma value for the left hand in x-direction that was obtained from the smoothed map and finally α is a tuning parameter that can be used to fine-tune our system. If at least one of the four displacements exceeds its respective threshold, a gesture sequence is initiated. The accuracy of this segmentation is thoroughly discussed in section 5.

As a final point, it is important to mention that this analysis is automatically written into an XML file, containing a the gesture segmentation annotations. Next to this basic segmentation, our gesture analysis tool also generates information regarding the usage of the gesture space as described in the next subsection. This file type is compatible with existing multimodal annotation tools such as ELAN or ANVIL.

4.3 Gesture space annotation

Researchers in gesture studies are interested in the spatial distribution of gestures, i.e. where in the gesture space a gesture occurs. A commonly used methodology for this purpose is the gesture space as defined by McNeill (1992) and illustrated in figure 7. He proposed to divide, the space into sectors using a system of concentric squares. The sector directly in front of the chest is the center-center sector, surrounding this, we the center sector is defined. Then the periphery, which is sub divided into upper, lower, right and left. Finally, the extreme periphery is defined, which is divided in even more sub sectors. Manually annotating the gesture space is extremely labor-intensive since ideally, one has to assign a specific gesture sector to each individual frame of a gesture sequence. In order to reduce this workload, we noticed that the manual analysis of the gesture space is often reduced to the allocation of a single sector for each entire gesture. For example: the annotation of the sector where the majority of the gesture occurs or the annotation of the sector where the gesture is the largest. It is clear that such an annotation reveals only a fraction of the spatial information. In order to overcome this problem, we can use our semi-automatic hand annotations as described in section 3 to automatically annotate the gesture space. As mentioned before, next to the hands both face and human torso are detected, as illustrated in figure 3. We defined a mathematical relationship between the face -, torso detection and the individual gesture sectors as defined by McNeill. This allows us to automatically define the gesture sectors on

