Behavior Research Methods (2018) 50:1374–1397
https://doi.org/10.3758/s13428-018-1050-7

A novel evaluation of two related and two independent algorithms for eye movement classification during reading

Lee Friedman[1] · Ioannis Rigas[1] · Evgeny Abdulin[1] · Oleg V. Komogortsev[1]

Published online: 15 May 2018

Abstract

Nyström and Holmqvist have published a method for the classification of eye movements during reading (ONH) (Nyström & Holmqvist, 2010). When we applied this algorithm to our data, the results were not satisfactory, so we modified the algorithm (now the MNH) to better classify our data. The changes included: (1) reducing the amount of signal filtering, (2) excluding a new type of noise, (3) removing several adaptive thresholds and replacing them with fixed thresholds, (4) changing the way that the start and end of each saccade was determined, (5) employing a new algorithm for detecting PSOs, and (6) allowing a fixation period to either begin or end with noise. A new method for the evaluation of classification algorithms is presented. It was designed to provide comprehensive feedback to an algorithm developer, in a time-efficient manner, about the types and numbers of classification errors that an algorithm produces. This evaluation was conducted by three expert raters independently, across 20 randomly chosen recordings, each classified by both algorithms. The MNH made many fewer errors in determining when saccades start and end, and it also detected some fixations and saccades that the ONH did not. The MNH fails to detect very small saccades. We also evaluated two additional algorithms: the EyeLink Parser and a more current, machine-learning-based algorithm. The EyeLink Parser tended to find more saccades that ended too early than did the other methods, and we found numerous problems with the output of the machine-learning-based algorithm.

Keywords Eye movement · Classification · Evaluation · Comparison

In 2010, Nyström and Holmqvist published a major article describing a new method for classifying eye movements during reading. The object was to classify such eye movement signals into periods of fixation, saccades, and postsaccade oscillations (PSO). (Note that another eye movement type—i.e., smooth pursuit—is not present in a typical text-reading task and was not classified by the original method [Nyström & Holmqvist, 2010] as a separate category.) In addition to the eye movement categories, some data are also classified as "noise" (really, "artifacts"), and some are left unclassified. (By design, the algorithm does not classify all events in the recording.) The algorithm is very sophisticated and has many novel features, including the fact that it is the first algorithm to automatically detect PSOs. In describing their method, the authors emphasized its "adaptive" aspect, which involved setting various thresholds based on each individual recording's characteristics.

Several years later, we were in possession of a large database of recorded eye movements during reading. M. Nyström kindly made the Matlab (Natick, MA) code available to us. The coding style was remarkably sophisticated, elegant, and efficient. We did find one error in the many lines of code. Although in the present case this error played no role in our analysis, in other contexts it might be very important.[1] We made minor technical modifications to the code in order to better accommodate our data format and to report out the data as we preferred. One substantive change we had to make from the start was the sampling rate. The original code was written for data collected from an SMI HiSpeed eyetracker (Teltow, Germany) at 1250 Hz, but our data were collected with an EyeLink 1000 (SR Research, Kanata, ON, Canada) at 1000 Hz. Dr. Nyström informed us that his algorithm was adaptable

✉ Lee Friedman
lfriedman10@gmail.com

[1] Department of Computer Science, Texas State University, Dr. Komogortsev's Lab (Comal 307D), Comal Building, 601 University Drive, San Marcos, TX 78666, USA

[1] Documentation of the programming error is available in the file "DocumentationOfaProgrammingErrorInTheONH.docx" at https://digital.library.txstate.edu/handle/10877/6874.

to the lower sampling rate with a change to a single parameter in the code. The EyeLink 1000 automatically removes data it determines to be blinks and replaces these samples with "NaN" values. The SMI eyetracker, on the other hand, automatically replaces periods in which the eyes are closed with 0. The code also needed to be modified to deal with this difference.

The original algorithm (ONH) produced interpretable results with our data; we carefully inspected the resulting classifications but were not satisfied with the results. The first author (L.F.) of the present article spent approximately 3–4 months modifying the ONH to produce a new version (MNH). After some formal inspection and assessment, we decided that the MNH was quite usable for us, although we knew that it was not perfect.

Before one can evaluate an algorithm, one first needs to establish the goals of the algorithm. For example, one goal might be to classify the raw data, another might be to classify the smoothed data, and a third might be to classify the velocity trace. Because the ONH employs a Savitzky–Golay filter with a window length for our data of 19ms, and given the fact that this filter would delay the resulting smoothed position and velocity calculation by 9 ms and that this filter delay was never corrected in the original code, the algorithm could not be intended to classify the raw position signal. (Although Nyström & Holmqvist, 2010, claim that the delay is removed, this is not correct.[2] We provide more information about this issue below.) Although there is no explicit mention of this, it seems much more likely that the algorithm was designed to classify the smoothed position trace rather than the velocity trace, since classification of the position trace is typically what is desired.

We wanted to compare the performance of the two algorithms, but we were interested particularly in the issues that other algorithm developers would want to have addressed. As one example, for saccade initiation, we wanted to count the number of times each algorithm detected the onset of a saccade too early, on the basis of human expert judgment. Another example would be to measure the number of fixation periods each algorithm failed to detect. This type of information would directly assist the algorithm developer in focusing exactly on specific problematic decisions the programs need to make.

Recently, Andersson, Larsson, Holmqvist, Stridh, and Nystrom (2017) published a major study on a new evaluation method for comparing the results of many algorithms. This approach started with 34 recordings that two human experts had classified and compared the expert classifications to those from many other algorithms. The results were expressed in terms of the degree of agreement on a sample-by-sample basis, or by comparing the resulting distributions of basic parameters such as fixation duration or peak saccade velocity. Although the evaluation approach described by Andersson et al. was not used to detect the kinds of classification errors we will emphasize in this report, it seems likely that their approach could be modified to provide such data. However, since this evaluation method requires human experts to classify, in many recordings, every sample as belonging to one event type or another, this is a very cumbersome and inefficient method to get the kind of classification error counts that we seek, especially when comparing only two or three algorithms. However, we concede that the approach of Andersson et al. would probably be faster when comparing more than three algorithms.

During the review process, it was agreed that we would also compare both the ONH and the MNH to two additional algorithms, the EyeLink Parser (SR Research Ltd, Ottawa, Ontario, Canada) and a modern, machine-learning-based eye movement classification algorithm (Zemblys, Niehorster, Komogortsev, & Holmqvist, 2018). The EyeLink Parser is provided with an EyeLink eye movement device similar to the one employed by the EyeLink 1000 herein. This comparison allowed us to document the performance of the MNH as compared to a very commonly employed alternative algorithm. It is important to note that the EyeLink Parser is a real-time, online algorithm and was not designed to classify PSOs.

With respect to the machine-learning-based algorithm, the concern was that the ONH was not considered to be a state-of-the-art algorithm at this time, and that an improvement on it would possibly also not be competitive with state-of-the-art algorithms. The algorithm of Zemblys et al. (2018) was chosen because it is considered a modern, state-of-the-art algorithm, because it classifies PSOs in addition to fixations and saccades, and because it was simple for us to have our data scored with this method. The Zemblys et al. algorithm is based on a machine-learning approach termed "Random Forest," and thus Zemblys et al. refer to the algorithm as the IRF. At the time that the decision to also evaluate these two comparisons was made, two members of the algorithm evaluation panel (authors E.A. and I.R.) were no longer involved, so these two additionial comparisons were based on a single evaluator (author L.F.). For comprehensive reviews of eye movement methods, see Andersson et al. (2017), Hein and Zangemeister (2017), and Salvucci and Goldberg (2000).

We began with a taxonomy of error types ($N = 32$, Table 1) that would cover all the important decisions that an algorithm would logically need to make. However, using this taxonomy of errors, it is possible that the exact same samples might be misclassified by different raters in more than one way. For example, imagine the transition from the end of a saccade to a PSO. One error could be that the saccade ends too late, so

[2] For a detailed discussion of the filter delay issue, see "DiscussionOfTheFilterDelayIssueWithTheONH.docx" at https://digital.library.txstate.edu/handle/10877/6874.

Table 1 Taxonomy of eye movement classification errors during reading

Type	Noise Detection	Type	Noise Timing
1	Noise Misclassified as Fixation	17	Noise Starts Too Early
2	Noise Misclassified as Saccade	18	Noise Starts Too Late
3	Noise Misclassified as PSO	19	Noise Ends Too Early
4	Noise not Detected	20	Noise Ends Too Late
Type	**Fixation Detection**	**Type**	**Fixation Timing**
5	Fixation Misclassified as Noise	21	Fixation Starts Too Early
6	Fixation Misclassified as Saccade	22	Fixation Starts Too Late
7	Fixation Misclassified as PSO	23	Fixation Ends Too Early
8	Fixation not Detected	24	Fixation Ends Too Late
Type	**Saccade Detection**	**Type**	**Saccade Timing**
9	Saccade Misclassified as Noise	25	Saccade Starts Too Early
10	Saccade Misclassified as Fixation	26	Saccade Starts Too Late
11	Saccade Misclassified as PSO	27	Saccade Ends Too Early
12	Saccade not Detected	28	Saccade Ends Too Late
Type	**PSO Detection**	**Type**	**PSO Timing**
13	PSO Misclassified as Noise	29	PSO Starts Too Early
14	PSO Misclassified as Fixation	30	PSO Starts Too Late
15	PSO Misclassified as Saccade	31	PSO Ends Too Early
16	PSO not Detected	32	PSO Ends Too Late

too many samples are labelled as saccade. But the same result could be interpreted as a PSO that starts too late. We resolved this problem by creating a hierarchy of decisions, described below. In this hierarchy, saccades precede PSOs, so a late transition from a saccade to a PSO would always be classified as a saccade that ends too late. In this case, the error "PSO Starts Too Late" would never occur.

We developed software to allow a rater to classify and count the error types made by each algorithm. The goal of the present report is to document the error types and rates for each algorithm. In this process, we introduce a new algorithm for eye movement classification and a new method for evaluating any eye movement classification algorithm.

General method

Subjects for eye movement data

We report on 20 recordings from 20 different subjects, randomly selected from a larger set of over 300 eye movement recordings. The original data set was collected as part of an effort to develop a biometric approach to identify subjects on the basis of their eye movement characteristics (Rigas, Komogortsev, & Shadmehr, 2016). The study subjects, all undergraduate college students, were recruited using several methods. Several instructors in the Computer Science, Engineering, Math, and Psychology departments of Texas State University mentioned the project to students and offered extra credit for participation. Also, an e-mail was sent to all first- and second-year students explaining the study and requesting participation. Students were not paid for their participation for the round of data collection employed in this work. We also selected four additional recordings to serve as a training sample for the manual eye movement classification procedure described below. The mean (with *SD*) age of the 20 subjects (seven females) was 22.5 years (5.5).

The reading task

The subjects in the original study viewed seven different tracking tasks. Only the text-reading task is relevant to the present report. Each subject was asked to read, silently, an identical pair of quatrains from the famous nonsense poem "The Hunting of the Snark," written by Lewis Carroll (written from 1874 to 1876). The text was displayed in Times New Roman 20-point bold font and was single-spaced. The mean letter interval for each piece of text was approximately 0.50 deg of visual angle. The height of the line of the text was 0.92 deg of visual angle.

Eye movement recording

All recordings were conducted by trained operators who were present with the subjects for the entire testing period. The subjects were seated 55 cm in front of a computer monitor with their heads resting on a chin/head rest. The monitor subtends ± 23.3 deg of visual angle in the horizontal direction, 11.7 deg to the top and 18.5 deg to the bottom. The EyeLink 1000 (SR Research Ltd., Kanata, ON, Canada), a video-oculography system that employs detection of both the pupil and the corneal reflection to determine gaze position, was used to record eye movements. It records both vertical and horizontal eye movements, binocularly. In the present study, only left eye movements were collected. For 298 subjects, we have a mean spatial accuracy of 0.50 (*SD* = 0.17, min = 0.20, max = 1.06). For further specifications, see the SR Research website (www.sr-research.com). The sampling rate for our data was 1000 Hz. Prior to each task, a calibration dataset, consisting of nine screen positions (primary position and eight peripheral points), was collected by the EyeLink 1000. A position calibration validation procedure was run immediately after calibration. If there was little error (maximum error < 1.5 deg, average error < 1.0 deg), the calibration was accepted. Otherwise, the equipment was readjusted and the calibration was redone. During the recording, before each task, the operator explained the next task (stimulus). The EyeLink 1000 transformed the raw records into gaze position data, in visual angle units, using the calibration data collected at the start of each task. The Stampe (1993) heuristic spike removal algorithm was employed. (It was also used on the data processed for the ONH.) We set the parameter File Filter Setting = 2 on the Eyelink 1000. This means that the File Sample Filter was set to Extra, meaning that the data recorded would be filtered using a two-stage recursive heuristic filter of the type described in the Stampe article. Over the course of several generations of EyeLink eyetrackers, SR Research has made some small, proprietary changes to the exact heuristics to further improve their performance, but the fundamental approach remains the same as the one outlined in the Stampe article. Our EyeLink software version was 4.48. Also, blinks were detected and removed from the data by the EyeLink 1000. The eye movements were analyzed offline. Only the first 26 s of recordings for each subject were chosen for this study, because one of the subjects chosen randomly finished reading the poem in approximately 26 s, and we wanted the same amount of data from each subject to be represented in each recording.

