Exploration of webcam-based eye tracking by comparing gaze behaviour of Dutch Sign Language users during interpretation of sign language

Mark Wessel

Bachelor Artificial Intelligence

Faculty of Humanities, Utrecht University

July 2, 2021

Mark Wessel
m.wessel@students.uu.nl
Student ID: 6008283

Supervisor: Jessica Heeman, MSc.
2nd review: Dr. Chris Paffen

Abstract

The present study investigates properties of webcam-based eye tracking for online experiments using the WebGazer JavaScript library implemented in Gorilla Online Experiment Builder. Participants (n=15) completed a calibration sequence to optimize the eye tracking quality of their webcam, after which they were they were tasked to attentively watch 9 video's in which a story was signed in Dutch Sign Language. Between video 4 and 5 calibration was repeated. The main research question is: what are the accuracy and precision of a webcam-based eye tracker for online experiments? The accuracies (reported in degrees of visual angle) of the webcams averaged 3.65° (SD = 0.77°) and the precisions (reported in inter-sample distance root mean square (RMS)) averaged 0.89° RMS (SD = 0.34° RMS). Furthermore, the data was of sufficient quality to filter out typical properties of gaze behaviour such as fixations, smooth pursuit eye movements and blinks. These results show promise that online experiments with a webcam-based eye tracker can be a viable alternative to lab-based experiments with classic infrared eye trackers in cases where very accurate data quality is not required. The eye tracking data drifted during the experiment causing a loss in accuracy. Attempts to improve accuracy by post-hoc compensation of the drift were unsuccessful. The author proposes further research attempts to determine the origin of this drift and possible remedies.

Keywords: webcam-based eye tracking, Dutch Sign Language, accuracy, precision

Introduction

Eye tracking is the act of estimating where a person is looking based on the position of their eyes. This information can be used to study human overt visual attention as first shown by Yarbus (1967). It is used in a range of departments, e.g. psychology (Lai et al., 2013), computing science (Sharafi et al., 2015), medicine (Ashraf et al., 2018), advertisement (Santos et al., 2015), and aviation (Peißl et al., 2018).

Eye tracking can be executed with different techniques. One of the most widely-used techniques is video eye tracking using feature-based gaze estimation. With this technique, a camera detects the position and direction of the eyes, and maps this to a location on the screen. This detection is based on the position and shape of various features of the eye such as the iris, the cornea and the pupil. Then, the gaze position is estimated using one of two approaches: 1) model-based gaze estimation, or 2) interpolation-based gaze estimation. Eye trackers using model-based gaze estimation compute a 3D model of the eye using the detected features, and then calculate the intersection between that model and the screen. Eye trackers using interpolation-based gaze estimation determine a relationship between the features of the eye and a known gaze position on the screen (e.g. a neural network linking a state of the eye features to the position of a fixation cross). (Hansen & Ji, 2010).

Most video-based eye trackers use active lighting, such as infrared (IR) light, to illuminate the eyes. The wavelength of IR light is invisible to the human eye, thus the eye can be sufficiently illuminated without interfering with an experiment. For feature-based gaze estimation, illumination of the eye is necessary to create an image with high contrast between the features of the eyes to detect those features using computer vision (Hansen & Ji, 2010).

When implementing an eye tracker in a study, multiple properties of the eye tracker must be taken into account. The most important properties are the setup required (i.e. number of cameras, number of lights, possibly a head mount), whether the setup needs to be calibrated, and the accuracy and precision of the eye tracker (Hansen & Ji, 2010). The accuracy of the eye tracker is defined as the distance between the participant's actual gaze location and the estimated gaze location. Accuracy is expressed in degrees visual angle. Precision of the eye tracker is defined as the spread between multiple recordings of the same gaze location (i.e. when recording a stationary artificial eye with perfect precision, all gaze recordings will be at exactly the same position) (Holmqvist et al., 2011). Precision is expressed as the root mean square (RMS) of the distances between gaze estimations. A comparison between good- and poor accuracy and precision is shown in figure 1.

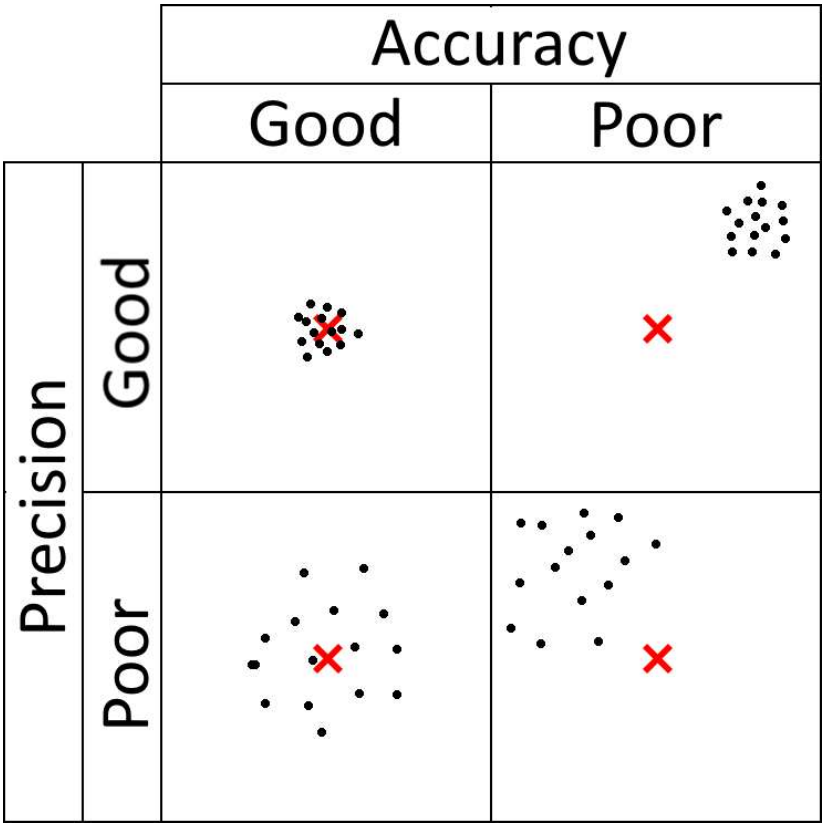

Figure 1. An overview of good and poor accuracy and precision. The red cross represents the fixation point for the participant, and the black points represent gaze estimations.

Multiple challenges occur when attempting to implement eye tracking in a study. For one, a choice must be made between high cost and high accuracy. For example, model-based gaze estimation is more accurate than interpolation-based gaze estimation, but two cameras are needed to gain 3D vision of the participant and compute a 3D model of the eyes. On the other hand, interpolation-based gaze estimation needs only one camera to detect a 2D image of the eyes, but a higher quality camera will detect a higher quality image thus enabling a better relationship between the features of the eye and the screen. Furthermore, additional specific hardware is required by most eye trackers, such as 1 or more IR light(s), additional cameras to detect the tilt and pan of the participant's head, or a head mount. These additional hardware requirements make it difficult to use eye tracking outside lab-based environments. The difficulty to use eye tracking outside of lab-based environments is further enhanced by IR light being unreliable in situations where the lighting conditions are suboptimal (e.g. outdoors) (Hansen & Ji, 2010). Lastly, a choice of sampling frequency of the eye tracker must be made. Sampling frequency is the amount of times per second that the eye tracker records the features of the eyes. An eye tracker with a sampling rate of 100Hz will make 100 recordings per second, i.e. 1000/100 = 10ms between recordings. The higher the sampling rate, the higher the cost of the hardware. When trying to record the duration of gaze events such as fixations and saccades, an inverse relationship exists between the sample rate of the eye tracker and the amount of datapoints required for a given variance. The higher the sampling rate, the less datapoints are necessary (Andersson et al., 2010). Saccades take 30-40ms, and the lower the saccadic distance is, the higher is the required sampling rate to record the peak velocity of that saccade (Holmqvist et al., 2011). Micro saccades are 50 times smaller than saccades (Engbert, 2006) and thus require even higher sampling rates to record. Studies that investigate

micro saccades always use an eye tracker with a sampling rate of at least 200Hz (Holmqvist et al., 2011).