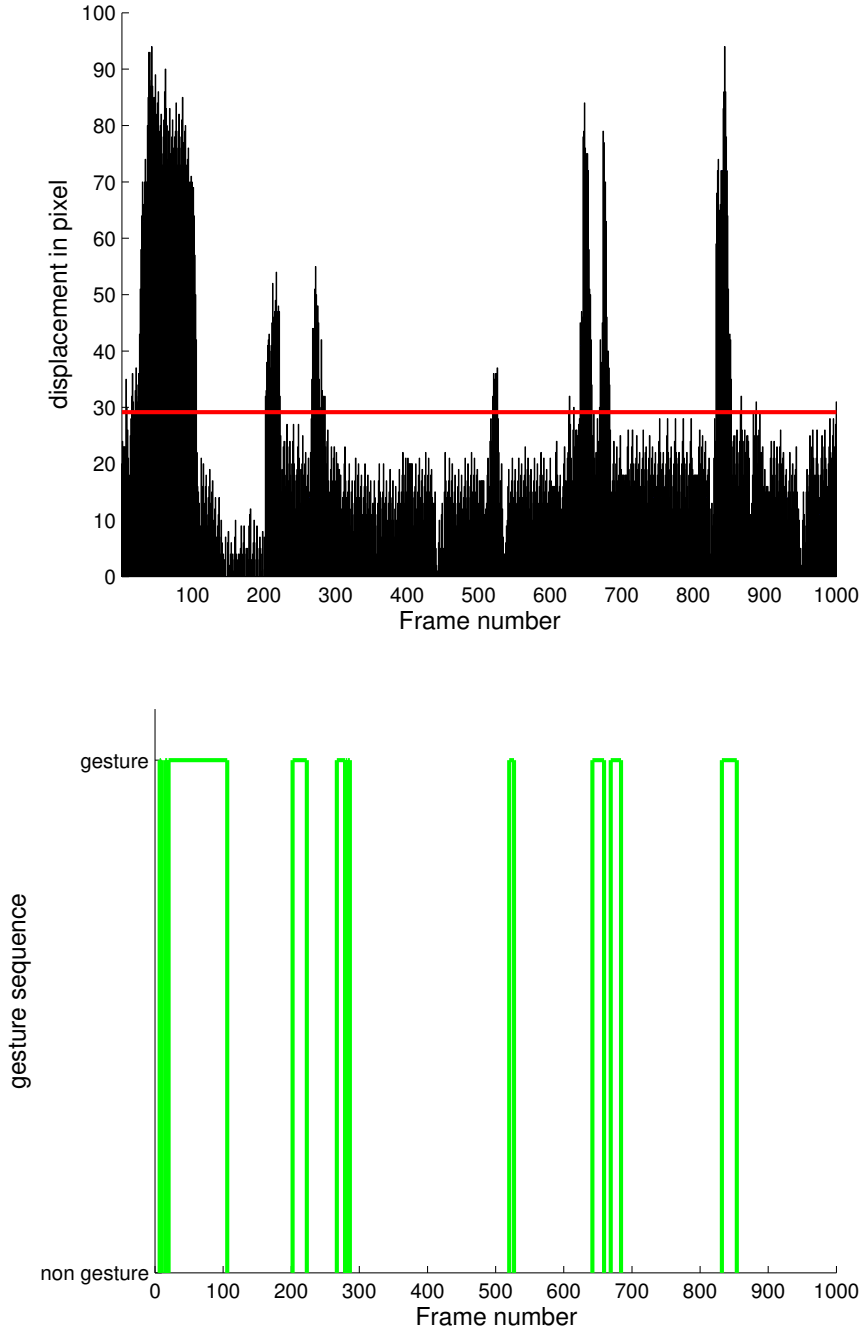


Fig. 6: Top part: displacement in of the right hand w.r.t the rest position. Red line indicates the applied threshold. Bottom part: gesture segmentation that is generated using this displacement.

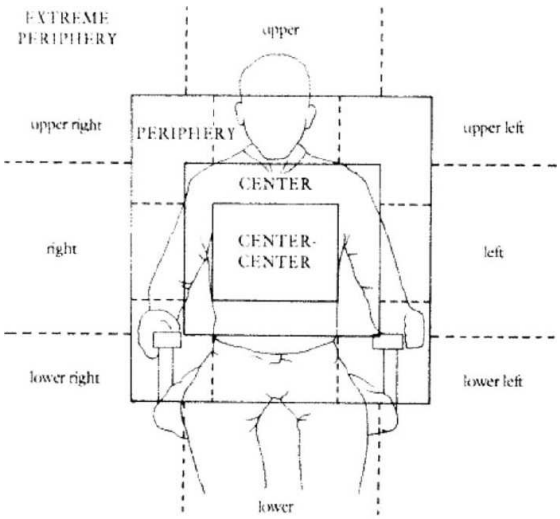


Fig. 7: Gesture space as defined by McNeill (1992). We can distinguish 4 larger sectors as represented by capital letters as well as the respective sub sectors.

each individual image as shown in figure 8. Here we distinguish four larger sectors: center-center, center, periphery and extreme periphery as well as the subdivisions for both periphery and extreme periphery. Using these automatically generated sectors we are able to easily determine in which sector each hand is located at each moment. Therefore our system compares automatically the hand coordinates with the coordinates of each sector. Similar to the previously mentioned approach, this analysis is also stored in an XML file. For each hand a unique tier is added in which, for each gesture sequence, the usage of the sectors is expressed. Compared to manual analysis, our approach always provides a frame-based analysis of the gesture space, which is in case of manual analysis practically infeasible. Since our automatic analysis generates gesture sectors based on both face and torso detection, we are able to ensure a consistent and non-subjective definition of the sectors across several recordings. In manual labelling on the other hand, significant differences exist in the exact definition of the sectors between several annotators. Our automatic system excludes these unwanted side effects.

4.4 Gesture directionality

A final type of analysis to be included in our system is the directionality of gestures as discussed in this subsection. Using the above-mentioned approaches, we are able to automatically segment a recording in gesture and non-gesture sequences. Furthermore, we can automatically provide information regarding the gesture space for each individual frame. Another vital aspect of gesture analysis is the directionality of gestures. Researchers are interested in the direction and movement of each gesture, resulting in a specific trajectory of each hand (see e.g. the gesture annotation scheme presented by Bressem (2013)). Although this is of great importance

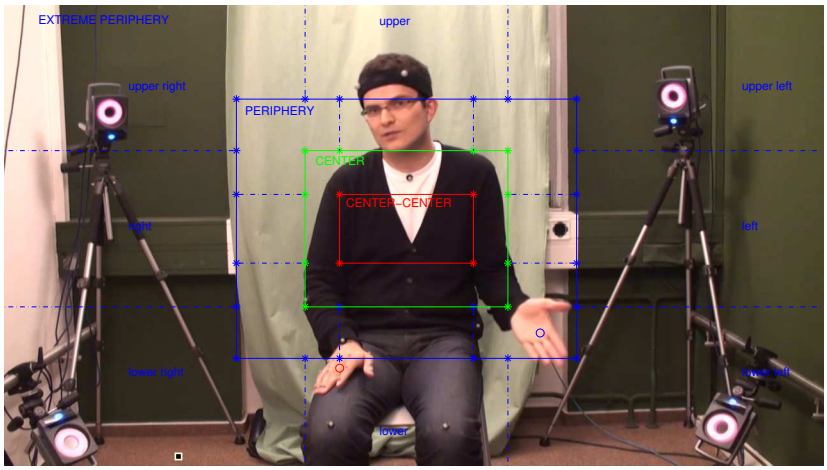
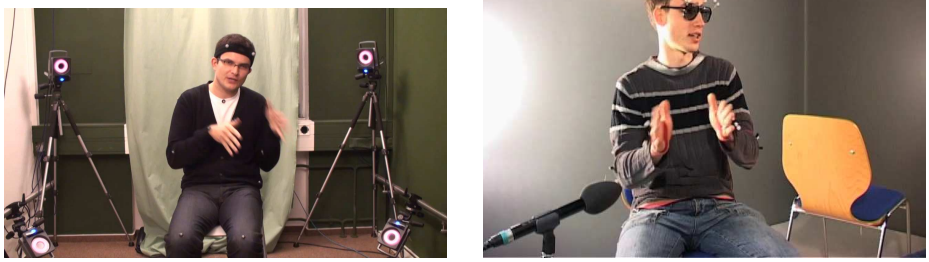