Experiment 1: ONH versus MNH

Method

Algorithm narrative

This is a description of the steps taken by the ONH (Nyström & Holmqvist, 2010) and the MNH algorithms to classify eye movements.

Load the data For the ONH, the data have 0 values for both horizontal and vertical position during periods of eye closure. For MNH, periods of blinks have NaN values. In addition, the spikes in the MNH data were filtered using the Stampe (1993) filter. [We were informed by Dr. Holmqvist (personal communication) that the ONH data were also filtered by the Stample filter.] As we will document below, it appears from our results that the data used in the original Nyström and Holmqvist (2010) were substantially noisier than ours.

Filter the data Smooth the data and compute the smoothed position trace, smoothed velocity trace, and smoothed acceleration trace. Both the ONH and the MNH use the Savitzky–Golay filter function **sgolay** to compute filter coefficients (Both algorithms were written in Matlab code and run in Matlab [Version R2015b; The MathWorks, Natick, MA].) Both algorithms employ the **filter** routine to apply the filter coefficients, to compute the smoothed position signal, the velocity signal and the acceleration signal. Note that despite the claims by Nyström and Holmqvist (2010, p. 192), the filter delay is not removed. Use of the **conv** function (convolution), as in **conv(X,g(:,1),'same')**, where **X** is the signal and **g** are the filter coefficients, would have removed the delay. Use of the **sgolayfilt** function would also have removed the delay. Note that, for 1000 Hz, the ONH filter width was 19 samples (i.e., 19 ms), but MNH changes this to 7 ms. Both algorithms used the same filter order—that is, 2.

Compute radial velocity and radial acceleration If Vel_X is the velocity for the horizontal eye movement signal and Vel_Y is the velocity of the vertical eye position, then

$$\text{radial velocity} = \sqrt{Vel_x^2 + Vel_y^2} \tag{1}$$

If Acc_X is the acceleration for the horizontal eye movement signal and Acc_Y is the acceleration of the vertical eye position, then

$$\text{radial acceleration} = \sqrt{Acc_x^2 + Acc_y^2} \tag{2}$$

From this point forward, all references to velocity refer to radial velocity and all references to acceleration refer to radial acceleration.

Detect and classify noise and artifact Artifact occurs in the ONH when the velocity signal is greater than 1,000 deg/s or when the acceleration signal is greater than 100,000 deg^2/s. For the ONH, signal values of 0 in the raw horizontal and raw vertical position trace indicate artifact (eyes closed), and these samples are therefore marked as artifact. For the MNH, we defined an artifact as any period when the velocity signal was greater than 1,500 deg/s or the acceleration signal was greater than 100,000 deg^2/s. In the case of the MNH, blinks are automatically detected by the eyetracker, and signal values are replaced with the code NaN. For the MNH, these samples are also marked as artifacts.

The MNH also used a routine to remove a distinct type of noise that we call "Rapid Irregularly Oscillating Noise of Eye Positional Signal" (RIONEPS) (Fig. 1) (Abdulin, Friedman, & Komogortsev, 2017). Our evidence suggests that this type of noise is caused by intermittent failures of the eye-tracking system to properly detect either the pupil center of mass or the corneal reflection center of mass. An article describing this noise, providing examples from multiple vendors and presenting our algorithm to detect it, is available as Abdulin et al. (2017)

Extend artifact blocks backward and forward in time It is reasonable to assume that some signal before and after an artifact might also be contaminated by the artifact, so both algorithms extend the data marked as artifact forward and backward in time. The two algorithms use the identical method, but the thresholds used are completely different. For the ONH, a velocity threshold is set to be the median of the entire velocity signal (all samples in the recording) multiplied by 2.0. This is in contrast to the threshold described by Nyström and Holmqvist (2010), page 193, which states that the threshold is the median of the entire signal but not multiplied by 2. Artifacts are extended both forward and backward in time until the velocity is less than or equal to this threshold. For the 20 subjects in the present study, the median velocity threshold value was 5.59 (25th percentile = 5.03, 75th percentile = 6.43). We found that, for our data, this threshold led to the removal of much more signal before and after each artifact block than was reasonable when compared to the ground truth represented by visual inspection. For the MNH, we computed our own threshold. First, we divided the signal into two categories: samples in which the velocity was above 100°/s and samples in which the velocity was below 100°/s. We considered the velocity samples below 100°/s to be a crude estimate of the velocity during fixation. Any "fixation" blocks of continuous samples that were shorter than the minimum fixation duration (for MNH: 0.030 s) were excluded from further analysis. To obtain the most stable fixation velocity values, for each remaining fixation block, data were rejected that were within 5 ms of the start and end of each fixation block. This procedure would leave in place only the central portions of each fixation block, where the signal would likely be most stable. We combined the velocities from these remaining fixation samples and computed the 90th percentile of the velocity noise distribution. This is the only adaptive threshold in the MNH. We used this as our threshold for deciding how much data to exclude before and after each artifact block. (The MNH also uses this threshold to determine the end of a PSO [see below].) In the 20 subjects in the present study, the median 90th-percentile velocity threshold was 14.66°/s (25th percentile = 10.81, 75th percentile = 16.87). This threshold is 2.6 times greater than the threshold used by the ONH algorithm, and therefore excludes much less data before and after each artifact block. This produced much better results with our data.

Description of the general approach to finding saccades in the ONH The general approach to finding saccades for the ONH is illustrated in Fig. 2A. All nonartifact velocity signals greater than the "saccade peak velocity" threshold are labeled as potential saccades, assuming there are a minimum of two contiguous samples above this threshold (for the data used in Nyström & Holmqvist, 2010, three contiguous samples). To identify the start of a saccade, the algorithm steps backward in time, one sample at a time, to determine whether the velocity crosses below what we call the "saccade subthreshold," which is always substantially lower than the saccade peak velocity threshold. If the event crosses below the saccade subthreshold, then the algorithm steps back in time to find the first local velocity minimum. This is the saccade start. To identify the end of the saccade, the algorithm steps forward time, to determine whether the velocity crosses below what is called the "local velocity noise threshold," which is determined for each saccade on the basis of a weighted sum:

$$\begin{aligned}\text{Local Velocity Noise Threshold}\\ &= 0.7{*}\text{Saccade Subthreshold} \\ &+ 0.3{*}\text{Local Velocity Noise}\end{aligned} \tag{3}$$

The velocity signals that contribute to local velocity noise are from the saccade start backward for 0.040 s (at 1000 Hz, 40 samples; see open circles in Fig. 2A).

$$\begin{aligned}\text{Local Velocity Noise} &= \text{mean(velocity signals)}\\ &+ 3{*}SD(\text{velocity signals})\end{aligned} \tag{4}$$

If the event crosses below the local velocity noise threshold, then the algorithm steps forward in time to find the first local velocity minimum. This is the saccade end. A saccade is rejected if the local velocity noise is greater than the saccade peak velocity threshold.

Computing the adaptive thresholds for saccade peak velocity and saccade subthreshold in the ONH For the ONH, the peak velocity of a saccade must exceed the saccade peak velocity threshold. This threshold is computed iteratively in a loop. Initially, the saccade peak velocity threshold is set at 100°/s. During each iteration of the loop, values below this threshold are considered as potential fixation periods. Any "fixation" blocks of continuous samples that were shorter than the

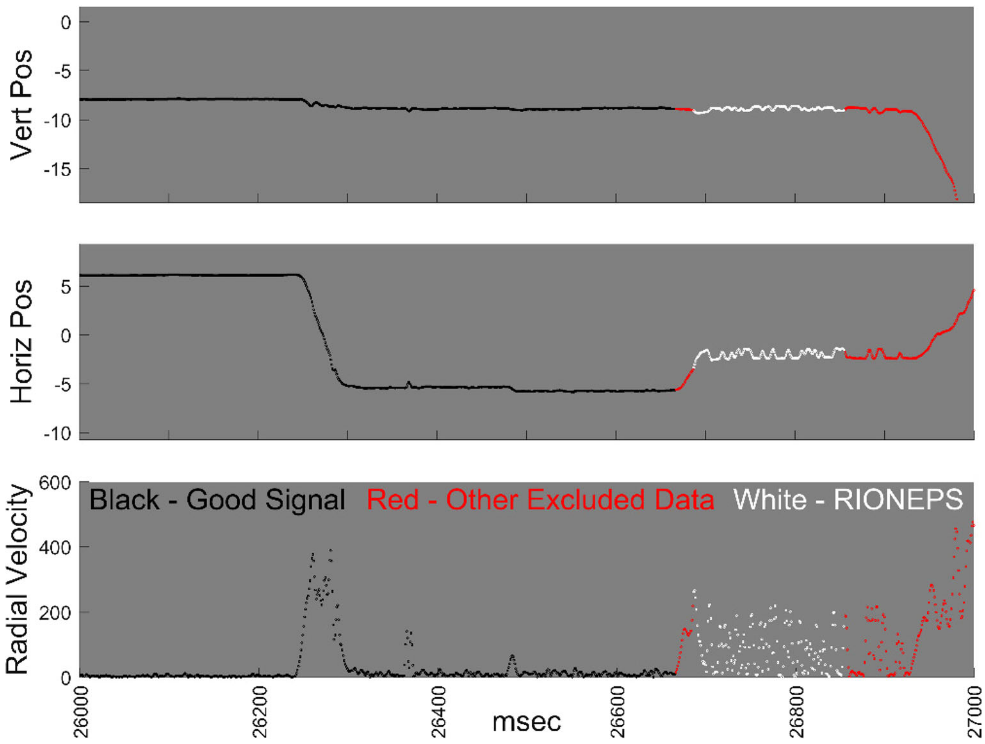

Fig. 1 Illustration of "Rapid Irregularly Oscillating Noise of Eye Positional Signal" (RIONEPS), Subject 8

minimum fixation duration (for ONH: 0.040 s) were excluded from further analysis. To obtain the most stable fixation velocity values, for each remaining fixation block, data were rejected that were within 5 ms of the start and end of each fixation block. This procedure would leave in place only the central portions of each fixation block in which the signal would likely be most stable. The mean (μ) and standard deviation (σ) of the distribution of the velocities during these "fixation periods" were calculated. The updated peak saccade velocity threshold was defined as:

$$\text{Updated Peak Saccade Velocity Threshold} = \mu + 6*\sigma \quad (5)$$

The loop continues to iterate until the difference between the last threshold calculation and the current calculation is less than or equal to 1.0. The saccade subthreshold is calculated as:

$$\text{Saccade Subthreshold} = \mu + 3*\sigma \quad (6)$$

This approach assumes that the velocities during these potential fixation periods are reasonably normally distributed, or at least unimodal and symmetric. With our data, this

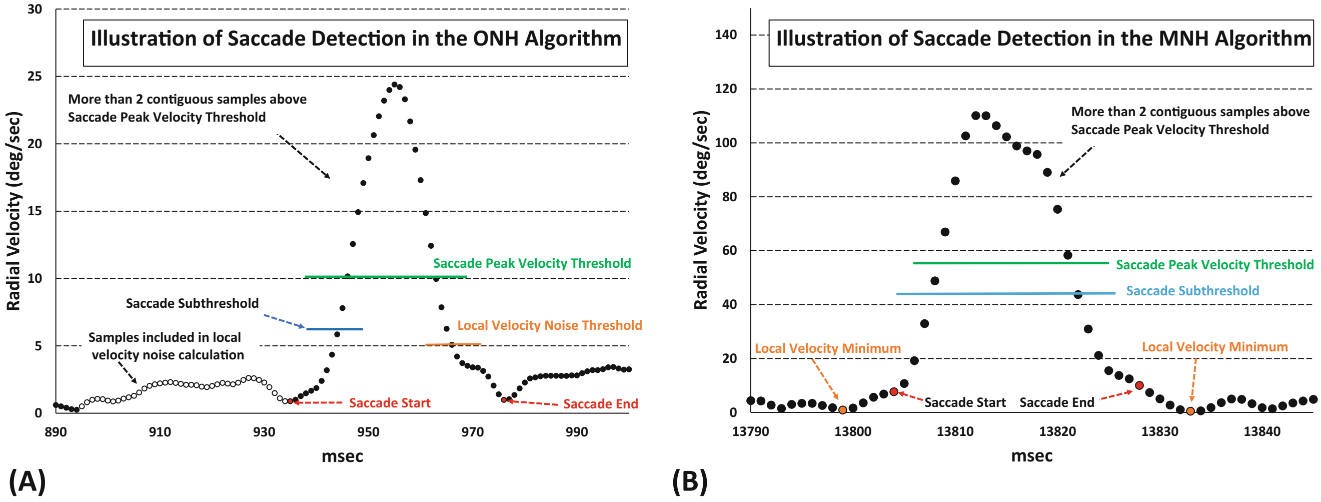

Fig. 2 Illustrations of saccade detection in the ONH (A) and in the MNH (B)

assumption was completely unsupported. These distributions were highly skewed. We hypothesize that any such distribution, from any similar dataset, would be similarly skewed. A measure of skewness is kurtosis. For a normal distribution, the kurtosis should be near 3.0. Below we present the kurtosis for each of the 20 datasets employed in this study. The median kurtosis was 7.56. Also, tests of normality (Pearson chi-square goodness-of-fit tests) indicated that these distributions were not normal (all *p* values were essentially 0.0, meaning that we can reject the null hypothesis that the distribution was drawn from a normal population). In Fig. 3, we present the distribution of the velocities in question for the subject closest to the median kurtosis of the set.