In recent years, attempts have been made to minimize the challenges to implementing eye tracking and to make the use of eye tracking in studies more accessible. One possible attempt is the use of more commercially available hardware such as an ordinary consumer-grade webcam. Although relatively little work has been done to explore the quality of eye tracking using webcam-based eye trackers, it has gained more attention in recent years. For example, Burton et al. (2014) compared data from an experiment taken with both an IR video-based eye tracker and a webcam-based eye tracker to show that webcam-based eye tracking performs similarly to a classic video-based eye tracker when presented with larger images on fixation tasks, although the video-based eye tracker outperformed the webcam-based eye tracker in cases with smaller images or images further from the centre of the screen.

Papoutsaki et al. (2016b) made eye tracking less hardware-specific by developing a JavaScript-based eye tracking library, named WebGazer, which can be used with commercial-grade webcams. WebGazer is a feature-based eye tracker using interpolation-based gaze estimation. WebGazer establishes a relationship between a person's features of their eyes and their gaze position is established via interactions the person makes on a webpage. When the person interacts with an object (e.g. clicks a button), WebGazer assumes the person is looking at that object, and then links the state of the features of the person's eyes to that position on the screen. Semmelmann & Weigelt (2018) carried out an experiment both online and in-lab, using the same webcam-based eye tracking hardware in both cases, and compared the data to show that data from the online experiment was only slightly less accurate than data from the in-lab experiment. The findings of Semmelmann & Weigelt (2018) indicate that online experiments would become a

viable option with webcam-based eye tracking, and defeat the challenge of being tied to in-lab settings.

In the present study, I aim to build on the WebGazer library of Papoutsaki et al. (2016) to explore webcam-based eye tracking by quantitatively and qualitatively illustrating multiple properties of webcam-based eye tracking using WebGazer. I will answer the following main research question: 'what are the accuracy and precision of a webcam-based eye tracker for online experiments?'. I hypothesize that webcam-based eye trackers will have considerably worse accuracy and precision than typical video-based eye trackers used in research, but that the precision will suffice to detect fixations and saccades. Moreover, given the hypothesis holds, I will try to improve the accuracy of the webcam-based eye tracker by introducing validations of the relationship between eye features and gaze location between trials. I hypothesize that the webcam-based eye tracker predictions will drift over time and that interim validations can be used to compensate for this drift to keep consistent accuracy. Finally, I will determine whether the webcam-based eye tracker detects participants' blinks and if these blinks can be filtered. I hypothesize that the eye tracker will not detect blinks and therefore that it will not be possible to filter them from the data.

To determine the answer to my research questions the current experiment was embedded in an experiment investigation the differences in gaze behaviour between Beginner and Expert users of Dutch Sign Language (DSL) in collaboration with Mick Richters, Puck Rutten & Bjorn van Vliet (Richters, Rutten, van Vliet, 2021). The entire code of the experiment can be found in appendix A.

DSL is a visual language primarily used by the deaf community of the Netherlands. At the core of the deaf community stand people who grew up using sign language and use it as their

primary language (albeit not necessarily learnt from their parents). However, multiple groups of people can also be included in the deaf community, such as hearing people growing up in a deaf family and deaf/hard of hearing people with cochlear implants who still were exposed to sign language (e.g. by their deaf family) while growing up (Crasborn, 2016).

A person's gaze behaviour is typically related to information that person is processing. For example, when reading a text longer fixations on phrases indicate that the person is having difficulty processing that phrase . Likewise, beginner- and native users of American Sign Language (ASL), while interpreting a story signed in ASL, fixate on different regions of the signer. (Emmorey et al., 2009)

Emmorey et al. (2009) had participants interpret a story signed by a live native ASL signer while wearing a head-mounted eye tracker. This resulted in videos of the participants' fields of view, which were post-hoc labelled with regions of interest (ROIs) of the signer's upper face, eyes, mouth, lower face and hands.

In the present study, we had participants interpret videos containing stories signed in DSL by a native signer. Our videos contained no distractors, and participants are instructed to pay attention, therefore we can assume participants are focusing on the signer during the videos. Given this assumption, we mimicked Emmorey et al. (2009) by post-hoc labelling the videos with ROIs twice. Once with large ROIs (face, hands) and once with small ROIs (eyes, mouth). The decision to identify both small and large ROIs is based on the fact that we do not know the precision of the webcam-based eye tracker, and thus if it is possible to differentiate between small ROIs. Because of the assumption that the participant is focused on the signer, we sized the ROIs according to the maximum region that a given region (e.g. hands, eyes) of the signer moved in (figure 2). Therefore,

when a gaze estimation point is in a given ROI we can assume the participant was looking at the corresponding part of the signer.

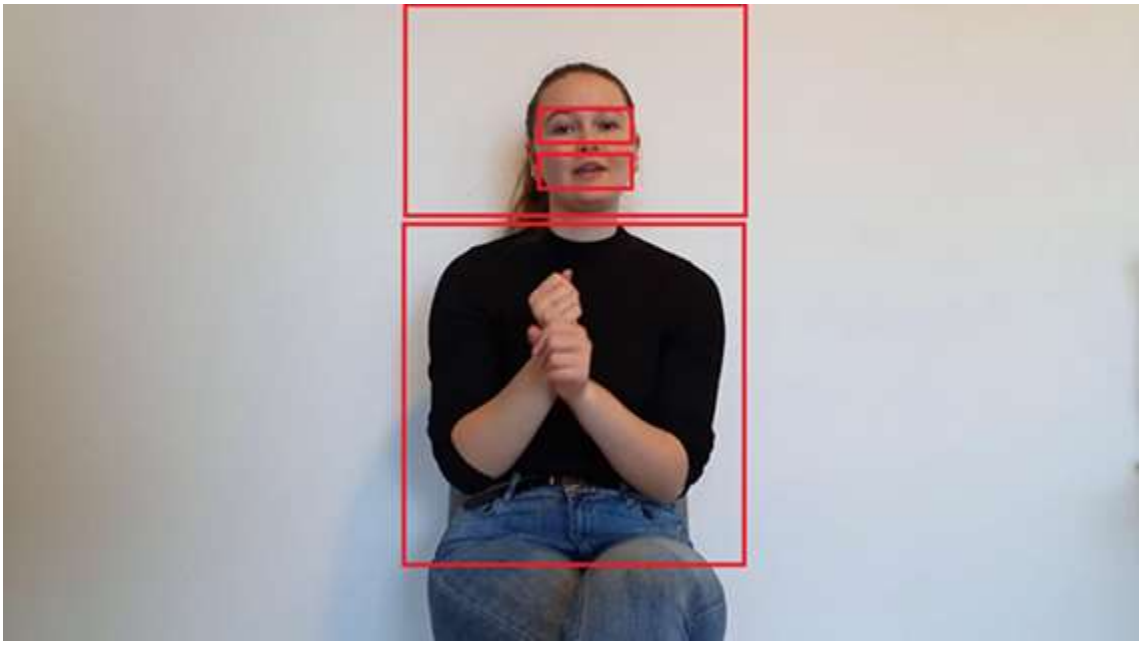

Figure 2. An example frame from a video of the present study. The red regions represent the ROIs for the hands, face, eyes and mouth. The ROIs are not visible to participants during the experiment.

One of the largest challenges for the Dutch deaf community has been the lack of recognition of DSL. In 1880 sign language was banned in education. Deaf children in schools were to be taught using oral methods, i.e. speaking and lip-reading, a movement called oralism. When a child was caught using sign language, punitive measures were taken. Naturally, people did not abandon their use of sign language, instead switching to only using sign language in secret, resulting in different dialects. (Morgans, 2008).

In 1980 Stokoe (1980) published an in-depth analysis on sign language, arguing that sign language is processed the same way as spoken language and that prohibiting deaf children from learning sign language had a negative effect on their development. This publication led to the nuancing of the general views on sign language, and members of the dead community started

working on greater general recognition of DSL, however the Dutch government did not yet recognize DSL as an official language.

To reach official recognition of the language, the Gebarencentrum was established in 1996. Gebarencentrum is an organization that works on standardizing DSL by developing a national dictionary containing standardized signs for words. These signs are picked from variants from one of the schools mentioned earlier. This standardisation of DSL serves to allow education material for the deaf community to be standardised like the education material on hearing schools (Doof, n.d.).