Fig. 8: Automatically generated gesture space based on the human torso and face detections. Here the left hand is located in the *periphery*, while the right hand is located in the *extreme periphery*. Image obtained from the NeuroPeirce corpus.

in several aspects of gesture analysis, we often notice that manual annotation is restricted to a partial analysis. For example, the directionality of an entire leftward pointing gesture is often annotated as *left*, since this is the major direction of movement. Again, this partial analysis arises from the tremendous amount of work that manual, frame-based, analysis requires. To further support the annotation of this type of recordings, we propose an automatic alternative. Here we calculate the direction of movement for each frame by comparing the hand positions of the current frame and the positions in the previous frame. Thereafter we apply a temporal smoothing by using a 1-D convolution filter to reduce jitter on the annotations. Currently our approach automatically annotates four directions: left,right, upwards and downwards for each hand in each frame of a recording. And again, the results of this analysis can exported to an annotation file for further processing in ELAN and other tools such as ANVIL.

As mentioned before, the focus of this paper was on the accurate detection of gesture sequences rather than the recognition of specific gestures. Next to this gesture detection, we also extract relevant features from each gesture sequence such as usage of the gesture space and directionality of the gestures. Since our goal was to develop a tool for simplifying the work of manual annotators, our system needs to achieve high accuracy. Therefore we performed a profound validation of the above-mentioned approaches. In the next section, we present the results of this validation in terms of accuracy, usefulness and cost-effectiveness compared to manual analysis.



(a) Frame of NeuroPeirce corpus (b) Frame of the SaGA corpus

Fig. 9: Example frames of both corpora that we used in our validation.

5 Results

To validate our approach, we searched for existing pre-annotated corpora as a basis for comparison between manual and semi-automatic annotation. We found two independent institutes willing to provide their corpora as well as their annotations. The first corpus, the NeuroPeirce corpus (Brenger and Mittelberg 2015), was created by the research group of Professor Irene Mittelberg (University of Aachen) in the context of a larger research project. Within this project several recordings were made of participants during natural gesturing i.e. non-elicited speech and gesture production. For our validation, we got access to one recording of approximately 7 minutes (10500 frames) as well as the corresponding annotations in ELAN. These annotations include the above-mentioned parameters of gesture phases, position in gesture space, etc. An example frame can be found in the left part of figure 9. The second corpus was created at the university of Bielefeld and is known as *The Bielefeld Speech and Gesture Alignment Corpus* (SaGA) (Lücking et al 2010). Here direction-giving dialogues were recorded using multiple cameras. For our validation we used a recording lasting approximately 8.5 minutes (in total 13000 frames). Again, this recording was annotated in ELAN and includes annotations of the gesture phases. In the right part of figure 9, an example frame of this recording can be found. Since both corpora contain participants are that are sitting, we also evaluated our approach on a self-recorded corpus of a lecture recording in which a speaker gives a PowerPoint presentation. This introduces additional challenges since the speaker if free to move around in front of the presentation screen, making the calculation of the rest positions much more challenging.

We processed each recording separately using the above-mentioned approaches. First we used our semi-automatic hand annotation tool in order to retrieve the hand, face and torso locations in each frame. Some measurements regarding this annotation can be found in table 1. Here we notice that the amount of manual interventions required by the system (based on the predefined threshold) is negligible. In less than 3% of the frames manual annotation was required, i.e a reduction of manual work with a factor of 37 as compared to fully manually annotating

Table 1: Measurements of the semi-automatic hand annotation of the used recordings.