The mean is a poor choice to represent the central tendency of such a distribution. The median is a better choice. The standard deviation is likewise a poor choice to represent the spread of such a distribution. Some sort of percentile measures (e.g., the interquartile range or something similar) would be a better choice. We believe that the failure of the ONH algorithm to properly characterize the central tendency and spread of these distributions is responsible for a number of classification errors produced by the ONH.

To illustrate the variability in these adaptive thresholds and the kurtosis from subject to subject, we present the adaptive saccade thresholds and kurtosis for the 20 subjects in the present study in Table 2. As a general matter, the saccade peak velocity thresholds computed for our data are in the range of 10°–20°/s, which is substantially lower than the 33°/s reported by Nyström and Holmqvist (2010). Our data are substantially less noisy than the data they report on. Note that this is true even though we employed a much shorter, and thus potentially noisier, smoothing window for the calculation of velocity. Note the high intersubject variability in these thresholds, as well as the extremely high thresholds and the extremely high measure of kurtosis for one subject.

Description of the general approach to finding saccades in the MNH The general approach to finding saccades in the MNH is illustrated in Fig. 2B. The saccade peak velocity threshold was fixed at 55°/s, and the saccade subthreshold was fixed at 45°/s. (We question the need for the saccade subthreshold, but have not studied its role sufficiently to decide whether it could be dropped without affecting classification performance.) There must be two samples above the saccade peak velocity threshold for a set of samples to be labeled as the peak portion of a saccade. To find the start of a saccade, the algorithm steps backward, one sample at a time, to see if the velocity crosses below the saccade subthreshold. If the samples do cross this threshold, then the algorithm steps backward, one sample at a time, until a local velocity minimum is found. We have found that using the velocity local minimum as the start of a saccade makes sense if one is trying to classify the velocity trace, but if one is trying to classify the position trace, the result of using the local velocity minimum is that many, if not most, saccades will appear to start too early in the position trace. Therefore, after we have found the local minimum, we step forward a variable number of samples (depending on the number of samples between the point at which a saccade passes below the saccade subthreshold and the point of local minimum) to mark the start of each saccade. In the case of Fig. 2B, we stepped forward five samples. To find the end of a saccade, the velocity must cross under the saccade subthreshold, and then, moving forward one sample at a time, the local velocity minimum is once again found. If the saccade is followed by a PSO, then this local velocity minimum is the saccade end point. If the saccade is not followed by a PSO, then we step backward a variable number of samples (also depending on the number of samples between the point at which a saccade passes below the saccade subthreshold and the point of local minimum) to mark the end of the saccade. In the case of Fig. 2B, we stepped back five samples.

Detecting PSOs with the ONH algorithm The ONH algorithm for detecting PSOs is illustrated in Fig. 4A and B. The ONH defines two types of PSOs, "weak" (Fig. 4A) and "strong" (Fig. 4B). The ONH searches for evidence of a velocity peak from the end of the saccade to 40 ms (minimum fixation duration) after the end of the saccade. This does not mean that the entire PSO must be completed in this window, just that a potential PSO is not a PSO if it does not meet the peak velocity criteria in this window. Weak PSOs occur when the velocity during this period crosses above the local saccade velocity threshold. Strong PSOs occur when the velocity during this period crosses above the saccade peak velocity threshold. If the velocity during this period does not cross above the local saccade velocity threshold, then no PSO is present. Each time the signal crosses above and below this threshold, one potential weak PSO is defined. The velocity during this period can cross above and below the local saccade velocity threshold more than once, as in Fig. 4A. In this case, the event is labeled as a weak PSO, with two weak PSO peaks. The end of the PSO is the first local velocity minimum after the velocity crosses below the local saccade velocity threshold for the last time within this 40-ms window. This local velocity minimum must occur within 80 ms of the end of the previous saccade.

In Fig. 4B, we illustrate a strong PSO. Note that the velocity during the peak window crosses above and below the saccade peak velocity threshold, and subsequently crosses above and below the local saccade velocity threshold twice. This event would be labeled a strong PSO, with two weak PSO peaks and one strong PSO peak. Note that the end point of this PSO occurs after the 40-ms peak PSO window.

For both types of PSOs, the event is rejected if the peak velocity during the 40-ms peak PSO window is greater than the peak velocity of the prior saccade, if the local minimum PSO end occurs more than 80 ms after the end of the prior saccade, or if artifact data are present within the 80-ms window.

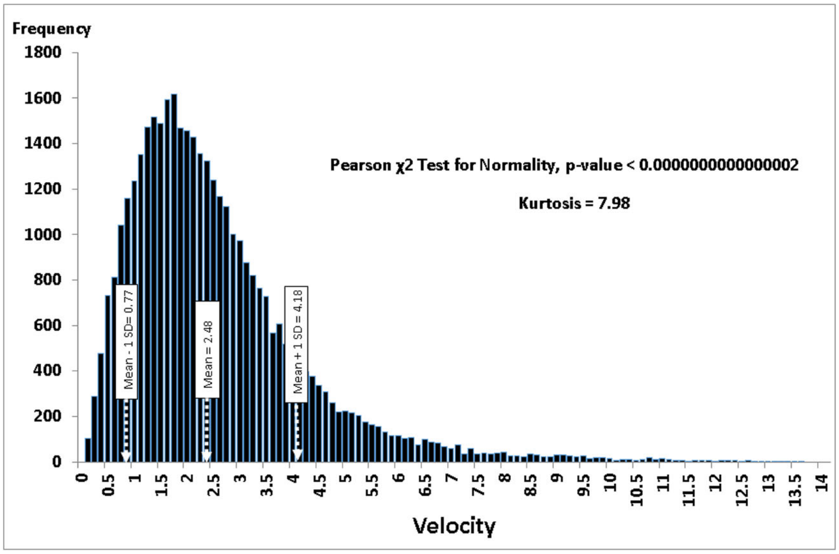

Fig. 3 Distribution of velocities during potential fixation periods

Detecting PSOs with the MNH algorithm Figure 5 illustrates the PSO detection steps for the MNH algorithm. The start of every potential PSO is the sample following the end of the saccade. To find the end of each PSO, we search for five consecutive points that are all below the 90th percentile of the velocity noise distribution during fixation (defined in Step 5 above). The first of the five consecutive points is the PSO end. PSOs are classified as small, moderate, and large. If the velocity during a PSO crosses above the saccade peak velocity threshold, then the PSO is large. If the velocity during a PSO crosses above the saccade subthreshold and below the saccade peak velocity threshold, then the PSO is moderate. If

Table 2 Saccade detection thresholds for 20 datasets scored by the ONH

	Subject ID	Saccade Peak Velocity Threshold	Saccade Subthreshold	Mean Local Velocity Saccade Threshold	Kurtosis
	1	10.06	5.95	5.26	8.71
	35	10.70	6.40	5.84	6.48
	23	10.93	6.49	5.90	7.12
	2	11.13	6.64	5.99	7.13
	17	12.41	7.57	7.24	5.25
	34	12.54	7.58	7.07	6.09
	29	12.65	7.57	6.83	7.98
	31	12.80	7.63	6.86	6.65
	10	14.39	8.55	7.73	8.82
	3	14.58	8.68	7.99	8.45
	12	14.70	8.76	7.92	9.38
	19	15.51	9.44	8.93	5.89
	33	15.59	9.22	8.20	8.45
	14	15.82	9.58	8.99	6.30
	37	15.86	9.26	8.15	8.38
	40	17.10	10.02	8.70	8.57
	38	17.78	10.89	10.31	4.83
	26	21.76	12.89	11.54	8.11
	13	24.15	14.58	13.55	6.97
	22	102.34	55.08	40.49	12.19
25th percentile		12.51	7.57	6.85	6.44
Median		14.64	8.72	7.96	7.56
75th percentile		16.17	9.69	8.945	8.48

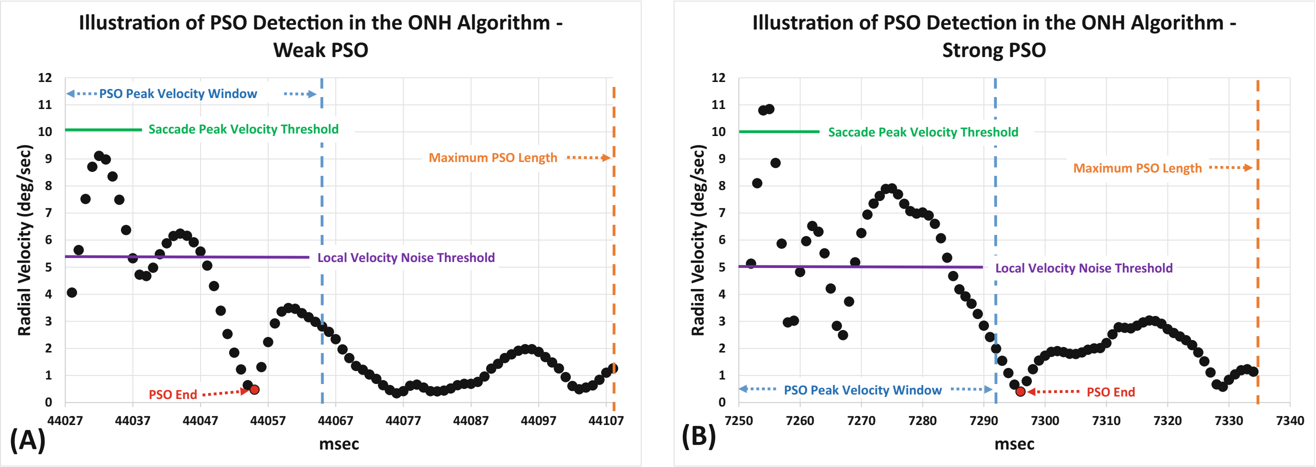

Fig. 4 Illustration of PSO detection [“weak PSO” (A) and “strong PSO” (B)] in the ONH algorithm

the velocity during a PSO crosses above the small PSO velocity threshold (fixed at 20°/s) and does not cross the saccade subthreshold, then the PSO is small.

Detection of fixation with the ONH Nyström and Holmqvist (2010), in their Table 1 (p. 191) state that potential fixations are defined by samples that are neither saccades, PSOs, nor noise. However, this is not how the code is written. In the code, all samples not considered saccades or PSOs are considered to be potential fixations. The presence of noise (artifact) is not considered. Potential fixations that do not include more samples than the minimum fixation duration (40 samples of 1 ms each, with the ONH at 1000 Hz) are left unclassified. If the peak velocity during the fixation is greater than the saccade peak velocity threshold, then the event is also left unclassified. Finally, if a potential fixation contains any artifact, the event is left unclassified. All potential fixations passing these filters are considered to be true fixations.