It was only recently, the 13th of October 2020, that DSL was officially recognized by the government of The Netherlands. Official recognition gives members of the deaf community the right to be presented information in DSL. Moreover, the government is now responsible for upholding this right. (Gebarencentrum, n.d.)

Because of the long repression the Dutch deaf community, the community is very isolated (N. Fluitman, personal communication, March 23rd, 2021). The embedding of the present study in the investigation of the differences in gaze behaviour between Beginner and Expert users of Dutch Sign Language (DSL) allows for insight into the way DSL users interpret DSL, and thus insight into potential development strategies for software to bridge the gap between the Dutch deaf- and hearing communities (e.g. artificial intelligence to translate DSL into spoken Dutch.)

Methods

Participants

Participants included 15 adults (3 male, 12 female) between 20 and 49 years old. All participants were either native DSL signers, a certified interpreter or teacher in DSL, or studying DSL at the University of Applied Sciences Utrecht or the University of Amsterdam. All participants were fluent in Dutch (spoken and written) or in DSL.

Participants were recruited on social media platforms such as Facebook or LinkedIn, and through word of mouth in the DSL community. Participants could take part in the experiment by clicking on a link. Every participant was automatically assigned a random ID upon clicking on the link, therefore the link was generic and could be reused by every participant. Participants were post-hoc classified as beginner or expert based on their answers on the questionnaire prior to participating in the experiment. Participants were asked to self-assess their DSL skill level from a list of options containing multiple identifiers (i.e., "first year DSL student", "DSL minor student", "native signer") (see appendix C). Participants were offered to enter a raffle for 3 €10,- gift cards as compensation for their time in the 20 minute study. Ethical approval was obtained from the Ethics Review Board of the Faculty of Social & Behavioural Sciences of Utrecht University before recruiting participants.

Materials

Participants used their own hardware (i.e. their own computer/laptop, webcam, and in case of a computer their own monitor). Due to software limitations on certain browsers (e.g. the webcam not working properly), participants were limited to the browsers Google Chrome, Microsoft Edge and Mozilla Firefox (any version).

The experiment is comprised of 9 trials. A trial consisted of a fixation cross, a video and two questions. Each video contained one short story signed in DSL by a native signer. The signer wore neutral dark clothing, and sat in front of a white background to eliminate any possible distractors. The trials were preceded by an introduction video that served as a practice trial for the participant in order to get used to the format of the trials. The remaining 9 videos each contained a short story about different topics. The videos were equally split into three difficulty groups based on the Common European Framework of Reference. The first three videos were of the level A1-A2, the fourth through sixth videos were of the level B1-B2 and the last three videos were of the level C1-C2. The difficulties of the topics of the videos were determined based on the level of the

videos. For example, an A1-A2 video consisted of a story with simple vocabulary about going to the zoo, while a C1-C2 video contained a news item with more extensive vocabulary about political hardships at a Dutch embassy.

After each video participants had to answer two multiple-choice questions about the topic of that video. The participants were told to answer the questions as accurately as possible to stimulate maximum attention when watching the video. Every question had four options, one of which was always “I don’t know”. Because the videos contained no distractors and the participants were stimulated to pay maximum attention while watching the videos, we can assume the participants are looking at the signer while watching a video.

The experiment was hosted on the Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). The experiment was programmed in JavaScript using the JsPsych library (de Leeuw, 2015). Some plugins were modified for the use-case of this experiment, for example some hardcoded instructions in English were changed to hardcoded instructions in Dutch.

Eye tracking was carried out using the JsPsych modified version of WebGazer (Papoutsaki et al., 2016b). WebGazer is a client-side eye tracking library written in JavaScript that makes use of consumer grade webcams to track gaze behaviour of participants. WebGazer uses cursor-gaze relationships to self-calibrate, by assuming a person is typically looking at the places of cursor interactions. While the original article of WebGazer pointed out that initial calibration is not necessary due to WebGazer’s ability to continuously self-calibrate using the participant’s cursor clicks, in this study an initial calibration without continuous self-calibration was used instead to ensure the data between videos remained constant.

WebGazer combines three facial feature detection libraries to detect the face and eyes, namely js-objectdetect (Tschirsich, 2012) and tracking.js (Lundgren et al., 2014) to detect the face and eyes and encapsulate those into rectangles, and clmtrackr (Mathias, 2014) to fine-tune the detection. When the eye region is known, WebGazer detects the pupil by looking for the highest contrast location in the eye region. Each eye is turned into a 6x10 pixel image which represents that eye. These 2 images then undergo grayscaling and histogram equalization in order to increase the contrast. The two processed images are then turned into a feature set (i.e. a 120-dimensional vector) representing both eyes. This feature set is then input into a series of linear regression algorithms to first determine the x and y position of the eyes and then map these locations to pixels on the screen. The algorithms assume that a person is fixating on the click locations.

Procedure

All participants were informed that they would see several fragments in which a story in DSL would be signed, and that every fragment would be followed by two questions about the content of the fragment. The participants were told to answer the questions as accurately as possible, and to be honest if they did not know the answer. Participants were informed their eye movements would be recorded using their webcam, and that the experiment would take about 20 minutes.

After consenting to take part in the study, participants were asked to answer short questionnaire for demographic purposes (see appendix C). These questions give an indication of the participant's DSL level (N. Fluitman, personal communication, March 23rd, 2021). Participants were asked to self-asses their skill level according to several categories. The levels were "first year DSL student" through "fourth year DSL student", "minor DSL student", "interpreter", "teacher" and "native signer". These questions served to assign the participant into the beginner or expert group. Lastly, the participants were asked whether they were wearing glasses. The answer to this question may explain abnormalities in eye tracking accuracy.

When the questionnaire was completed the participants were instructed to complete a calibration sequence consisting of five steps: 1) positioning of the participant's face in front of the webcam; 2) determining the size of the participant's screen; 3) determining the distance between the participant and their screen; 4) calibrating the eye tracker; 5) validating the calibration. A visual overview of the entire calibration procedure can be seen in figure 3.

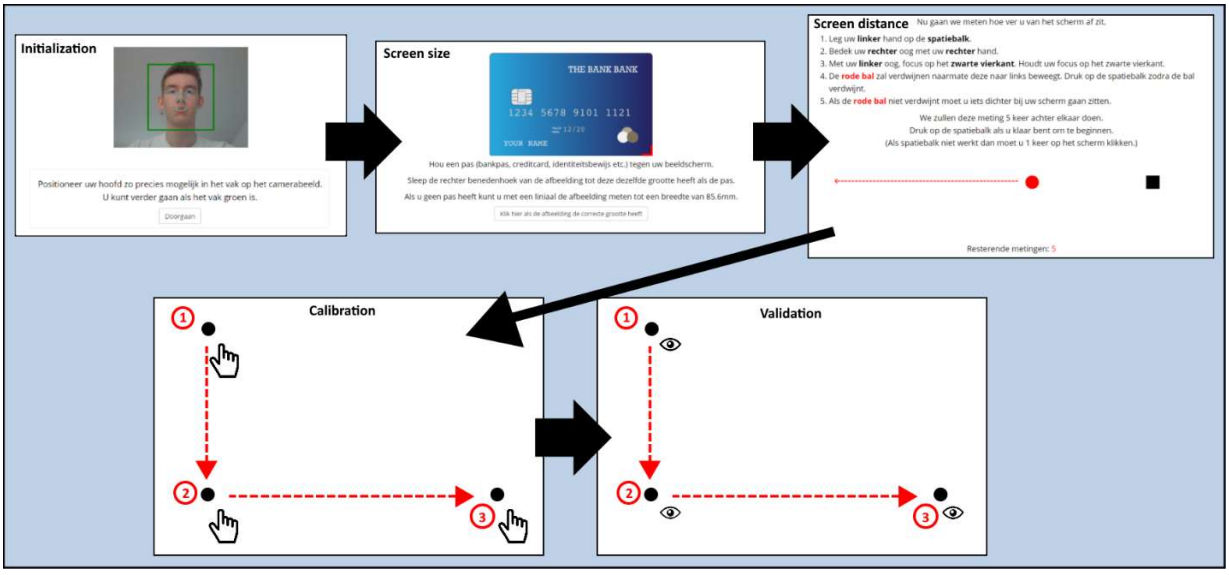

Figure 3. An overview of the calibration procedure of the present experiment

The first step initialized the webcam. Participants were shown their camera feed and informed to position their head in the middle of the visible square. This positioning ensures that the webcam has the best possible view of the eyes of the participant, and encourages the participant to sit still during the trials.