	NeuroPeirce	**SaGA**	**Lecture**
length	7 min	8,5 min	4,8 min
#frames	10500	13000	6700
#manual annotated frames	253	358	120
pct manual annotated frames	2,41%	2,75%	1,7%
processing time candidate generation	40 min	34 min	?
processing time automatic filtering	29 min	26 min	?
amount of manual intervention	8 min	12 min	4 min
total processing time	77 min	72 min	29 min

each frame. The total processing time includes the face and torso detection, the generation of the skin segments, filtering the candidates as well as the manual interventions. Starting from the hand, face and torso detections in each frame, we can apply our gesture analysis tool to both recordings. In the remainder of this section we compare our automatic analysis and the manual annotations. First, we evaluate the accuracy of the semi-automatic hand annotation tool 5.1, after which we present the results of an accuracy measurement of the gesture phase segmentation 5.2. In a third and fourth section, we review the gesture space annotation 5.3 and directionality 5.4. Finally, in section6, we elucidate some limitations of our analysis approach.

5.1 Accuracy of the hand annotations

Since the accuracy of our gesture detection algorithm relies on the accuracy of the semi-automatic hand annotations, it is important to validate their accuracy. For this, we manually labeled the hand positions in the first 1000 frames of the NeuroPeirce recording. Then we compared our semi-automatic hand annotations to this ground-truth in terms of accuracy. A hand annotation is considered valid if the distance between the detection and the manual annotation was below half face width, which is a commonly used measure for hand detection algorithms (Zhang et al 2014). This comparison reveals that in 97% of the hands, the position was obtained correctly using our semi-automatic approach. Furthermore, for the entire set of 2000 annotated hand positions, the average distance was only 10 pixels, which indicates that our approach is indeed highly accurate and can be used as a basis of our gesture analysis tool.

5.2 Accuracy of gesture phase segmentation

To validate our gesture phase segmentation, we propose a frame-based comparison between our automatically generated gesture sequences and the manual annotations of both recordings. Based on the manual gesture phase annotations in ELAN, we assigned a label to each frame: **1** if the frame was part of an annotated gesture phase, **0** otherwise. The same frame-based information was extracted from our automatic gesture phase segmentation. Finally a validation scheme as illustrated in figure 10 was used. For each frame we compare both manual and automatic

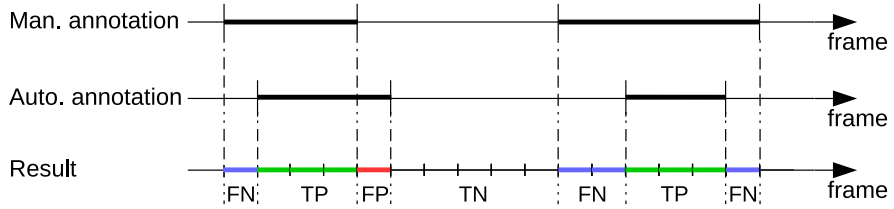


Fig. 10: Validation methodology that is used for the gesture phase segmentation.

Table 2: Accuracy of the optimal working point i.e. $\alpha = 0.9$. Bottom row of the table shows the best accuracy of the AUVIS tool tested on the same corpora.

	NeuroPeirce	**SaGA**
Precision	98,85%	94,38
Recall	83,41%	80,22
F_1**-score**	90,48%	86,73%
Best F_1**-score AUVIS**	82.23%	62.17%

annotations, resulting in one out of four labels per frame: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Using these labels, we can determine the accuracy of our approach as shown in equation 3. The precision (P) is the fraction of retrieved instances that are relevant, while the recall (R) is the fraction of relevant instances that are retrieved. By combining them into the F1-score (F_1), we obtain a single value that expresses the accuracy of our system, which is a commonly used approach for measuring the accuracy of a computer vision algorithm.

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \\ F_1 &= 2\frac{PR}{P+R} \end{aligned} \tag{3}$$

In equation 2, we already introduced the tuning parameter α. This parameter is used to find the optimal fraction of the σ thresholds. For validation, we varied α in the range from 0 up to 3 in steps of 0.1 in order to find the optimal setting. The results of these experiments are shown in figure 11, in which we plot precision versus recall. The most optimal point on such a graph is the upper-right corner (both P and R equal to 1). It is clear that the curves of both recordings approach this point. Both curves reach their best accuracy at the same α: 0.9. The corresponding accuracy measurements for this α are given in table 2. Here, we notice that our approach achieves very high accuracy on both recordings.