Detection of fixation with the MNH All samples not considered saccades or PSOs or artifacts are considered to be potential fixations. Note this important difference between the two algorithms: For the ONH, samples that are not part of saccades or PSOs are considered potential fixations, whereas for the MNH, samples that are not part of saccades, PSOs, or *artifacts* are considered potential fixations. In this way, in the MNH, periods of fixation can be either preceded or followed by artifacts. This allows for many fixation periods, missed by the ONH, to be classified as such by the MNH. Since the ONH rejects any potential fixation period in which an artifact occurs, this precludes the possibility that long periods of true fixation might precede or follow an artifact. In our view, this

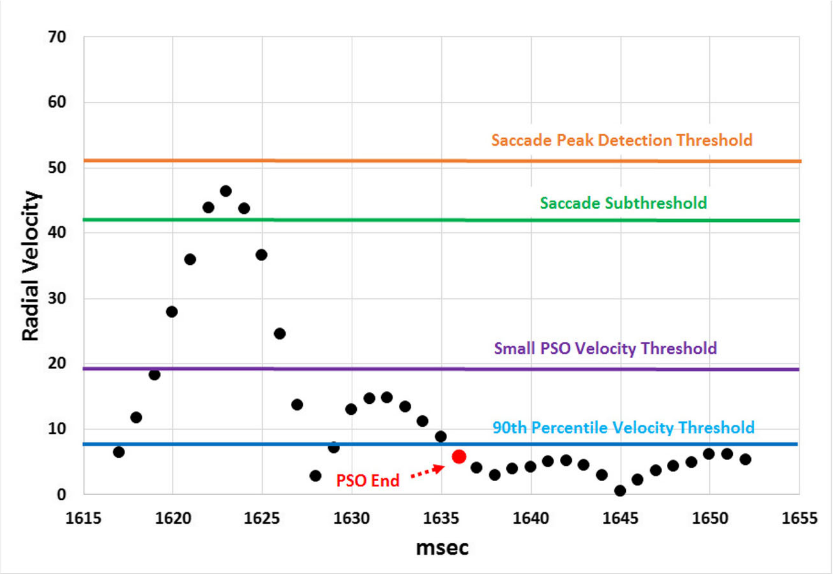

Fig. 5 Illustration of PSO detections in the MNH algorithm

exclusion is not justified, since the MNH scores many periods that look like perfectly good fixation periods as fixations, whereas the ONH leaves such periods unscored. Below we will see examples of fixations not detected by the ONH for this reason. For the MNH, potential fixations that do not have more samples than the minimum fixation duration (30 samples of 1 ms each, with the MNH at 1000 Hz) are unclassified. We chose 30 ms as the minimum fixation duration, as opposed to the 40-ms minimum employed by the ONH. We made this choice after inspecting fixations in the range of 30–40 ms in duration. All the fixations we observed in this range appeared to be valid fixations. The maximum difference between any two points in the potential fixation block must be smaller than 2 deg, for both the horizontal and vertical position signals. All potential fixation blocks that pass these criteria are considered true fixations.

For a list of all the parameters set for the MNH, see Table 3.

Error classification guidelines

As we mentioned above, we had a hierarchy of error types, based on the order in which the original Nyström and Holmqvist (2010) algorithm works. In that algorithm, artifactual events (often referred to as "noise") are first classified as such. Then saccades are classified, followed by PSOs, if they are present. Everything else is classified as fixation. So, our guidelines state the following:

> In what follows, "misclassification" refers to both misclassification and mistiming. Any misclassification error that can be attributed to noise misclassification should be scored as noise misclassification. After deciding if there are any noise misclassifications, any misclassification that can be attributed to saccade misclassification should be attributed to saccade misclassification. Any misclassification that cannot be attributed to noise misclassification or saccade misclassification, but can be attributed to PSO misclassification should be attributed to PSO misclassification. Any misclassification that cannot be attributed to noise misclassification or saccade misclassification, or PSO misclassification should be attributed to fixation misclassification. With this hierarchy, some of the 32 classification failures should be relatively rare, especially timing errors. We do not expect many, or any, errors of the type "PSO starts too early" or "PSO starts too late," since these would, under normal circumstances, be classified as "saccade ends too early" or "saccade ends too late." Further, under normal circumstances, given this hierarchy, we do not expect any fixation timing errors.

We defined a "saccade that starts too early" or that "ends too late" as a saccade whose timing was off by more than three samples (i.e., 3 ms) from the human expert judgement. The same applies to saccades that end too late.

There is also a special rule for short, unclassified fixation periods that can occur amid an artifact. In the original ONH (Nyström & Holmqvist, 2010) publication, a fixation needed to be at least 40 ms in duration. Therefore, any fixation period that is unclassified, that is shorter than 40 ms, was not classified as "fixation not detected."

We did not classify the first event in any recording, since neither algorithm would have sufficient data to make a classification at this starting stage.

Finally, especially for the ONH, there were many unclassified periods that probably should have been classified as an artifact. We did not score this error. Of course, any unclassified signal that looked like a fixation, saccade or a PSO was classified as an error.

We defined a saccade according to the appearance of the position channels. We believe that there is general agreement about what the profile of a saccade in the position signal should looks like. PSOs were defined on the basis of their appearance in the velocity trial. A PSO must:

Table 3 Fixed parameters used in the MNH

Parameter	Value
Length of Savitzky–Golay filter window	0.007 s
Velocity above which data are rejected as noise	1500°/s
Acceleration above which data are rejected as noise	100,000°/s^2
Initial temporary threshold for fast determination of noise distribution during fixation.	100°/s
RIONEPS threshold	100
Minimum fixation duration	0.030 s
Saccade peak velocity threshold	55°/s
Saccade subthreshold	45°/s
Minimum saccade duration	0.010 s
Minimum threshold for a small PSO	20°/s
Within a single potential fixation, the maximum allowable amplitude difference between 2 points	2 deg

(1) start immediately after a saccade
(2) include any velocity peaks that are contiguous
(3) have a peak velocity less than the previous saccade
(4) have a peak velocity greater than the following fixation noise.

Software for error classification

We employed a custom graphical user interface program to view the vertical and horizontal position traces as well as a radial velocity trace. The program displayed 1 s of data per page. Fixations were shown in red, saccades in green, PSOs in blue, artifacts in white, and unclassified data in black. The program was designed with a built-in magnifying function—if you clicked the left mouse button at any point in the signal trace, the region where you clicked was magnified by a factor of 10 so that individual samples could be discerned. Pushbuttons brought up dialogs to count the number of any error type per page. All the error classification results were written to an output file for further analysis.

Human raters

There were three human raters, all of whom are authors on this report. Rater "L" has been studying eye movements, on and off, for more than 25 years. He has 12 peer-reviewed publications that report on eye movement results. He is also the programmer who created the MNH. Rater "I" has 5 years of research on eye movements, with special focus on biometrics, human-computer interaction, and computer vision. Rater "E" has been working with eye movements for 4 years. He has implemented and tested several classification algorithms, developed novel noise detection methods for eye movement signals, and worked on eye movement biometrics.

Rater training

Prior to classifying the main set of 20 subjects, all raters practiced on four "training subjects," each scored by both algorithms (eight studies). During this period, open discussion between the raters was encouraged. After each rater scored all eight training studies, their results were compared in detail and, in multiple discussions, an attempt at consensus was sought. Then the three raters scored the training set again, and the results and any discrepancies were again discussed. At this point, the raters began independently rating the main data set, and no further interrater communication was allowed.

Statistical analysis

The number of errors for each error type was not normally distributed, so interrater reliability was assessed with the Kendall coefficient of concordance (KCC; Siegel & Castellan, 1988). This statistic evaluates the reliability of the three raters taken together. If the KCC was above .7, the error type was included in the final analysis. Only four error types had a sufficient number of errors (more than a total of 30 across all raters and subjects) and had KCCs above .70. These four error types accounted for approximately 90% of all errors.

With 20 subjects and paired data, we had more than a .8 statistical power to detect effect sizes (Cohen's *d*) of 0.68 and above. Most of the key effect sizes in this report were gigantic (>17.0), but some were in the range of 1.0 or so. Thus, we had adequate statistical power, considering the effect sizes we found.

To compare the error rates between methods, we employed the sign test (Siegel & Castellan, 1988). This is a paired test that converts each comparison between an ONH scoring and an MNH scoring for the same subject (a set) as a series of ones and zeros. If the ONH is greater than the MNH, this is represented as a 1. If the MNH is greater than the ONH, this is represented as a 0. The sign test evaluates the probability of getting X number of ones for Y trials (in this case, Y = 20 trials). Effect sizes (Cohen's *d*; Cohen, 1988) were calculated for these tests by converting from an exact *p* value (15 digits) to a *Z* score. The *Z* score was converted to an equivalent Pearson *r* correlation coefficient, which was then converted to Cohen's *d* (Cooper & Hedges, 1994). Given the precision of the computer system, the maximum *d* that could be estimated was 17.1, a truly giant effect size.

To compare the numbers of events (fixations, saccades, PSOs, artifacts, and unclassified) scored by each algorithm and the lengths of time spent occupied by these events, paired *t* tests were employed. Effect sizes (Cohen's *d*) were computed by transformation of the *t* values (Cooper & Hedges, 1994).

Results: ONH versus MNH

Signal examples showing typical errors

Example recordings are shown in Figs. 6, 7, 8, 9, 10, and 11. Figure 6 illustrates a near total failure of the ONH system. Note the long stretch of saccades that are not detected by the ONH. In this case, there were many such examples in the recording, so we considered this case a failure of classification and completely removed the data from this set for all further calculations. (A "set" consists of one subject scored by two algorithms.) We also found such an example in the training set of four subjects we evaluated. So, our best estimate is that this occurs in two out of 24 subjects, or approximately 8.3% of the time. In both cases of total failure, the distribution of velocity noise during fixation was extremely skewed (kurtosis values = 12.1 and 12.29, the highest kurtosis values observed in this study). This led to saccade velocity peak detection thresholds of either 102.3°/s or 104.0°/s, respectively. In our view, these total failures

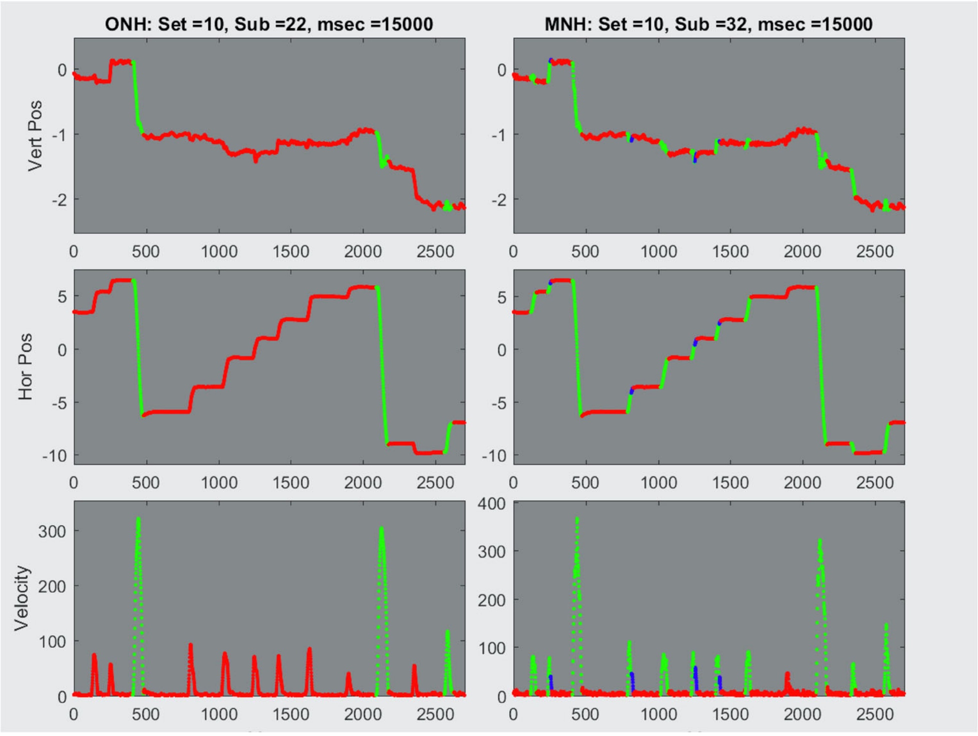

Fig. 6 Horizontal and vertical position traces, plus velocity traces from the same section of eye movement data, classified by two different algorithms (left, ONH; right, MNH). Each true subject is scored twice, once by the ONH and once by the MNH. For purposes of blinding the raters, these two scorings were referred to as different subject numbers. A "set" consists of these two scorings. For example, in this figure true Subject 1,297 is scored by the ONH and referred to as Subject 22, and also scored by the MNH and referred to as Subject 32. Both samples start 15,000 ms into the recording, and the numbers on the *x*-axis are increments from this starting point. In this case, both samples end at about 17,500 ms. The data classified in red are fixations, those in green are saccades, and those in blue are PSOs. Unclassified samples are shown in black and are not present in this example. Artifact samples are shown in white and are also not present here

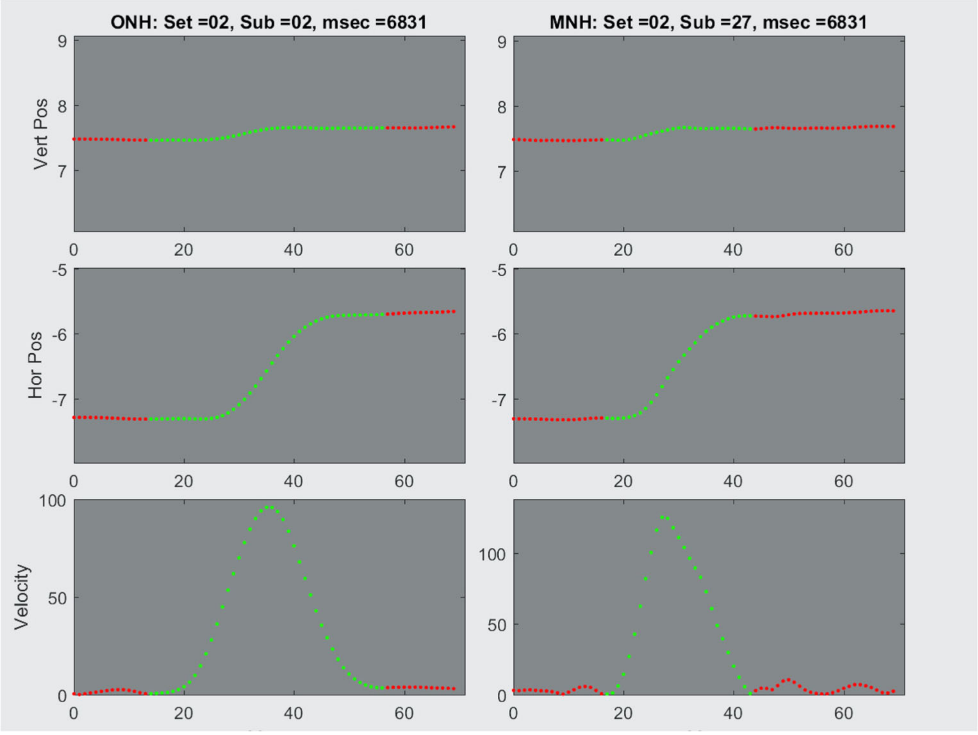

Fig. 7 Illustration of a saccade that starts too early and ends too late according to the ONH. See the caption to Fig. 6 for details on the figure conventions. On the left is a saccade classified by the ONH, and on the right is the same saccade classified by the MNH