The second step measured the participant's screen size. An image of a credit card was shown on the screen and participants were instructed to resize the image until it was the same size as an actual card (i.e. membership card, debit card etc.) With this information the screen size was calculated.

The third step was a blindspot test to assess the distance between the participant and their screen. This distance can be determined with an average error of 3.25cm (Li et al., 2020). During this test participants covered their right eye and focused on a black square on the right side of their screen. A red dot which starts positioned slightly left of the black square starts moving to the left. When the dot entered the participant's blindspot, the participant pressed the space-bar.

Using information about the participant's screen size and the distance between the participant and their screen from steps 2 and 3, a scaling factor was determined to scale the videos such that 50px of the videos represented 1° of visual angle for the participant.

The fourth step was a calibration during which the participant was shown nine points, one at a time, spread out over the screen, in random order. A point remained visible until the participant clicked on it. During calibration WebGazer learns the relationship between the eyes of the participant and the fixation point on the screen, by assuming a person is looking at a point when clicking on it (this assumption holds due to the instruction to look at the points while clicking them).

The last step was validation which consisted of the same nine points sequence, but this time the participant was instructed to only focus on the points. Every point was visible for two seconds, one second of which was to allow the participant to change their gaze location (i.e. no data was collected during the first second), and again the points appeared in random order.

When the calibration sequence was completed, participants were informed the introduction and the trials would begin. After the first 4 trials all steps of the calibration sequence were repeated before the last 5 trials could be completed.

When all trials were completed, participants were informed the experiment was over and they were offered the ability to enter their email into a remote document (i.e. google forms) in order to prevent the possibility to couple their email with their study participation ID, ensuring complete privacy of the participants.

Data analysis

For every participant data was saved in a JavaScript dictionary. Gorilla converted this dictionary to a .csv file containing data for every participant. The entire analysis was carried out in the language R using RStudio (RStudio, PBC, v1.3.959).

During pre-processing of the data participants that did not complete the experiment were excluded from the analysis. Then, because WebGazer returns one data table (.json) per participant per video containing all corresponding gaze points, the gaze points were parsed into horizontal screen position in pixels (x), vertical screen position in pixels (y), and time of recording in milliseconds (t) for each participant for each video.

The ROIs (hands, face, eyes, mouth) were manually defined for every video. Then during pre-processing, for every participant, the ROIs were scaled and positioned according to the scaling factor determined using the second and third steps of the calibration sequence and the location of the videos on the screen of the participant.

The main research question concerns accuracy and precision. These measures are well-defined terms in the field of eye tracking (Holmqvist et al., 2011). Accuracy was calculated as the mean of the distance between each recorded gaze point and their corresponding validation

point. Precision was calculated as the root-mean-square of the distances between gaze recordings for each validation point. Every calculation was done in pixels, and due to the rescaling of the stimuli (mentioned in procedure) final values can be transformed to degrees of visual angle. A visual overview of the definitions of accuracy and precision and their visualization in the data can be seen in figure 4.

In order to quantitatively assess whether the webcams of the participants had consistent accuracies and precisions, the accuracies and precisions of the first and second calibrations were compared using an unpaired two-sample t-test.

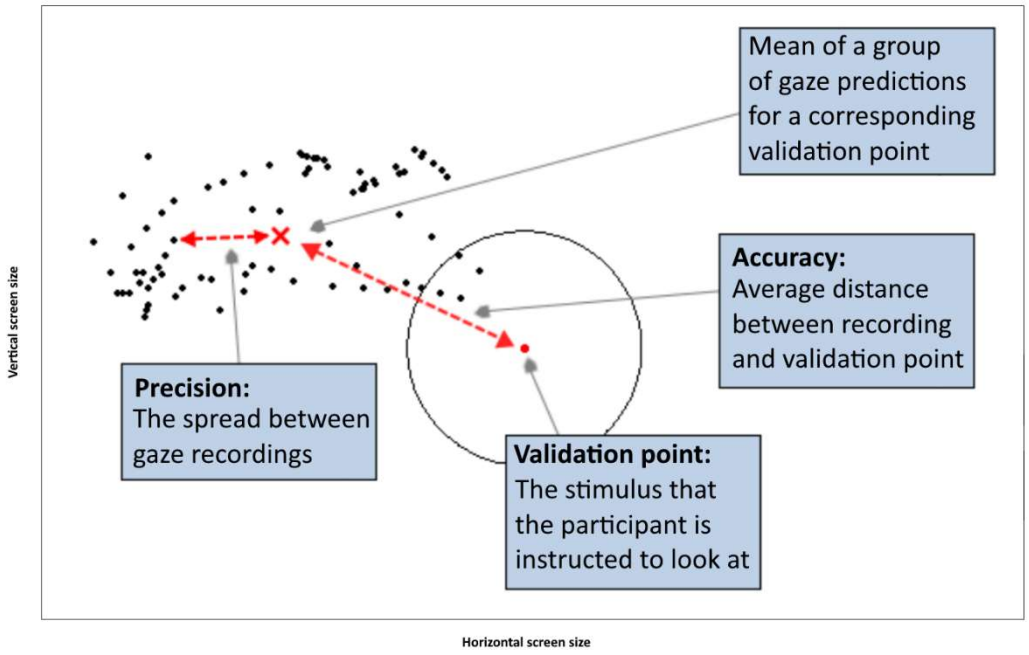

Figure 4. Schematic of accuracy and precision in as defined in the current experiment

The sub research questions are answered using a qualitative approach, as these questions served to explore other properties of webcam-based eye tracking on which further research can focus.

Due to the very graphical nature of eye tracking data (i.e. 2d positions), raw data was plotted in a 2d graph to visualize the data. This qualitative approach led to further steps that

could be taken to possibly improve the data, such as transforming the points by an offset like in the first sub research question. Moreover, this approach led to insights about identifying blinks in the data, even though it was expected that the webcam-based eye tracker would not be able to record blinks due to its low refresh rate. When extracting fixations of participants an Identification by Velocity Threshold (IVT) filter was applied, using a Savitzky-Golay filter as smoothing function.

Results

Participants

Out of the 15 participants who took part in the experiment, 7 (46.7%) participants were excluded from analysis because they did not complete the experiment. Specifically, 6 (85.7%) participants did not watch a single video and 1 (14.3%) participant only completed the first half of the experiment.

Participants' screen resolutions ranged from 1280x720px to 1920x1080px and screen sizes ranged from 11 to 20 inches. The refresh rate of the participants' webcams ranged from 11.83 Hz to 26.09 Hz (M = 18.76 Hz, SD = 5.06 Hz), and the distance between the participants and their screen ranged from 38.16cm to 66.74cm (M = 51.1 cm, SD = 10.5 cm).

Of the 8 included participants, 6 (75%) used Chrome (version 91.0.4472.77 or later) and 2 (25%) used Firefox (version 83.0 or later).

Accuracy and precision

The accuracies and precisions of the webcams are determined for every participant during both calibration sequences in the validation step.

The t-test on accuracy showed that the accuracy of the first calibration (M = 3.65°, SD = 1.23°) was not different from the second calibration (M = 3.64°, SD = 0.97°) ($t(13) = .023$, $p = .982$) which means the webcams had consistent accuracy.

Likewise, the t-test on precision showed that the precision of the first calibration (M = 0.86° RMS, SD = 0.40° RMS) was not different from the second calibration (M = 0.93° RMS, SD = 0.33° RMS) ($t(13) = -.383$, $p = .708$) which means the webcams also had consistent precision.