Next to the validation of our own approach, we also compared our accuracy against another gesture analysis algorithm. We opted to use the AUVIS gesture analysis tool as presented by Schreer and Masneri (2014), since a) it formulates the same goal and b) it is directly available in the ELAN annotation software

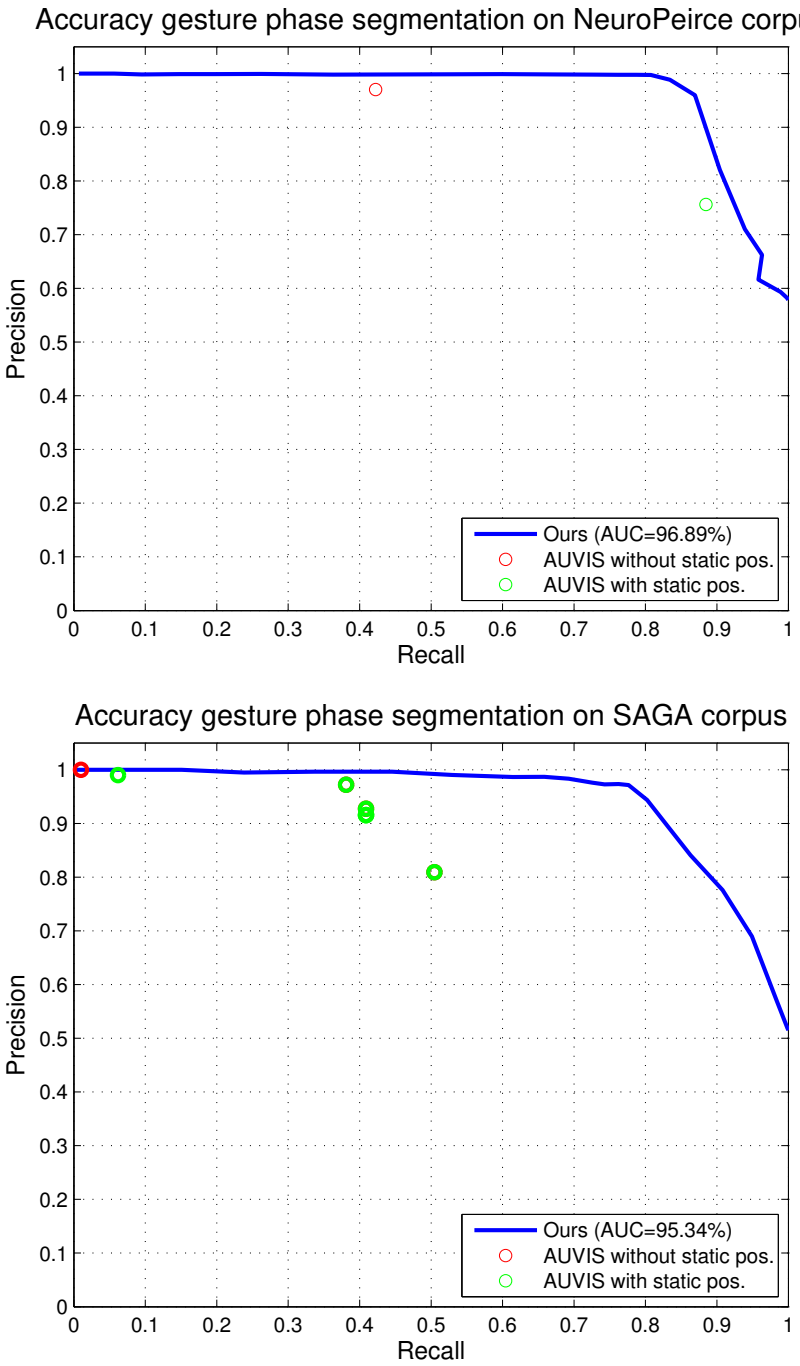


Fig. 11: Precision-recall curves of our approach for both recordings. Colored circles represent the accuracy of the Schreer and Masneri (2014) method.

(the latter experiment was performed in ELAN version 4.9.4). For more information regarding this integration, we refer to[2]. As mentioned in section 2.3, their approach relies on manual tuning of the skin-segmentation parameters. We asked 5 participants this tuning for each recording in order to provide a fair comparison, both experienced and non-experienced annotators. In this publicly available implementation of their approach, two methods for gesture segmentation are available: taking the distance to the static position into account or not using the static position and only relying on the velocity of the hands. The results of their approach are shown in figure 11 using the colored circles. These circles reveal that the skin segmentation of the NeuroPeirce recording was rather easy since the same accuracy was achieved for each skin segmentation setting. On the other hand, the skin segmentation of the SaGA recording was far more complex as shown by the diverse accuracy results. This was mainly caused by the presence of the wooden chair in each image as shown in figure 9, which has more or less skin tone. In table 2, the best F_1-score of the AUVIS approach is given for both corpora. Overall it is clear that our approach outperforms the method of Schreer and Masneri (2014). Furthermore our approach results in more consistent accuracy over multiple recordings without time-consuming and subjective parameter tuning.

Since there were no independent annotations of the gesture sequences were available of our own-recorded corpus, another validation technique was used. First, our gesture segmentation approach identified the segments in which the speaker is gesturing based on the retrieved information (i.e. face, upper body and hand locations). In total, our approach found 91 gesture sequences in this recording.