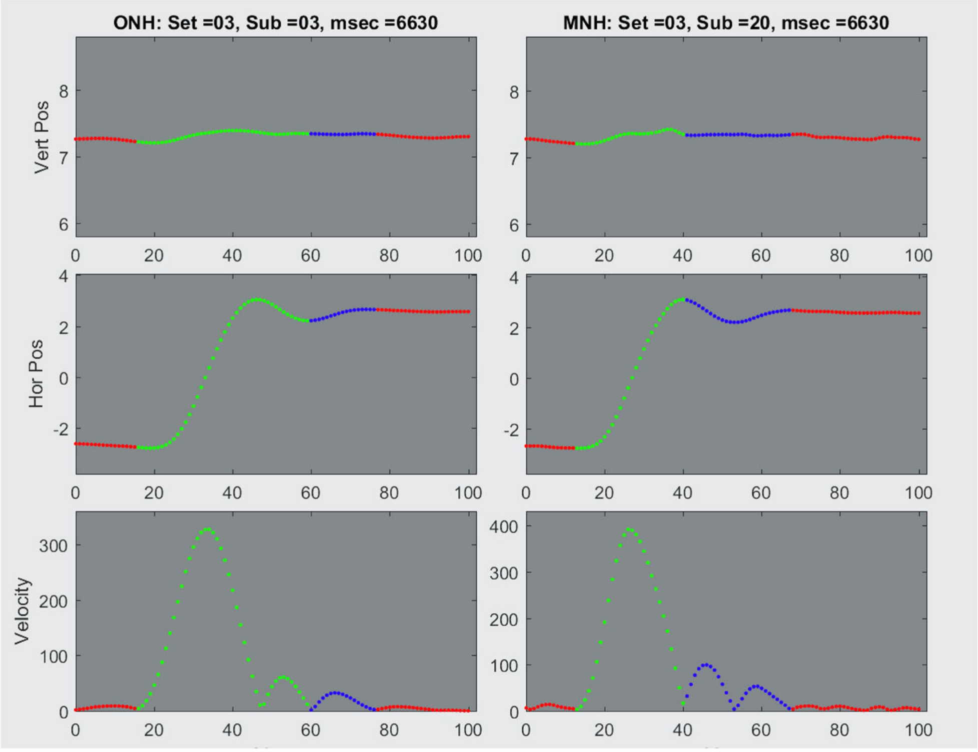

Fig. 8 A saccade that, as scored by the ONH, includes a portion of the subsequent PSO. See the Fig. 6 caption for the figure conventions

stemmed from the fact that the ONH algorithm uses metrics appropriate for a unimodal symmetric distribution, whereas the actual distributions are highly skewed.

Figure 7 illustrates a saccade that starts too early and ends too late as scored by the ONH. Note how early the saccade on the left starts and how late it ends. Although the saccade on the

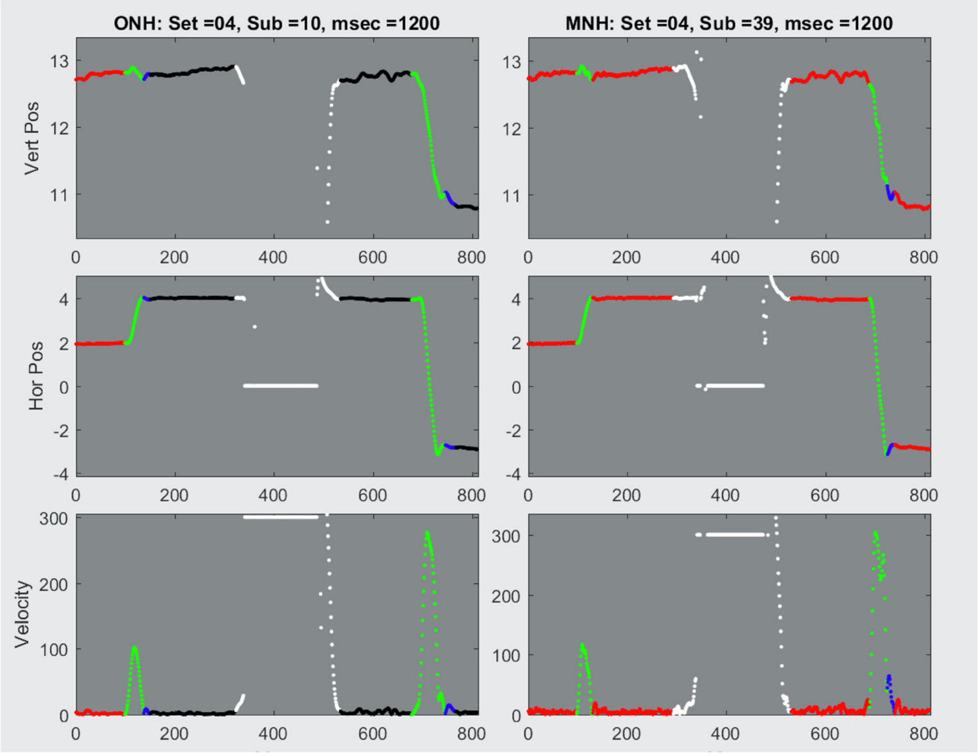

Fig. 9 Different performance of the two algorithms near a blink artifact. See the Fig. 6 caption for the figure conventions

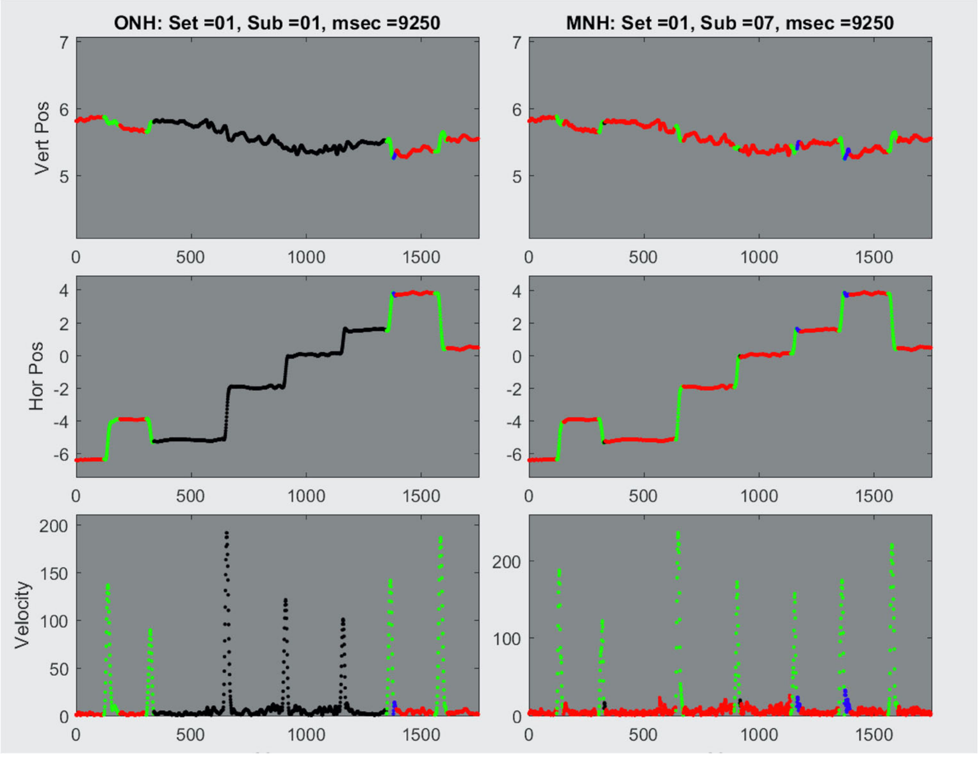

Fig. 10 Examples of saccades not detected by the ONH. See the Fig. 6 caption for the figure conventions

right might start a little too early, it would not have been classified as starting too early, since the true estimate of the start is less than 3 ms different from the detected saccade start determined by the human observer. This was the rule for all timing decisions: Classified events had to be more than 3 ms off from the choice a human expert might make. In this figure,

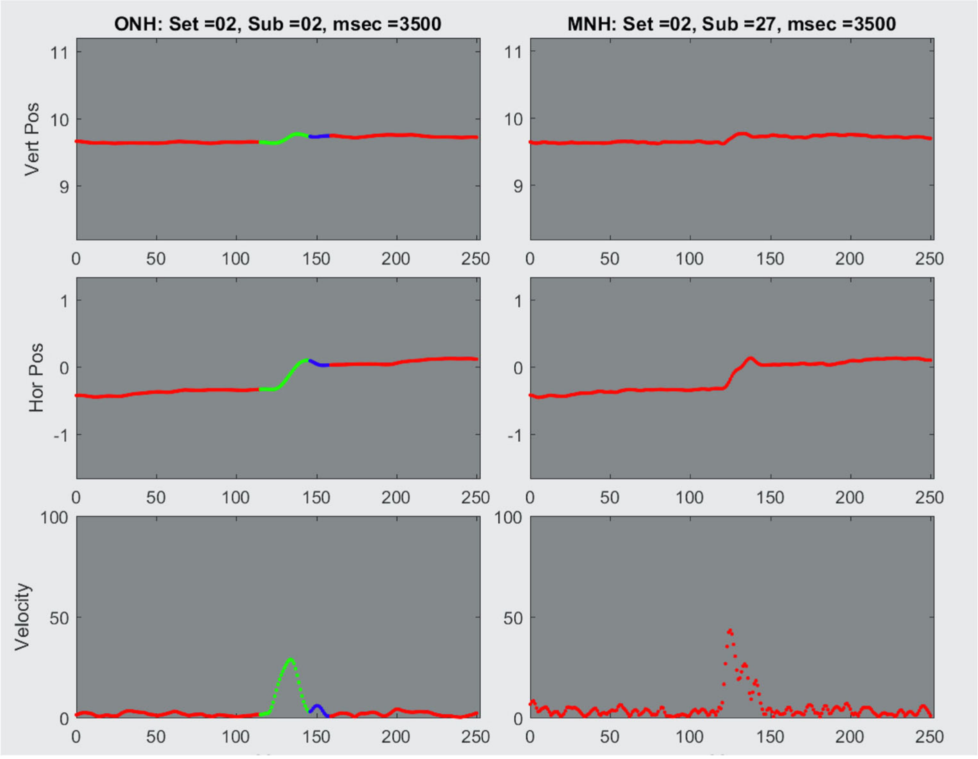

Fig. 11 A small saccade (0.43 deg) that is correctly classified by the ONH but missed by the MNH. See the Fig. 6 caption for the figure conventions

you can also see that the more sensitive and responsive velocity trace is for the MNH as compared to the ONH. There were many, many such saccades, especially when the recordings were classified by the ONH. These two errors ("saccade starts too early" and "saccade ends too late") were, by far, the most frequent of all errors. We suggest that these errors occur because the saccade starts and ends were defined by the ONH as the local velocity minima. Although these points fit the velocity data well, in the position trace this mark of the start of a saccade starts too early, and this mark of the end of the saccade ends too late.

In Fig. 8, we show a saccade that, when scored by the ONH, ends so late that the saccade includes a portion of the subsequent PSO. Here we see a saccade that is classified as ending too late for the ONH algorithm but does not end too late when classified by the MNH algorithm. Note that in this case the ONH-classified saccade includes part of the adjacent PSO, but the MNH-classified saccade ends before the PSO. This occurs because the PSO immediately following the saccade begins at a velocity above the saccade subthreshold (in this case, 8.7°/s). We think that this is an unreasonably small saccade subthreshold that is produced by the adaptive methods for the ONH. The MNH algorithm, which uses more reasonable and fixed thresholds for saccade detection, has no problem recognizing the PSO as a PSO.