Transforming the data to increase accuracy

Recall the assumption that a participant is focused on the signer while watching a video. Then, an increase in accuracy is in this case defined as an increase of gaze time spent in one of the ROIs (figure 5). Because every stimulus was preceded by a fixation cross, these fixation crosses can serve as interim singular validation points. The mean offset between the recorded gaze positions and the fixation cross position was determined (figure 6) and using this mean offset all gaze recordings were transformed towards the position of the fixation cross (figure 7).

Before adjusting by the offsets from the interim validations, the participants gaze points fell in any ROI on average 65.57% of the time with a standard deviation of 29.94%. After adjusting for these offsets, gaze points fell in any ROI on average 50.74% of the time with a standard deviation of 35.06%. Out of the 72 videos (9 videos per participant, 8 participants) the total ROI coverage increased only 23 (31.94%) times. A detailed overview of ROI coverage before and after adjusting for interim validations is displayed in table 1.

Table 1

Region of interest coverage before and after adjusting for interim validation offset

	Raw		Adjusted	
	Mean	SD	Mean	SD
Hands	46.00%	28.23%	35.99%	32.54%
Face	16.69%	17.11%	13.24%	18.68%
Mouth	1.63%	2.76%	1.02%	2.24%
Eyes	1.25%	2.57%	0.48%	1.10%
All	65.57%	29.94%	50.74%	35.06%

Figure 5. A comparison of good and poor accuracy in ROIs

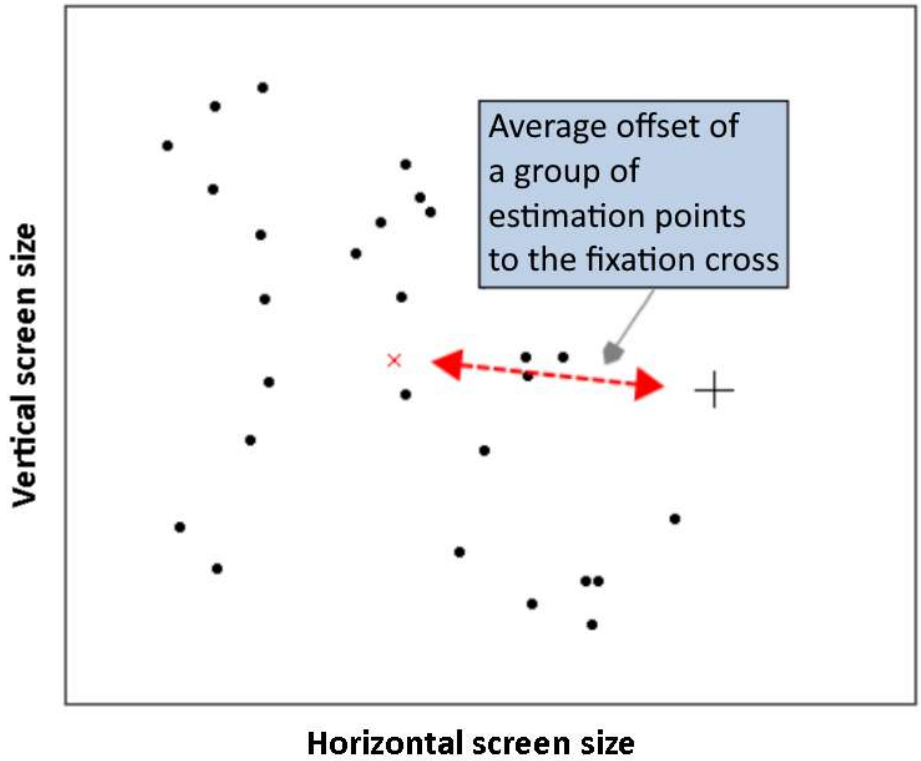

Figure 6. An overview of the interim validation offset. Black points are gaze estimations, the red cross is the mean position of this group of gaze estimations. The black plus is the fixation cross.

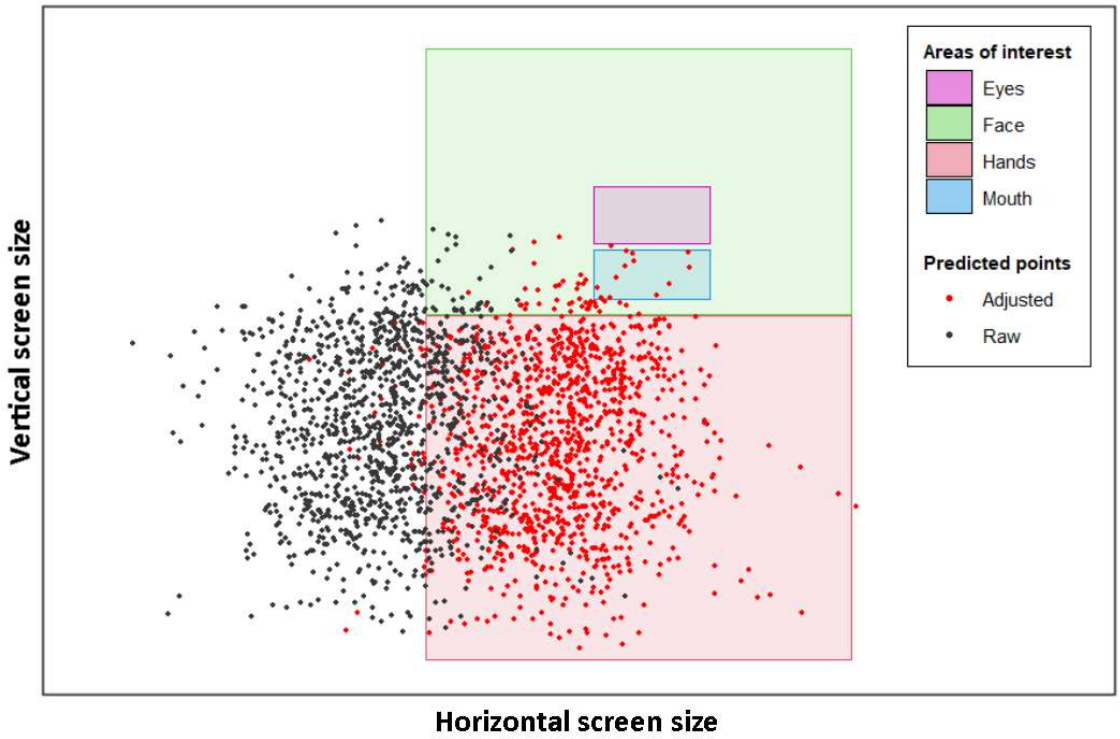

Figure 7. Raw prediction data vs. data adjusted for interim validation offset.

Qualitative results

Fixations and saccades

Data analysis was split between data from the validation task and data collected during each video. Figure 8 shows data from a validation task, and figure 9 shows data from one trial.

As can be seen in figure 8, a software limitation during the validation tasks resulted in gaps in the data. To allow for saccades, the software stopped recording eye tracking data for 1 second as the next validation point was shown (i.e. the participant would already be fixating on the validation point by the time the data recording had started again). Despite this limitation, figure 8 shows that data from a webcam-based eye tracker can be adequately filtered to find fixations. The means of the velocities for the validation points in the bottom plot are similar (M = 542 px/s, SD = 85 px/s). Likewise, the horizontal gaze position plot shows a clear distinction between points at the left- and right side and middle of the screen. The vertical gaze position plot also shows a distinction, though it is not as significant as the horizontal gaze position plot.

Figure 9 illustrates that most eye movements during the videos were smooth pursuit eye movements. This is indicated by the fluctuating gaze position combined with the low velocity. Since participants are instructed to interpret the signs in the video as well as possible (i.e. it is necessary to follow the face and hands with their gaze), smooth pursuit is expected. Notable are the peaks in velocity combined with a sharp shift in the horizontal position of the participants gaze. These jumps are not saccades (despite some of them reaching the saccade velocity threshold), but blinks which will be acknowledged in the next section. Unfortunately, this means we lack data of saccadic eye movements and thus we cannot determine whether saccades could also be labelled in webcam-based eye tracking data.