For validation of our gesture segmentation approach, we asked an independent annotator to manually assign a label to each extracted segment. The annotator could choose between either *gesture* or *non-gesture*. Then, we compared the automatically generated annotations to the manually assigned labels to measure the reliability of our gesture segmentation approach. As shown in the leftmost columns of table 3, the reliability of our approach is again satisfying, although the overall score is somewhat lower as compared to measurements of the previous experiments.

Manual inspection of this result revealed that the majority of disagreements occur in short gesture sequences. Indeed, it might happen that the Kalman filter of at least one hand floats away before our hand detection approach requests manual intervention, resulting in false positives. Furthermore, as shown in figure 12, there are some translations in the upper body detections. These are mainly caused by the deformable aspect of the model that we use. Since the relative hand positions are calculated using the center of the upper body detection, their position is affected by these translations. As a result, it might happen that the distance between the rest position and the relative position of some hands exceeds the threshold, causing an erroneous gesture sequence. Besides these translations, we also noticed slight scale variations in the upper body detections, which also affect the obtained hand positions.

As an additional validation step, we removed the gesture sequences which have a duration less than 500 ms, which is indeed relatively short for a gesture, and we repeated the reliability measurements. The rightmost column of table 3 shows the improved results of this additional validation. It is clear that the lower reliability

[2] `https://tla.mpi.nl/projects_info/auvis/`

(a) (b)

Fig. 12: Variations in upper body detections that may cause changes in relative hand positions.

Table 3: Reliability of gesture analysis.

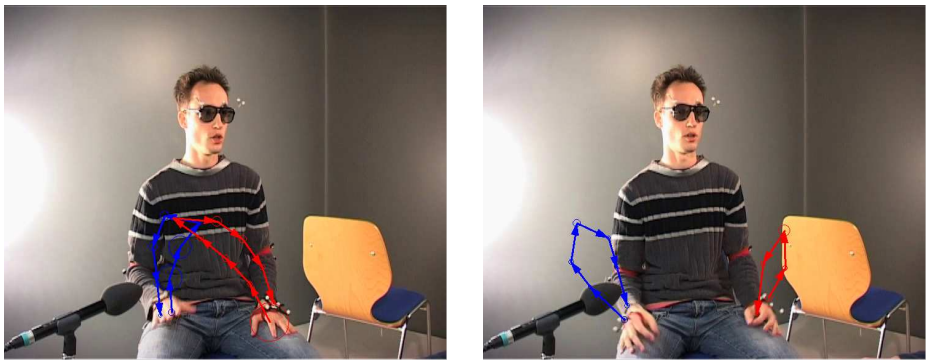
	Level	**Level without short segments**
Agreement	79.1%	87.5%
Scott's Pi	78.7%	87.3%
Cohen's Kappa	78.7%	87.3%
Krippendorff's Alpha	78.7%	87.4%

scores in the left part of this table are indeed mainly caused by the shorter gesture segments.

5.3 Accuracy of gesture space annotation

The validation of the gesture space is far more complex since the available annotations are inadequate. As mentioned before, the majority of existing gesture space annotations cover only a small portion of the data. Indeed, the annotation of the gesture space in the NeuroPeirce recording was restricted to a single label for each gesture stroke, whereas our system provides a frame-based gesture space annotation for an entire gesture phase (preparation, stroke and retraction). This imbalance in level of precision made it difficult to perform a meaningful comparison between the manual annotation and our automatic labeling. The SaGA recording did not include annotations of the gesture space at all.

Since our gesture space analysis provides a frame-based annotation, we needed to transform this output for validation. We extracted the frame sequences from each gesture stroke and calculated the most occurring label in each stroke. Then, we compared these extracted labels to the manual annotations resulting in an accuracy of 64.42%. Again, this comparison is suboptimal since our automatic



(a) (b)

Fig. 13: Examples of two gestures captured into a single image.

analysis produces a much more fine-grained annotation compared to the manual annotations.

Since there is a mathematical relation between the hand positions and the gesture sectors, one can assume that if the positions of the hands are obtained highly accurately, our gesture space analysis is likewise accurate. As mentioned above, the accuracy of our semi-automatic hand annotation tool is 97%, thus we might suppose our gesture segmentation is as accurate as well.

5.4 Output of gesture directionality

In this last subsection, we discuss the analysis results of the gesture directionality. As mentioned in subsection 4.4, our system defines the direction of each hand in each frame. Comparing our automatic direction labels to the manual annotations was impossible since none the corpora provided such annotations, since it is practically unfeasible to perform this type of annotations manually. Nevertheless, we might assume that our directionality analysis is highly accurate since it is directly extracted form the semi-automatic hand annotations.