In Fig. 9, we show an example of the different performance of the two algorithms near a blink artifact. The EyeLink 1000 automatically detects blinks and replaces the data during the blink with NaN values. For display purposes, during blinks, we set the horizontal and vertical position to 0.0, and set the velocity to 300. These blink artifacts are shown in white. Both algorithms also typically exclude some data before and after such blinks. These periods are also shown in white. Note how, for the ONH, the fixation periods before and after an artifact period are not scored. These fixation periods are properly scored with the MNH. The ONH defines the set of all potential fixations as those periods that are not saccades and not PSOs. Therefore, the stretch of the recording that starts near 1,350 ms and ends near 1,875 ms is not considered as a potential fixation period because it contains an artifact. The MNH defines the set of all potential fixations as those periods that are not saccades and not PSOs and *not artifacts*. Therefore, the periods from 1,350 to 1,520 and from 1,730 to 1,875 are considered as potential fixations. We think that these are fixation periods and should be counted as such.

In Fig. 10, we see that the ONH has not only not classified fixation periods near an artifact but also has not classified many obvious saccades. The MNH classifies these data correctly. The problem here is really the problem of allowing adaptive thresholds to obtain such small values. The saccade peak velocity threshold for the ONH scoring on the left is 10.06, which is the smallest in the study and, to us, is unreasonably small. In the case of the three saccades that are not scored by the ONH in the middle of this figure, these saccades are rejected because the local velocity noise (11.20, 14.96, and 12.94, respectively) is greater than the saccade peak velocity threshold. These are clearly saccades, and their rejection is a by-product of the adaptive estimation of the saccade peak velocity threshold. The MNH uses a fixed saccade peak velocity threshold of 55°/s and scores these events accurately.

In Fig. 11 we show a small saccade (0.43 deg) that is correctly classified by the ONH algorithm but is missed by the MNH algorithm. Since the MNH uses a fixed velocity threshold (55°/s) for the peak of a saccade, this saccade was not detected.

Results of the error classification

Many errors occurred so infrequently that they were not analyzed ("noise misclassified as fixation," "noise not detected," "noise starts too late," "noise ends too early," "noise ends too late," "noise starts too early," "noise misclassified as PSO," "fixation ends too late," "fixation ends too early," "fixation misclassified as noise," "fixation starts too early," "fixation starts too late," "fixation misclassified as saccade," "saccade misclassified as PSO," "saccade starts too late," "saccade misclassified as noise," "PSO misclassified as noise," "PSO misclassified as saccade," "PSO starts too early," and "PSO starts too late"). The total number of errors is a sum of each error for all raters and sets (pairs of subjects). We considered that if there were fewer than 30 errors (across 20 subjects scored by both the ONH and the MNH), then there was insufficient data to perform a statistical analysis.

A number of errors occurred with sufficient frequency but were not rated reliably ("noise misclassified as saccade," "fixation misclassified as PSO," "saccade ends too early," "saccade misclassified as fixation," "PSO misclassified as fixation," "PSO not detected," "PSO ends too early," and "PSO ends too late"). For these errors, at the least, more specific guidelines and rounds of practice, including consensus discussions, would be required in order to improve the reliability of error detection. It is also possible that human raters might not be able to classify some of these events reliably, regardless of practice.

Table 4 lists the error types that account for many of the errors and that were detected reliably. Although there are only four such error types, they account for nearly 90% of all errors. All have very high interrater reliability (>.70) as assessed by the KCC. The scale run runs from 0 to 1.0, and anything above .7 is considered good to excellent reliability.

Figure 12A presents the rates for the error "saccade starts too early." For all the boxplots in Fig. 12, the heavy horizontal lines

Table 4 Errors that were detected reliably

Count	Error Number	Error Names	Total Errors	Percentage of All Errors	Kendall Coefficient of Concordance*
1	12	Saccade not detected	105	0.93	.75
2	8	Fixation not detected	711	6.29	.88
3	25	Saccade starts too early	5,914	52.33	.94
4	28	Saccade ends too late	3,420	30.26	.95
		Sum	10,150	89.82	

* Kendall coefficient of concordance comparing all raters

are at the median, and the diamond shapes represents the mean. A few extreme outliers are not shown, so as to enhance clarity, but all values were entered into the statistical evaluation (sign test) and the estimation of effect size. The effect sizes (Cohen's *d*) for these comparisons are either large or gigantic, considering that an effect size of 0.8 is, in many circumstances, considered "large" (Cohen, 1988).

The "saccade starts too early" error type accounted for 52% of all errors. The median number of saccades that start too early is always above 75 for ONH and less than 26 or so for the MNH. Given the very high numbers of saccades that start too early when scored by the ONH, we wondered what percentage of all saccades were scored with this error. These results are presented in Fig. 12B. A median of 78% of all saccades started too early for the ONH, whereas for the MNH this result was 19%. For all raters, more than 70% of all saccades started too early for the ONH, and fewer than 30% for the MNH.

Figure 12C presents the data for the error "saccade ends too late." There were many more of these errors when scored by the ONH than when scored by the MNH. These errors were 30.26% of all errors. The differences between the ONH and MNH are great and highly statistically significant, and the effect sizes (*d*) are huge. The median number of saccades that end too late is above 40 for all raters for ONH, but below 20 for MNH. Some of these saccades that ended too late included some of the adjacent PSO. The numbers of this error were rated by a single rater (rater "L") after the main ratings had been finished (Fig. 12D). Significantly more such saccades of this type were found for the ONH than for the MNH ($p <$ .002).

Figure 12E presents the data for the error "fixation not detected." These represent 6.29% of all errors. An average of approximately ten fixations were not detected per recording for the ONH. This error almost never occurred for the data scored with the MNH, but many more fixations went undetected when scored by the ONH than when scored by the MNH. A substantial number of these occurred near an artifact. The ONH leaves unclassified what we would consider perfectly good fixation periods before and after artifacts.

Figure 13F presents data for the error "saccade not detected." These accounted for only 0.93% of all errors. More saccades were not detected by the ONH than by the MNH, but this difference was only statistically significant for rater "E." Even in the worst case, the median number of undetected saccades for the ONH was two. So, this error type is not important.

Comparisons of numbers of fixations, saccades, PSO, and unclassified events and total lengths of time spent in each event type for both systems

Figure 13A and B compare the numbers of periods of fixation as scored by both algorithms (A) and the amounts of time spent in fixation (B). The means and *SD*s for these calculations are in Table 5. The MNH scores significantly more fixations and results in significantly more time spent in fixation. Figure 13C and D compare the numbers of saccades as scored by both algorithms (C) and the time spent in the saccades (D). The MNH scores significantly fewer saccades and results in significantly less time making saccades. Figure 13E and F compare the numbers of PSOs as scored by both algorithms (E) and the time spent in making PSOs (F). The MNH scores significantly more PSOs, but the lengths of time spent making PSOs were not significantly different between the two algorithms. Figure 13G and H compare the numbers of unclassified periods as scored by both algorithms (G) and the time spent in unclassified periods (H). The MNH has significantly fewer unclassified events and spends significantly less time in unclassified periods. The MNH spends 1% of the time that the ONH does in unclassified events (Table 5). Figure 13I and J compare the numbers of artifact periods as scored by both algorithms (I) and the time spent in artifact periods (J). The two algorithms are not statistically different in terms of the number of artifact periods, but the MNH spends significantly more time in artifact periods.

It is interesting to note that for the ONH, 49% of all saccades were followed by PSOs, whereas for the MNH, 61% of all saccades were followed by PSOs (Table 5).

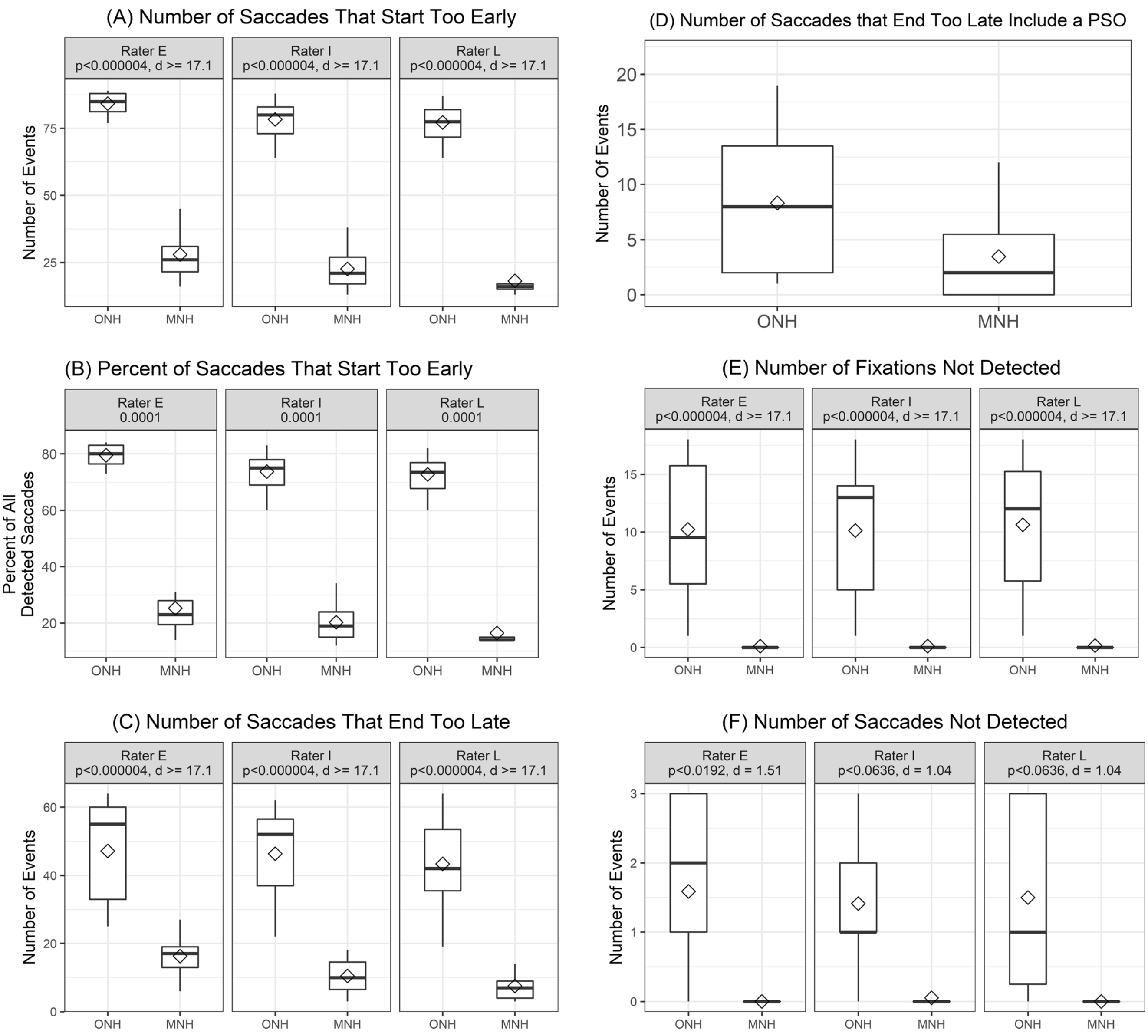

Fig. 12 Boxplots of error types for each rater

Experiment 2: IRF[3]

Method

Raimondas Zemblys kindly provided classification of our data by his algorithm (Zemblys et al., 2018), for which we are most grateful. As we noted above, the algorithm is based on machine learning, but Zemblys's article claims that some logical postprocessing of the machine-learning results is also performed as part of the IRF algorithm. These postprocessing steps are listed on the eighth page of that report.

A single rater (first author, L.F.) evaluated the same 20 recordings as in Experiment 1, as classified by the Zemblys et al. (2018) algorithm (the "IRF" algorithm). Since this evaluation occurred after the ONH–MNH comparison, the rater was not blind to the classification method. Reducing the evaluation to one versus three raters would lower the generalizability of the results to a potential population of raters. Also, with only one rater, the issue of interrater reliability was not relevant. All of the single rater's evaluations of every type of error in the ONH and MNH algorithms were compared to all of the single raters' evaluations of the IRF algorithm.

[3] This section is abbreviated due to space limitations. In particular, the tables and figures illustrating the results have been removed. For the complete section including tables and figures, see Full_Report_On_the_Classification_of_the_Zemblys_etal_2017...docx at https://digital.library.txstate.edu/handle/10877/6874.