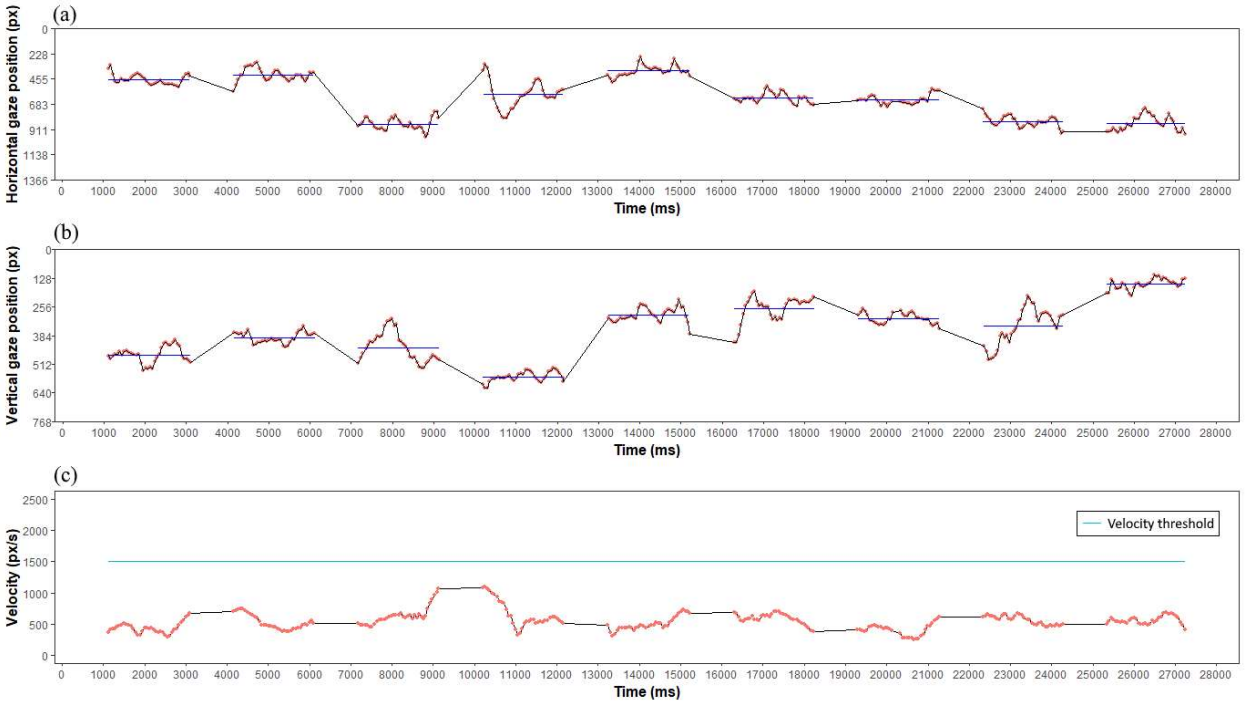

Figure 8. Gaze estimation data for a participant during a validation task. The dark blue lines indicate the mean horizontal/vertical position of the gaze estimations for the corresponding validation point. The cyan line represents the chosen IVT saccade threshold of 30° per second (1500px/s).

(a) Horizontal gaze position over time
(b) Vertical gaze position over time
(c) Gaze velocity over time

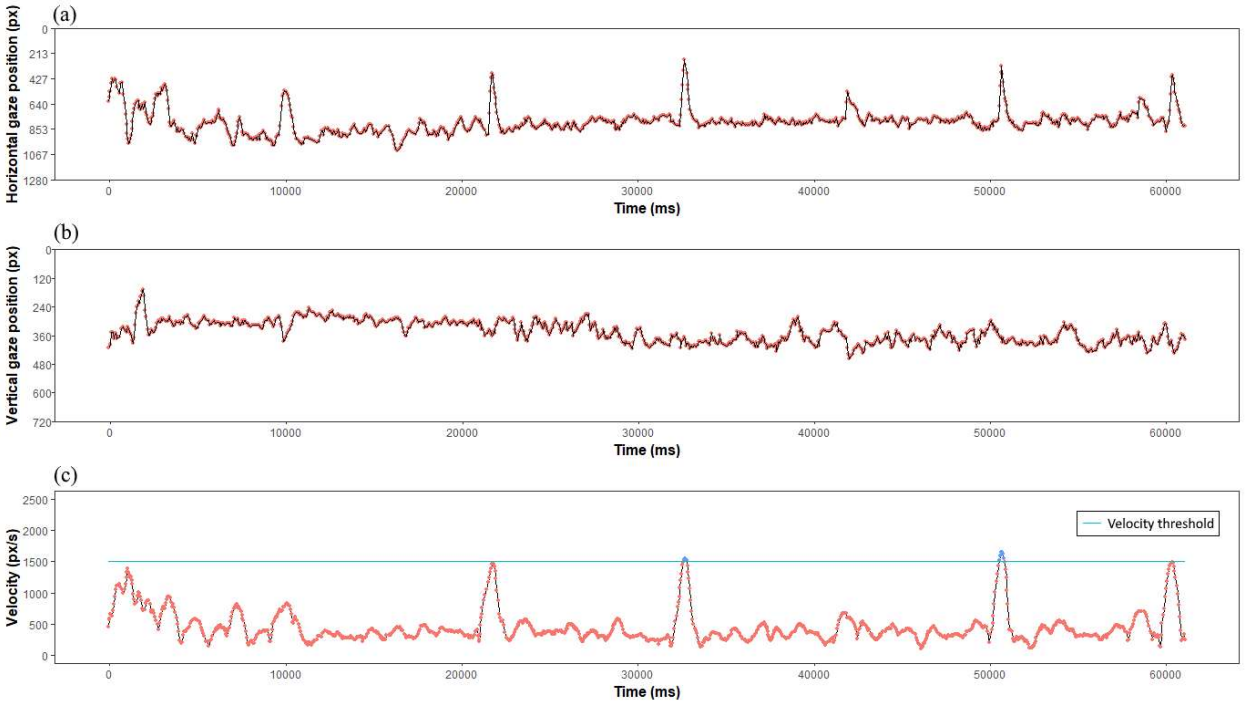

Figure 9. Gaze estimation data for a participant during a video. The dark blue lines indicate the mean horizontal/vertical position of the gaze estimations for the corresponding validation point. The cyan line represents the chosen IVT saccade threshold of 30° per second (1500px/s).
(a) Horizontal gaze position over time
(b) Vertical gaze position over time
(c) Gaze velocity over time

Blinks

The velocity plot in figure 9 shows multiple similar peaks accompanied by sharp movement shifts in the horizontal gaze position plot. Since it is shown previous section on fixations and saccades that the videos call for smooth pursuit eye movements, these peaks are most likely the result of participants blinking. When the eyes close during blinking, WebGazer sees a distorted version of the pupil and assumes this means the eyes are rapidly moving. Sharp movements in the prediction data caused by blinks occur in classical eye trackers, although typically they result in a series of prediction points that go straight down to the bottom of the screen.

A notable point in this experiment is that for every participant, blinks were predicted as eye movements in different directions. In the case of figure 9, blinks are interpreted as movements to the left (towards the negative X), while in other participants blinks were interpreted as movements to a range of left-right up-down movements.

In order to filter these blinks to reduce noise in the data, we again used the assumption that the videos call for smooth pursuit eye movements, as that would imply that fast eye movements are not from the participant and thus are blinks. Thus we can filter blinks by filtering every datapoint above a certain velocity threshold, much like IVT. However, since every webcam has a different representation of the pupil and thus a different direction and speed for the blink, not all blinks can be filtered using the same velocity threshold. Figure 10 illustrates this, by comparing unfiltered- and filtered data from the same video from a participant where the chosen threshold worked very well with data from a participant where the chosen threshold worked very poorly.

The chosen threshold in this case was to filter every point with a velocity of more than 500px/s (10°/s) as a blink. Since participant B had more velocity noise to begin with, they would need a higher threshold to filter blinks because with the chosen threshold of 500px/s too many points are filtered.