A final advantage of this automatic frame-based analysis is the possibility to represent an entire gesture into a single frame. Examples are given in figure 13. Here we show the first frame of each gesture sequence and we plot a circle for the individual hand positions of the entire gesture onto this frame. The displacement between two frames is illustrated using arrows. In case a hand was held still, we indicate this by increasing the radius of the corresponding circle. Such a representation of each gesture can be used in a graphical representation of a recording, where for example the gesture images represent interesting moments on a time line.

At last, it is worth to mention that the total processing of our entire gesture processing algorithm lasts only a few minutes.

6 Conclusion and future work

In this paper we presented a semi-automatic approach for the annotation of gestures that are relevant in research on human-human interaction e.g. in the field of psychology, linguistic or behavior analysis. Our focus lies on minimizing the workload that is related to this type of annotations. To provide a useful alternative, it is of vital importance that our approach produces highly accurate annotations. Therefore, our first step is the extraction of relevant body parts including face, human torso and hands in each image of the recording. This is achieved using our own semi-automatic hand detection algorithm in which manual intervention is required in case if and only if the confidence of a hand detection is too low. Several validation experiments revealed that this method is capable of detecting the hands highly accurate, which makes it a valid basis for the gesture analysis. On top of that we are able to reduce the amount of manual analysis by a factor of 37 as compared to fully manually annotating each frame. The gesture analysis starts by defining the rest position of each hand during an entire recording. Once this location is known, we calculate for each frame the distance between each hand and its respective rest position. Based on this displacement, we are able to segment a recording in gesture and non-gesture segments. On top of that, we automatically analyze the usage of the gesture space according to the McNeill (1992) sectors. A final analysis is done on the directionality of the hands. Here, we analyze the trajectory of each hand during gesturing. Each analysis generates automatically a unique tier in an XML compatible file, making our approach integratable with existing annotations. We performed a thorough comparison between our automatic gesture analysis and the annotations of two existing corpora revealing that our approach is highly accurate.

Although our analysis framework has proven to a valuable alternative for the traditional, manual and time-consuming, annotation of hand position, hand motion, gesture type, etc, there is room for improvement and future development. As mentioned in section 5.2, we noticed that the upper body detections are sometimes not perfectly aligned with the persons that are presented in the images. As a result, it might happen that some gesture segments are created erroneously. This issue can be solved by tracking the detection scale at which an upper body is detected. By doing so, one could limit the search space and therefore remove the fluctuations in the size of the detection windows. Furthermore, a more advanced tracking may overcome the translation issues.

Another straightforward enhancement consists of expanding our hand and gesture detection approach to multiple persons. Currently, our approach is developed to detect the hands of only a single person. The ability to analyse the gestures of multiple persons is relevant in for example the analysis of multi-party interactions. Furthermore, integrating a person re-identification step in the gesture analysis approach is inevitably linked with this expansion. Such an integration will bring the analysis of e.g. triadic conversations to a next level by automatically analysing the gestures of each participant as well. Building on our gesture detection approach, a gesture recognition approach can be developed allowing for a fine-grained gesture analysis. The automatic recognition of several basic gesture patterns, such as pointing or batons, is relevant in various application domains including research on gestural behaviour and research on sign language for the automatic subtitling of singers. Finally, our gesture analysis approach could be expanded as well. Cur-

rently we define a single rest position for each hand, but evidently it might occur that the rest position of the speaker changes during the experiment. Therefore, our approach can be modified by searching for multiple rest positions, based on the density of the relative hand positions. Another modification that could improve the accuracy of the gesture segmentation approach is combining the displacement to the resting position with the velocity of the hands. Such an integration would allow for a more accurate and finer segmentation, and therefore an accurate identification of a hold phase in a pointing gesture would be possible. Furthermore, it would be interesting to measure the influence of multiple annotators on the manual input of our semi-automatic analysis. We do expect that our system is highly robust since manual intervention is only required sporadically. Therefore, the manual annotation is no longer a repetitive task, reducing the chance of erroneous manual annotations.

Acknowledgements We would like to thank Prof. Irene Mittelberg and her research group for providing the annotations on the NeuroPeirce dataset (Brenger and Mittelberg 2015) as well as the authors of the SaGA dataset (Lücking et al 2010). The availability of the annotations was crucial in our validation process.