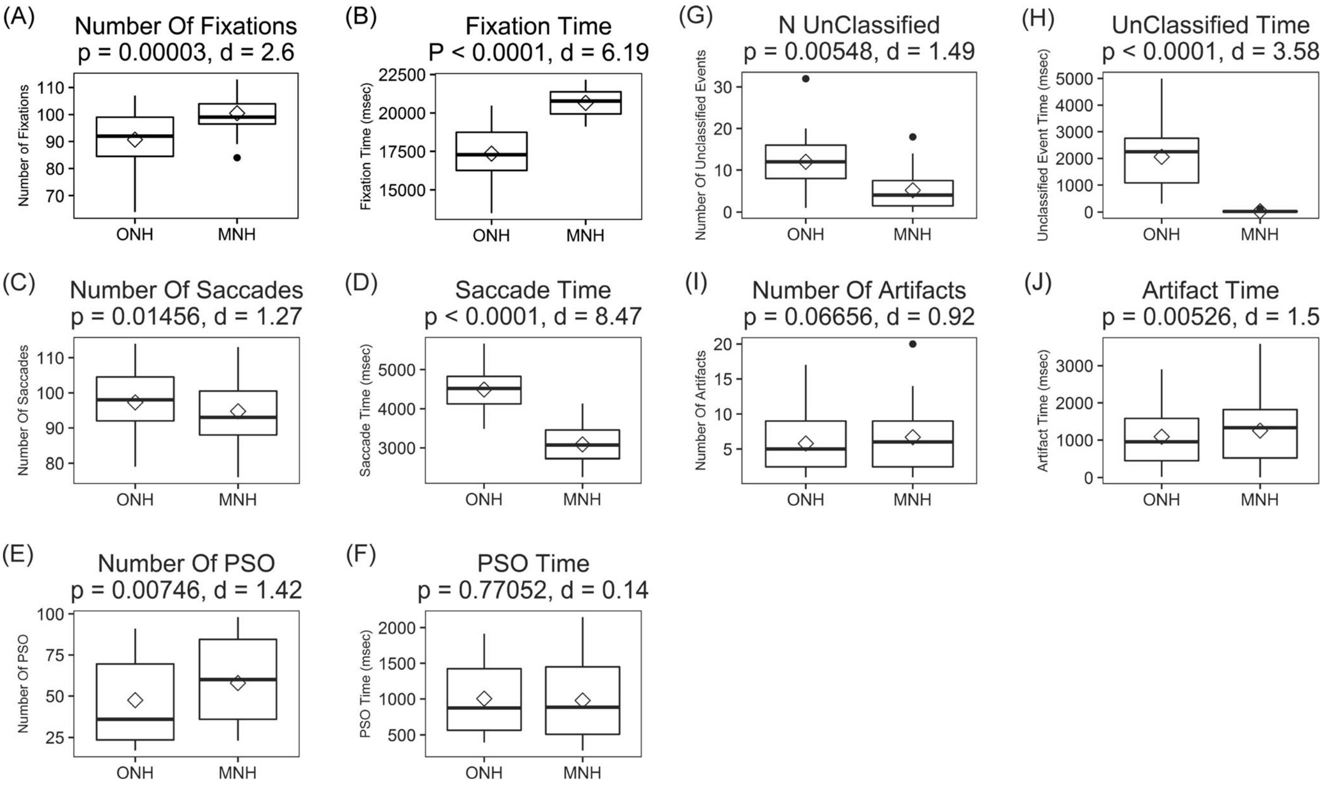

Fig. 13 Boxplots of basic statistics

The statistical significance of differences in error numbers and corresponding effect sizes were calculated as above for Experiment 1.

Results: IRF[4]

General impressions

There are several general impressions that one has after reviewing the classification results from the IRF. First, on the positive side, this algorithm is amazingly accurate for saccade timing. It is a remarkable fact that we did not find a single saccade with an onset timing that was either too soon or too late. And the offset timing of saccades, although not as perfect as the onset timing, was extremely good as well. For saccade timing, the IRF wins decisively. This will clearly be borne out when we review the evaluation results.

The second aspect that one notices is that this algorithm does a very poor job of rejecting unusual or artifactual events. As we noted above, there are periods, typically during blinks, when the EyeLink 1000 does not return a position value, but rather indicates missing data ("not a number," or NaN). The IRF interpolates across such blinks, treats the interpolated data as if they were good data, and attempts to classify such periods like all the good data in the recording. During postprocessing, the IRF removes the longer blink periods from classification. For our evaluation of the IRF, we simply declared all of these blink periods as noise/artifacts. Both the ONH and the MNH exclude some data before and after each blink ("peri-blink" data). Although the IRF claims to do this during postprocessing, our results indicate that no such postprocessing step was actually conducted (see below). The IRF results are severely contaminated because the IRF attempts to classify these peri-blink recording periods. We have many cases in which the IRF classifies what is obviously noise as fixation, saccade, or PSO. If these types of artifactual events are not handled properly, we see events in the midst of noise/artifact, where noise is classified as a saccade of 1-ms duration. These 1-ms "saccades" can have extremely high velocities (700°/s and above). Again, this is not supposed to happen with the IRF. Such artifactual "saccades" severely distort main sequence relationships. Finally, noise also affects the classification of fixations. With the IRF, we found six fixations that were less than 10 ms, even though the postprocessing steps claim to remove fixations less than

[4] This section is abbreviated due to space limitations. In particular, the tables and figures illustrating the results have been removed. For the complete section, including tables and figures, see Full_Report_On_the_Classification_of_the_Zemblys_etal_2017...docx at https://digital.library.txstate.edu/handle/10877/6874.

Table 5 Means (*SD*) for event counts and lengths of time

Measure	ONH		MNH		Ratio MNH/ONH
	Mean	*SD*	Mean	***SD***	
Number of fixations	90.68	10.95	100.47	7.72	1.11
Number of saccades	97.26	9.37	94.74	10.08	0.97
Number of PSOs	47.47	25.94	57.89	25.85	1.22
Number of artifacts	5.79	4.04	6.68	5.06	1.15
Number of unclassified events	12.00	7.19	5.16	4.79	0.43
Length of time spent in fixations (s)	17.36	1.91	20.65	0.94	1.19
Length of time spent in saccades (s)	4.49	0.60	3.09	0.49	0.69
Length of time spent in PSOs (s)	1.00	0.51	0.98	0.59	0.97
Length of time spent in artifacts (s)	1.09	0.85	1.25	0.98	1.15
Length of time spent in unclassified events (s)	2.06	1.16	0.03	0.03	0.01
Derived Statistics:					
Percent of saccades followed by PSO	48.81		61.10		1.25
Length of average fixation (ms)	191.42		205.51		1.07
Length of average saccade (ms)	46.15		32.61		0.71
Length of average PSO (ms)	21.16		16.91		0.80
Length of average artifact (ms)	188.58		187.69		1.00
Length of average unclassified event (ms)	171.38		5.72		0.03

50 ms. No such short fixations were noted for the ONH or the MNH. There were also a number of fixation periods with extremely high velocity samples with the IRF.

Also, the IRF has several problems classifying PSOs. For example, the IRF classifications have PSOs occurring after fixations, whereas, by definition, PSOs occur after saccades only. The IRF is supposed to prevent this, but apparently this postprocessing step also failed. Furthermore, there are some extremely short PSOs (even 1-ms PSOs) when classified by the IRF.

Likewise, many of the PSOs scored by the IRF would not meet our criteria for PSOs (see above). PSOs scored by the IRF often have no velocity peaks that are above those seen in the surrounding random fixation noise. Our criteria are similar to those used for the ONH, as indicated by the algorithm used to detect PSOs with the ONH and also with the example figures (their Figs. 1 and 9) illustrating PSOs in the original ONH article (Nyström & Holmqvist, 2010). Indeed, the IRF classifies more than twice as many PSOs as the ONH (2,210 vs. 1,051)—an increase of 110%.

Error classification for the IRF versus the ONH and the MNH

For the ONH–IRF comparison, the ONH had higher numbers of errors than the IRF for four error types, and the IRF had more errors than the ONH for four other error types. For the MNH–IRF comparison, the MNH has significantly more errors than the IRF for one error type, and the IRF had more errors than the IRF for six error types.

Given the absence of methods to handle unusual or artifactual events in the IRF, many more noise periods were classified as fixations by this method than by the ONH and MNH methods. The IRF also had more noise periods classified as saccades than the MNH, but not than the ONH. Many more fixation periods were misclassified as PSOs for the IRF than for the ONH or, more dramatically, for the MNH. The IRF had fewer fixations and PSOs that were not detected than the ONH. The IRF had no saccades that started too early, whereas the MNH and, especially, the ONH had many more such events. Although the numbers are small, the IRF had more saccades that ended too early than either the ONH or the MNH. The IRF has fewer "saccades that end too late" errors than either of the other algorithms, although the effect was statistically significant only versus the ONH. The IRF has significantly more PSOs that end too early than either other algorithm. The IRF also has more PSOs that end too late than both algorithms, but this increase was statistically significant versus the MNH only.

Experiment 3: EyeLink Parser

Method

The EyeLink Parser algorithm is an online, real-time classifier, which classifies eye movement data into fixations and saccades. It does not classify PSOs. This may or may not be an issue for online, real-time users. However, it is frequently employed offline for event classification, and for this task, not detecting PSOs is a real drawback, in our view. As we were informed by SR Research staff, the algorithm works as follows: Under the standard parser configuration, saccade onset is signaled when either velocity or acceleration go above thresholds of 30 deg/s and 8,000 deg/s^2, respectively, and the eye has traveled at least 0.1 deg. The velocity threshold is adjusted by an average velocity from the preceding 40 ms in order to ensure reliable detection of saccades during smooth-pursuit eye movements. Saccade offset is signaled when velocity drops below a threshold. Verification delays for saccade onset (4 ms) and offset (20 ms) ensure a stable state before saccade onset/offset is signaled. The EyeLink Parser does no filtering of the eye movement data prior to classification, other than the Stampe (1993) filter mentioned above. Since the details of the calculation of radial velocity employed by the EyeLink Parser are proprietary, we used the following simplified calculation, suggested by SR Research staff, as an approximation of the velocity calculation done by the EyeLink host software:

$$Vel_{T\ (x,y)} = (x_{t+2} + x_{t+1} - x_{t-1} - x_{t-2})/6 \tag{7}$$

where x is the x horizontal position signal, y is the vertical position signal, and t is the current sample (in milliseconds).

Radial velocity is:

$$Vel_{\mathrm{radial}} = SR * \sqrt{Vel^2_{T(x)} + Vel^2_{T(y)}} \tag{8}$$

where SR is the sampling rate (1000 Hz).

A single rater (the first author, L.F.) evaluated the same 20 recordings as in Experiment 1 as classified by the EyeLink Parser algorithm. Since this evaluation occurred after the ONH–MNH comparison, the rater was not blind to the classification method, as had also been the case with the IRF algorithm. Reducing the evaluation to one versus three raters would lower the generalizability of the results to a potential population of raters, as with the IRF algorithm. Also, with only one rater, the issue of interrater reliability was not relevant.

Statistical significance of the differences in error numbers and the corresponding effect sizes were calculated as above for Experiment 1.

Results: EyeLink Parser[5]

General impressions

There are several general impressions that one has after reviewing the classification results from the EyeLink Parser. First, the EyeLink Parser does not classify PSOs, and there were many PSOs in these recordings. We understand that PSOs are intended to be ignored by the EyeLink Parser. We also understand that, for certain real-time applications, the failure to detect PSOs may still allow for some meaningful online calculations. Nonetheless, PSOs are real events (although they may not be real eye movements; Hooge, Holmqvist, & Nystrom, 2016; Nystrom, Hooge, & Holmqvist, 2013), and we consider the failure of this algorithm to detect them is a flaw in the parser, especially when it is used for offline analysis. The ONH found 1,051 PSOs, and the MNH found 1,350 PSOs. Since our focus is on offline analysis, from our point of view the failure to detect these PSOs is considered to generate classification errors. Second, all blink periods (missing data) are both preceded and followed by "saccades." SR Research refers to these events as "blink-saccades," and makes no judgment regarding whether the saccades are true saccades or artifacts. In our sample, all such events were artifacts. SR Research informs users in its user manual that these events should be removed, so we will henceforth consider them removed. During online processing, there is not sufficient time to discriminate between the blink–saccades and regular saccades, and therefore these events remain classified as saccades. It is up to the user to remove them.

In general, the EyeLink Parser performs very well when determining saccade onset. However, a large number of saccades ended too early. The EyeLink Parser does a reasonably good job determining the end of saccades in the absence of PSOs. But in the presence of PSOs, we find many saccades that end too late and include a portion of the following PSO.

Error classification for the EyeLink Parser versus the ONH and the MNH

For the ONH–EyeLink Parser comparison, the ONH had higher numbers of errors than the EyeLink Parser for two error types, and the EyeLink Parser had more errors than the ONH for one type of error. For the MNH–EyeLink Parser comparison, the MNH has significantly more errors than the EyeLink Parser for one error type ("saccades that start too early"), and the EyeLink Parser had more errors than the MNH for one error type ("saccades that end too early").

The EyeLink Parser scores a trivially small number of saccades that start too early (or those that start too late, data not shown). So, one would have to conclude that the EyeLink Parser performs very well in determining the onset of each saccade. The ONH finds many saccades that end too late—far more than the EyeLink Parser—whereas the MNH finds fewer of these errors than the EyeLink Parser (but not significantly so).