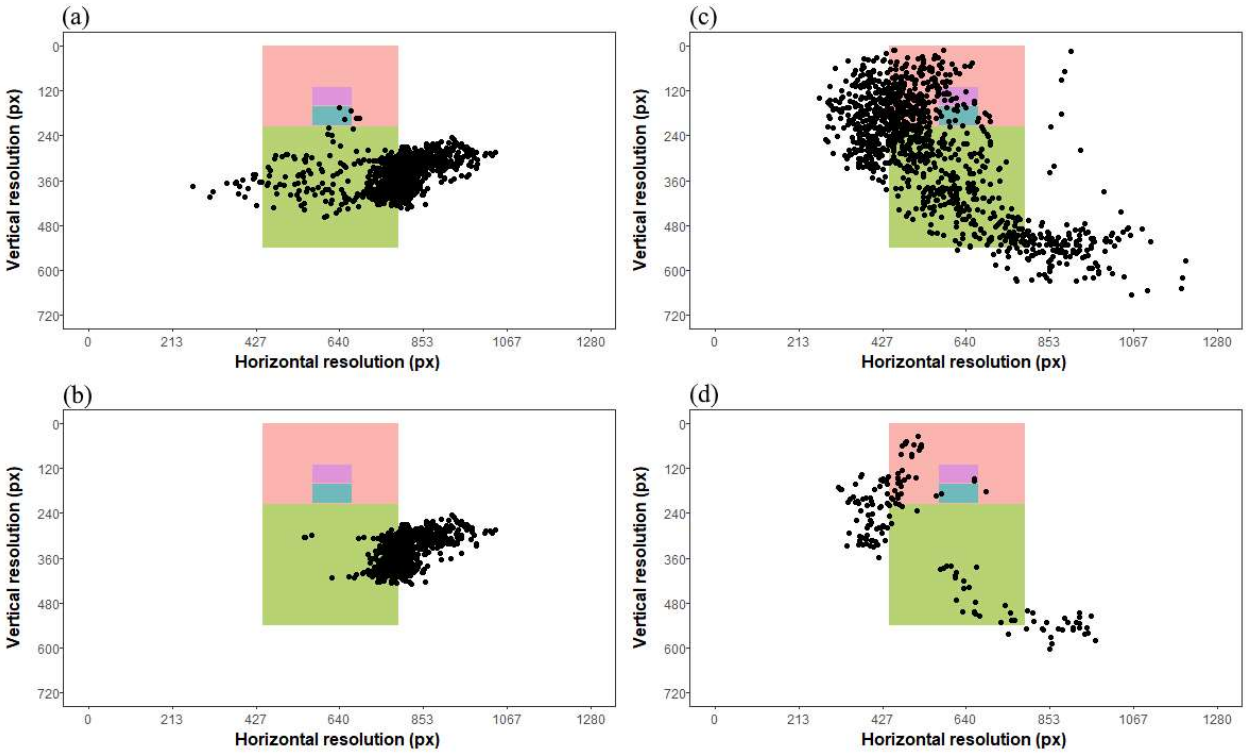

Figure 10. A comparison of datasets where blink filtering was applied. (a) shows unfiltered data from participant A on stimulus A. (b) shows data from participant A on stimulus A where gaze points above the velocity threshold of 10°/s have been filtered out. (c) shows unfiltered data from participant B on stimulus B. (d) shows data from participant B on stimulus B where gaze points above the velocity threshold of 10°/s have been filtered out. In the case of (b), the chosen velocity threshold of 10°/s worked well. In the case of (d), the chosen velocity threshold of 10°/s was too low and thus filtered out too many points.

Discussion

The present study was an exploratory study meant to give insight into the possibilities of webcam-based eye tracking as a possible low-cost, low-requirement alternative to classic video-based eye tracking. More precisely, the present study investigated the accuracy and precision of webcam-based eye tracking and whether the accuracy could be improved by introducing interim validations. Furthermore, the present study qualitatively explored the possibilities to label fixations and filter blinks from the data generated by a webcam-based eye tracker. I hypothesized that a webcam-based eye tracker would have considerably worse accuracy and precision than classic video-based eye trackers, but that the precision would be good enough to classify fixations and saccades. In order to test the hypothesis, WebGazer was implemented in an experiment in which participants calibrated their webcam before watching 9 videos in which a native DSL signer signed stories.

We report 4 main findings. First, webcam-based eye tracking using WebGazer averages an accuracy of 3.65° (SD = 0.77°) and a precision of 0.89° RMS (SD = 0.34° RMS). Second, introducing interim validations did not increase the accuracy of the webcam-based eye tracker. Third, it is possible to label fixations in the webcam-based eye tracker data, although the possibility of labelling saccades is still undetermined. Fourth, it is possible to filter blinks from the webcam-based eye tracker data.

To put the accuracy and precision of the webcam-based eye tracker into perspective, we can compare the accuracy and precision of the webcam-based eye tracker to the accuracies and precisions of other eye trackers often used in publications (appendix B). The accuracy and precision of the webcam-based eye tracker are 10- to 15-fold higher than the reported accuracies and precisions of video-based eye trackers. The difference in accuracy and precision between the

webcam-based eye tracker and video-based eye trackers is however somewhat nuanced by the fact that video-based eye trackers typically have worse accuracies and precisions than reported by the manufacturers in practice (Holmqvist et al., 2011). Despite the webcam-based eye tracker's very poor accuracy and precision when compared to video-based eye trackers, the webcam-based eye tracker's accuracy and precision were good enough for the use case of the present experiment (i.e. large regions of interest on the screen.) Moreover, the accuracy and precision of the webcam-based eye tracker were good enough to show a clear distinction between calibration points in figures 8a and 8b. The present study aimed to gain baseline insight into webcam-based eye trackers, without any additional optimization during the experiment. Future research could focus on improving data quality by optimizing the environment of the experiment, for example by mimicking the work of Nyström et al. (2012) and using a webcam-based eye tracker rather than a video-based eye tracker like Nyström et al. (2012) did.

The introduction of interim validations in the present experiment by way of fixation crosses before the videos was not enough to consistently improve the accuracy of the webcam-based eye tracker. The gaze estimations drifted as the experiment went on, but in seemingly random directions. No clear pattern of drift of the gaze estimations has been found, and it is unclear why interim validations sometimes improved the accuracy of the webcam-based eye tracker and sometimes did not. The present experiment had no camera feed of the participants, and thus it could not be determined whether the drift was caused by the webcam-based eye tracker or by external causes (e.g. the participant changing their position in front or the camera, or changes in room illumination.) Future research should determine whether the cause of the drift of the gaze estimations lies within the webcam-based eye tracker, and successively focus on possible remedies for that drift.

Fixations could consistently be labelled across all participants from the data from their calibrations. This finding supports the hypothesis that the precision of the webcam-based eye tracker is sufficient to label fixations, despite the precision of the webcam-based eye tracker being much lower than the precisions of video-based eye trackers used in practice (appendix B). Therefore, this finding suggests that webcam-based eye tracking can be used in experiments where optimal accuracy and precision are not required. Although it remains undetermined whether saccades could also be labelled from the webcam-based eye tracking data, we can form an expectation based on data from the present experiment. The very low average webcam-based eye tracker sample rate of 18.76 Hz (SD = 5.05 Hz) combined with the very low maximum webcam-based eye tracker sample rate of 26.09 Hz suggests that it is unlikely that webcam-based eye trackers can record maximum saccade velocity for most saccades. Future research could setup an experiment that stimulates saccades to confirm the expectation that maximum saccade velocity cannot be recorded with webcam-based eye trackers.

The webcam-based eye tracker did detect blinks, and the data was of sufficient quality to allow for filtering of said blinks of the participants by way of filtering by velocity of the gaze estimations. The ability to detect blinks in the data is an important quality for eye trackers because blinks introduce noise. Unlike video-based eye trackers, the webcam-based eye tracker did not detect blinks as a series of gaze estimations straight to the bottom of the screen. Instead, the blinks were detected in vastly different directions depending on the participant. Furthermore, the velocity of the blinks also differed per participant. These findings show promise that blinks can be filtered out of the webcam-based eye tracking data very accurately. Future research could attempt to determine a relationship between the webcam-based eye tracker's sample rate and the blink direction and velocity in order to filter blinks more accurately.