References

Abuczki Á, Esfandiari BG (2013) An overview of multimodal corpora, annotation tools and schemes. Argumentum 9:86–98

Alon J, Athitsos V, Yuan Q, Sclaroff S (2009) A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(9):1685–1699

Badi H (2016) Recent methods in vision-based hand gesture recognition. International Journal of Data Science and Analytics pp 1–11

Bennewitz M, Axenbeck T, Behnke S, Burgard W (2008) Robust recognition of complex gestures for natural human-robot interaction. In: Proceedings of the Workshop on Interactive Robot Learning at Robotics: Science and Systems Conference (RSS)

Brenger B, Mittelberg I (2015) Shakes, nods and tilts. motion-capture data profiles of speakers and listeners head gestures. In: Proceedings of the 3rd Gesture and Speech in Interaction (GESPIN) Conference, pp 43–48

Bressem J (2013) Transcription systems for gestures, speech, prosody, postures, and gaze. In: Proceedings of Body - Language - Communication: An International Handbook on Multimodality in Human Interactiong, vol 1, pp 1037–1059

Chang JY (2015) Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I, chap Nonparametric Gesture Labeling from Multimodal Data

De Beugher S, Brône G, Goedemé T (2016) Semi-automatic hand annotation making human-human interaction analysis fast and accurate. In: Proceedings of Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pp 552–559

Escalera S, Baró X, Gonzàlez J, Bautista MA, Madadi M, Reyes M, Ponce-López V, Escalante HJ, Shotton J, Guyon I (2015) Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I, Springer International Publishing, Cham, chap ChaLearn Looking at People Challenge 2014: Dataset and Results, pp 459–473

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1627–1645

Ferrari V, Marin-Jimenez M, Zisserman A (2009) Pose search: Retrieving people using their pose. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–8

Feyaerts K, Brône G, Oben B (2016) Multimodality in interaction

Gebre BG, Wittenburg P, Lenkiewicz P (2012) Towards automatic gesture stroke detection. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), Istanbul, Turkey

Jewitt (2009) The Routledge handbook of multimodal analysis. London; New York

Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vision 46(1):81–96

Kalman R (1960) A new approach to linear filtering and prediction problems 82:35–45

Kapuscinski T, Oszust M, Wysocki M, Warchol D (2015) Recognition of hand gestures observed by depth cameras. International Journal of Advanced Robotic Systems 12

Karlinsky L, Dinerstein M, Harari D, Ullman S (2010) The chains model for detecting parts by their context. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 25–32

Kipp M, Martin J, Paggio P, Heylen D (eds) (2009) Multimodal Corpora - From Models of Natural Interaction to Systems and Applications, Lecture Notes in Computer Science, vol 5509. Springer

Kuznetsova A, Leal-Taixe L, Rosenhahn B (2013) Real-time sign language recognition using a consumer depth camera. In: Procedings of The IEEE International Conference on Computer Vision (ICCV) Workshops

Lücking A, Bergmann K, Hahn F, Kopp S, Rieser H (2010) The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In: Kipp M, Martin JP, Paggio P, Heylen D (eds) Proceedings of LREC 2010 Workshop: Multimodal CorporaAdvances in Capturing, Coding and Analyzing Multimodality, pp 92–98

Marcos-Ramiro A, Pizarro-Perez D, Marron-Romera M, Nguyen LS, Gatica-Perez D (2013) Body communicative cue extraction for conversational analysis. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition

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

Mittal A, Zisserman A, Torr P (2011) Hand detection using multiple proposals. In: Proceedings of BMVC, BMVA Press, pp 75.1–75.11

Monnier C, German S, Ost A (2015) Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I, chap A Multi-scale Boosted Detector for Efficient and Robust Gesture Recognition

Müller C, Cienki A, Fricke E, Ladewig S, McNeill D (2013) Body - Language - Communication: An International Handbook on Multimodality in Human Interaction, Body - language - communication : an international handbook on multimodality in human interaction, vol 1. De Gruyter Mouton

Müller C, Cienki A, Fricke E, Ladewig S, McNeill D (2014) Body - Language - Communication: An International Handbook on Multimodality in Human Interaction, Body - language - communication : an international handbook on multimodality in human interaction, vol 2. De Gruyter Mouton

Neverova N, Wolf C, WTaylor G, Nebout F (2014) Multi-scale deep learning for gesture detection and localization. In: Proceedings of ECCV ChaLearn Workshop on Looking at People

Peng X, Wang L, Cai Z, Qiao Y (2015) Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I, Springer International Publishing, Cham, chap Action and Gesture Temporal Spotting with Super Vector Representation, pp 518–527

Rahim NAA, Kit CW, See J (2006) RGB-H-CbCr skin colour model for human face detection. In: Proceedings of M2USIC, Petaling Jaya, Malaysia

Rautaray SS, Agrawal A (2012) Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review 43(1):1–54

Schreer O, Masneri S (2014) Automatic video analysis for annotation of human body motion in humanities research. In: International Workshop on Multimodal Corpora in conjunction with 9th edition of the Language Resources and Evaluation Conference (LREC), pp 29–32

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 511–518

Yin Y, Davis R (2013) Gesture spotting and recognition using salience detection and concatenated hidden markov models. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI), ACM, New York, NY, USA, pp 489–494

Zhang Z, Conly C, Athitsos V (2014) Hand detection on sign language videos. In: Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA), ACM, pp 26:1–26:5