The present study gives insight into the strengths and weaknesses of webcam-based eye tracking. Based on the accuracy and precision, webcam-based eye tracking would work well in experiments which do not require optimal accuracy and precision. Based on the possibility to classify fixations combined with the low sample rate, experiments that are temporally less demanding play to the strengths of a webcam-based eye tracker (e.g. experiments focussing on visual spatial attention.) Furthermore, the fact that a webcam-based eye tracker is very suited for online experiments (since participants only need a device with a web-browser and a webcam) means that a very large participant population can be reached very quickly.

Future research should play to the strengths of webcam-based eye tracking by focussing on its limitations found in the present experiment, rather than focussing on finding more limitations. I suggest prioritizing insight on the cause of the drift of the webcam-based eye tracker, as that was its most significant weakness. Furthermore, future research should focus on finding a way to consistently filter out blinks given the seeming randomness involved in detecting them.

Conclusion

By setting up an online experiment using a webcam-based eye tracker, the present study has provided insight in the properties of webcam-based eye tracking, and into its strengths and weaknesses.

The present study has shown that the accuracy and precision of webcam-based eye trackers, although notably worse than video-based eye trackers, is a promising tool for experiments which can work around the limitations of webcam-based eye trackers (i.e. low sample rate, accuracy and precision.) and can play to the strengths of webcam-based eye tracking (i.e. online experiments).

While the field of eye tracking was already established in the field of artificial intelligence, the limitations of a webcam provide further opportunities to enhance eye tracking techniques around these limitations.

References

Andersson, R., Nyström, M., & Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research*, *3*(3), 1–12. https://doi.org/10.16910/jemr.3.3.6

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher*, *40*(1), 62–69. https://doi.org/10.1080/0142159X.2017.1391373

Burton, L., Albert, W., & Flynn, M. (2014). A comparison of the performance of webcam vs. infrared eye tracking technology. *Proceedings of the Human Factors and Ergonomics Society*, *2014-Janua*, 1437–1441. https://doi.org/10.1177/1541931214581300

Crasborn, O. (2016). What is a sign language? *Linguistic Approaches to Bilingualism*, *6*(6), 768–771. https://doi.org/10.1075/lab.6.6.04cra

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Doof. (n.d.). *Doof*. Retrieved June 16, 2021, from https://www.doof.nl/hoorbibliotheek/taal/gebarentaal/

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye gaze during comprehension of American sign language by native and beginning signers. *Journal of Deaf Studies and Deaf Education*, *14*(2), 237–243. https://doi.org/10.1093/deafed/enn037

Engbert, R. (2006). Microsaccades: a microcosm for research on oculomotor control, attention, and visual perception. *Progress in Brain Research*, *154*(SUPPL. A), 177–192. https://doi.org/10.1016/S0079-6123(06)54009-9

Gebarencentrum. (n.d.). *Gebarencentrum*. Retrieved June 16, 2021, from https://www.gebarencentrum.nl

Hansen, D. W., & Ji, Q. (2010). In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 478–500. https://doi.org/10.1109/TPAMI.2009.30

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye-tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., Lee, M. H., Chiou, G. L., Liang, J. C., & Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. In *Educational Research Review* (Vol. 10, pp. 90–115). Elsevier. https://doi.org/10.1016/j.edurev.2013.10.001

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-019-57204-1

Lundgren, E., Rocha, T., Rocha, Z., Carvalho, P., & Bello, M. (2014). *tracking.js: A modern approach for Computer Vision on the web*. https://trackingjs.com

Mathias, A. (2014). *clmtrackr: Javascript library for precise tracking of facial features via Constrained Local Models*. https://github.com/auduno/clmtrackr

Morgans, H. G. (2008). *Morgans (1999) Where did South African Sign Language Originate?*

30(1), 53–58. https://doi.org/10.1080/10228199908566144

Nyström, M., Andersson, R., Holmqvist, K., & Weijer, J. van de. (2012). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods 2012 45:1*, *45*(1), 272–288. https://doi.org/10.3758/S13428-012-0247-4

Papoutsaki, A., Daskalova, N., Sangkloy, P., Huang, J., Laskey, J., & Hays, J. (2016a). WebGazer: Scalable webcam eye tracking using user interactions. *IJCAI International Joint Conference on Artificial Intelligence*.

Papoutsaki, A., Daskalova, N., Sangkloy, P., Huang, J., Laskey, J., & Hays, J. (2016b). WebGazer: Scalable webcam eye tracking using user interactions. *IJCAI International Joint Conference on Artificial Intelligence*, *2016-Janua*, 3839–3845. https://webgazer.cs.brown.edu

Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-Tracking Measures in Aviation: A Selective Literature Review. In *International Journal of Aerospace Psychology* (Vol. 28, Issues 3–4, pp. 98–112). https://doi.org/10.1080/24721840.2018.1514978

Richters, M., Rutten, P. A. W., & van Vliet, B. Y. A. (2021). *Novice and expert gaze behavior in users of sign language*. Utrecht University.

Santos, R. D. O. J. dos, Oliveira, J. H. C. de, Rocha, J. B., & Giraldi, J. D. M. E. (2015). Eye Tracking in Neuromarketing: A Research Agenda for Marketing Studies. *International Journal of Psychological Studies*, *7*(1). https://doi.org/10.5539/ijps.v7n1p32

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Sharafi, Z., Soh, Z., & Guéhéneuc, Y. G. (2015). A systematic literature review on the usage of

eye-tracking in software engineering. *Information and Software Technology*, *67*, 79–107. https://doi.org/10.1016/j.infsof.2015.06.008

Stokoe, W. C. (1980). Sign Language Structure. *Annual Review of Anthropology*, *9*, 365–390. http://www.jstor.org.proxy.library.uu.nl/stable/2155741

Tschirsich, M. (2012). *Js-objectdetect: Computer vision in your browser - javascript real-time object detection.* https://github.com/mtschirs/js-objectdetect

Yarbus, A. L. (1967). *Eye Movements and Vision*. https://books.google.nl/books?hl=nl&lr=&id=kRf3BwAAQBAJ&oi=fnd&pg=PA1&ots=c8bdS7_2bO&sig=YsfYTyPtvopU2JYhdgBhRsBvc0g&redir_esc=y#v=onepage&q&f=false

Appendix A

This GitHub repository contains the entirety of the code used for the experiment. A detailed explanation of de code can be found in the README of the repository.

https://github.com/mwessel99/bachelor-ai-thesis

Appendix B

An overview of reported accuracies and precisions of eye tracking devices of 3 eye tracking device brands with the most publications as reported by https://imotions.com/blog/top-eye-tracking-hardware-companies/. Accuracies and precisions are retrieved from the manufacturer's websites, and are reported under optimal conditions, without any filtering applied.

Brand	*Model*	*Accuracy*	*Precision*
Tobii Pro	Fusion	0.3°	0.2° RMS
	Nano	0.3°	0.10° RMS
	Spectrum	0.3°	0.06° RMS
	Glasses 3	N/A	N/A
EyeLink	1000 Plus	0.15°	0.01° RMS
	II	N/A	< 0.01° RMS
SMI	Glasses	0.05°	N/A

Appendix C

The questions of the demographic questionnaire that participants completed before partaking in the present experiment. The questions were originally in Dutch, so a translation has been provided. Brackets indicate answer possibilities.

1. Wat is uw geslacht? [Man/Vrouw/Anders]

   What is your sex? [Male/Female/Other]
2. Wat is uw geboortejaar? [Getal]

   What is your year of birth? [Number]
3. Wat is uw horendheid? [Horend/Doof/Slechthorend]

   What is your level of hearing? [Hearing/Deaf/Hard of hearing]
4. Hoe vaak gebruikt u een tolk? (Alleen als doof/slechthorend) [Getal]

   How often do you utilize an interpreter? (Only if deaf/hard of hearing) [Number]
5. Wat is uw moedertaal? [Nederlands/NGT/Anders]

   What is your native language? [Dutch/DSL/Other]
6. Welke term(en) is/zijn toepasbaar op uw NGT niveau? [NGT student/Minor NGT/Tolk/Docent/Moedertaal/Anders]

   Which term(s) apply to your DSL level? [DSL student/DSL minor student/Interpreter/Teacher/Native signer/Other]
7. Draagt u een bril? [Ja/Nee]

   Do you wear glasses? [Yes/No]