PSO study

We also wanted to know what the EyeLink Parser does with PSOs, as defined by the MNH. There were 1,083 MNH-

[5] This section is abbreviated due to space limitations. In particular, the tables and figures illustrating the results have been removed. For the complete section, including tables and figures, see Full_Report_On_the_Classification_of_the_EyeLink_Parser.docx at https://digital.library.txstate.edu/handle/10877/6874.

defined PSOs, consisting of a total of 17,125 samples (i.e., milliseconds) of data, and only 2,093 samples were classified as saccades (12.2%). For each PSO found by the MNH, we determined what percentage of the PSO was classified as fixation by the EyeLink Parser. The percentage of all MNH-defined PSOs that were classified as 100% fixation was 56%. The percentage of all MNH-defined PSOs that were, in part, assigned to saccades was 44%. We also looked, subject by subject, at the percentage of all MNH-defined PSOs that consisted of 20% or more saccade, as classified by the EyeLink Parser. For three subjects, roughly 50% of all the MNH-defined PSOs consisted of 20% or more saccade according to the EyeLink Parser.

Discussion

Experiment 1: ONH versus MNH

The main findings of this part of the study are that our modified version of the Nyström and Holmqvist (2010) algorithm (MNH) has markedly fewer errors in the timing of the onset and offset of saccades and the number of undetected fixations and saccades than the original algorithm (ONH; Nyström & Holmqvist, 2010). In a small, but not insignificant, number of subjects, the ONH completely fails (our best estimate is in 8.3% of subjects). The MNH does underperform with respect to the ONH, however, in the detection of very small saccades. As a result, the MNH scores significantly more fixations, fewer saccades, more PSOs, and leaves fewer periods unclassified than does the ONH. Also, data scored by the MNH result in significantly more time in fixation, significantly less time in saccades, and much less time in unclassified periods than does the ONH.

We found the following issues with the ONH as applied to our data: (1) In our view, it uses too much smoothing of the position and velocity signals; (2) it uses a threshold for extending an artifact forward and backward in time that is much too small for our data; (3) it uses nonoptimal metrics to define the central tendency and spread of the highly skewed fixation velocity noise distribution; (4) the adaptive thresholds it computes for our data are unreasonably small, (5) by not considering the position trace, it consistently marks the start of saccades too early and the end of saccades too late; and (6) by not allowing fixation periods to start or end with artifact, it dramatically underscores fixation. Because of these issues with the ONH, as applied to our data, the ONH scores many more saccades as starting too early, or ending too late, and it fails to detect many fixations.

Eight error types were not reliably detected across raters. This indicates that, at least, more time writing guidelines, practicing, and having consensus discussions will be required before these error types can be reliably classified. Of course, there is no guarantee that these error types can, in the final analysis, be scored reliably. Although the raters were only highly reliable for four error types, these error types accounted for approximately 90% of all errors, so most of these error classifications were performed reliably. The reliability of ratings of saccade timing was excellent, and this was probably due to the ability of the rater to zoom into the recording to see each sample individually, as well as to the precise, 3-ms rule we used. Undetected fixations and saccades are easy to identify. Human classification of PSO-related errors was difficult and not reliable—more work is required in the future to resolve this issue.

In the original article, Nyström and Holmqvist (2010) raised a concern that the Stampe (1993) heuristic spike filter employed by the EyeLink system might have the effect of removing the number of PSOs. Since both data-recording systems applied the Stampe filter, and both found many PSOs, this concern was apparently unfounded. The algorithms did not differ in terms of the time spent making PSOs.

Although generally the MNH far outperformed the ONH, the MNH is clearly not perfect, and certain aspects of it could be improved. For example, although there were fewer saccade timing errors with this method, there were still a substantial number of such errors. In our view, to further improve the performance of the MNH for saccade timing, the position trace signals need to be taken into consideration as well as the velocity trace. The performance of the MNH in detecting very small saccades could be improved by the development of a special subroutine devoted to this task. A simple solution of lowering the fixed saccade velocity detection thresholds might work, or it might create unforeseen issues with the scoring of other events. In the future, we plan to try several approaches to deal with this issue.

We think that the present method used to evaluate error types provides an unusually rich source of information for an algorithm developer. Obviously, before others use this method, they might consider methods for enhancing the reliability of the detection of several error types. The creation of detailed instructions for how to detect certain errors may require an iterative process of drafting guidelines, scoring a sample of data, and having a follow up consensus discussion. But the method has the potential to provide information on all of the decisions such an algorithm needs to make, and in our evaluation, it is much less cumbersome than other methods that require human expert classification of every event in a recording (Andersson et al., 2017), which is an extremely time consuming and tedious procedure.

One caveat of the present study is that, given that the two algorithms smooth the data to different extents, it was possible for raters to determine which of the two algorithms classified the data being scored. We could not think of a way around this, but it is a threat to the extent to which the recordings were "blindly" rated. Furthermore, we make no claim that our algorithm is better, in every instance, with every kind of eye movement data. We had only our sample of subjects studied with the EyeLink 1000 at 1000 Hz to evaluate during a reading task, and, for the time being, our results must be considered limited in this way.

The MNH algorithm, with fixed thresholds for classifying saccades, performed much better than the adaptive method employed by the ONH algorithm. Although Nyström and Holmqvist (2010) emphasize the value of adaptive thresholds, it seems to the present authors that Nyström and Holmqvist do not provide either a theoretical or empirical foundation for the use of adaptive filters. It is not obvious that thresholds for saccade definition, or any event definition, should be adjusted for the level of fixation noise in a recording. The approach seems reasonable, and apparently works well with Nyström and Holmqvist's data. But the opposite notion, that the velocity thresholds for classifying eye movement events should be constant across subjects, is also reasonable. With adaptive thresholds, the definition of a saccade, for example, is formally different for each subject, and even for the same saccade occurring in different contexts in the same recording. Why should this be so? Using the same thresholds for each subject lead to events that are identically defined across subjects. In our context, in which we use eye movements to biometrically recognize humans at various time intervals, the same event definition will lead to more stable classification results and better biometric performance. As applied to our data, the adaptive thresholds for saccade peak velocity detection are often unreasonably small (10°–20°/s), whereas, Nyström and Holmqvist report a mean value of 33°/s. Since these adaptive thresholds are based on velocity noise during fixation, it would appear that our data were substantially less noisy than theirs. This is true despite our use of much less smoothing. If we were to try to reinstate the adaptive filter approach going forward, we would use statistical metrics that match the highly skewed distribution of velocity noise during fixation, and set a lower limit on the threshold levels.

We must acknowledge that this article was written in the belief that the classification of eye movements by human beings can be a gold standard. We are aware that Hooge, Niehorster, Nystrom, Andersson, and Hessels (2017) have stated that this view cannot be supported. We disagree with Hooge et al. (2017) and believe their analysis is based on an incomplete interrater reliability study. They studied the reliability of experienced oculomotor researchers at a single point in time, without any rater training or consensus discussions. There is a very large literature on the beneficial effects of rater training (Abaza & Ross, 2009; Abbo, Okello, & Nakku, 2013; Alcott, Swann, & Grafham, 1999; Angkaw, Tran, & Haaga, 2006; Bank et al., 2002; Buijze, Guitton, van Dijk, Ring, & the Science of Variation Group, 2012; Chan & Yiu, 2002; Chapman et al., 2016; Cusick, Vasquez, Knowles, & Wallen, 2005; Degenhardt, Snider, Snider, & Johnson, 2005; Haj-Ali & Feil, 2006; Istriana et al., 2013; Iwarsson & Reinholt Petersen, 2012; Lievens, 2001; Lou et al., 2014; Lundh, Kowalski, Sundberg, & Landen, 2012; Magnan & Maklebust, 2009; Mist, Ritenbaugh, & Aickin, 2009; Rosen et al., 2008; Sattler, McKnight, Naney, & Mathis, 2015; Schredl, Burchert, & Gabatin, 2004; Solah et al., 2015; Staelens et al., 2014; Store-Valen et al., 2015; Taninishi et al., 2016). There is also a literature on the beneficial effects of having consensus discussions and employing consensus guidelines (Beerbaum et al., 2009; Degenhardt et al., 2005; Foppen, van der Schaaf, Beek, Verkooijen, & Fischer, 2016; Iwarsson & Reinholt Petersen, 2012; Meade et al., 2000; Weinstock et al., 2001). The first author of this article has conducted many human interrater reliability studies, with humans classifying complex signals, images, or behaviors, over his career. In no case, upon first testing, were humans found to be reliable. However, using an iterative process of assessing reliability, having consensus discussions, and developing consensus guidelines has always led to much higher, and quite useful levels of interrater reliability. In his view, human raters are generally not reliable in rating complex phenomena without training and consensus discussion.

Recent findings have called into question the usefulness and accuracy of video-based eyetrackers, which rely on estimates of the pupil position and the corneal reflection position (Hooge et al., 2016; Nystrom et al., 2013). PSOs in particular are seen, at least partially, to be an artifact of the differing temporal dynamics of the pupil and the corneal reflection. The claim is also made that these devices are not suitable for studying the detailed dynamics of eye movements. We have nothing to contribute to this discussion, and we leave it to the user to decide whether or not to treat the PSOs detected by the MNH as eye movements. Perhaps some may choose to include PSO time in the postsaccadic fixation period. We want to simply state that, to define PSO events, we relied on the definitions provided in Nyström and Holmqvist (2010), particularly as displayed in their Figs. 1 and 9. We are not aware of any retraction on the part of these authors as to this definition of the appearance of PSOs in recordings from video-oculographic eye movement recording systems.

Experiment 2: Machine-learning algorithm

Our analysis of the IRF machine-learning algorithm of Zemblys et al. (2018) revealed several serious flaws in the output of this method. Although the IRF is amazingly accurate for saccade timing, there were a number of serious problems with the output of this method. Although the IRF algorithm states that it includes a number of postprocessing steps designed to remove unusual events, the evidence suggests that many postprocessing steps were either inadequately applied or not applied at all. Without these postprocessing steps, the output of the IRF is so replete with errors that we cannot recommend its use at this time.

The machine-learning methods employed in the IRF require training from human classifiers. It would appear that the human classifiers who trained this version of the algorithm have an unusual definition of what a PSO must look like—

certainly completely different from that of Nyström and Holmqvist (2010), the authors of the ONH. The IRF algorithm classified more than twice as many PSOs as the ONH, and many of these "PSOs" had no evidence of noticeable velocity peaks at all. Also, some PSOs occurred after a fixation.[6]

Experiment 3: EyeLink Parser

The Eyelink Parser does not classify PSOs. Although this may or may not be an issue for those who use the algorithm for online processing and decision making, in our view, it is definitely an issue for offline analysis. There were approximate 1,000 PSOs in the recordings analyzed herein for 520 total seconds analyzed, so approximately two PSOs per second were missed. The EyeLink Parser does an excellent job determining the timing of saccade onset, but it tends to classify more saccades that end too early than either the ONH or the MNH.

Conclusion

In conclusion, we have modified the original Nyström and Holmqvist (2010) algorithm (ONH) so that we now have a new algorithm, the modified Nyström and Holmqvist method (MNH). It makes dramatically fewer errors, with our data, in saccade timing, and it detects fixations and saccades that remain unclassified by the ONH. The MNH never has complete failures, whereas in some cases the ONH did completely fail. The MNH algorithm does not detect very small saccades, however. The IRF algorithm of Zemblys et al. (2018) does a very poor job of dealing with unusual or artifactual noise in recordings and produces PSO classifications that do not look like PSOs. At this stage, we cannot recommend its use. The EyeLink Parser misses all PSOs and tends to find saccades that end too early, but it does very well detecting the timing of the onset of saccades.

Author note We acknowledge the support to O.V.K. from the National Science Foundation (Grant CNS-1250718) and the National Institute for Standards and Technology (Grants 60NANB15D325, 60NANB16D293). We also express our gratitude to Raimondas Zemblys for scoring our data with his algorithm. Sam Hutton is a staff scientist at SR Research. He helped us describe the workings of the EyeLink Parser and review this section for errors and comments. None of the authors have any conflict of interest to declare.

[6] For a more full discussion of the PSO detection in the Zemblys et al. (2018) article, see the document labeled "Report on PSO Detection in the Zemblys et al. (2018) Paper.docx" at https://digital.library.txstate.edu/handle/10877/6874.

References

Abaza, A., & Ross, A. (2009). Quality based rank-level fusion in multibiometric systems. Paper presented at the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Washington, DC.

Abbo, C., Okello, E. S., & Nakku, J. (2013). Effect of brief training on reliability and applicability of Global Assessment of functioning scale by Psychiatric clinical officers in Uganda. *African Health Sciences*, *13*, 78–81. https://doi.org/10.4314/ahs.v13i1.11

Abdulin, E., Friedman, L., & Komogortsev, O. V. (2017). *Method to detect eye position noise from video-oculography when detection of pupil or corneal reflection position fails*. Unpublished manuscript. arXiv:1709.02700

Alcott, D., Swann, R., & Grafham, A. (1999). The effect of training on rater reliability on the scoring of the NART. *British Journal of Clinical Psychology*, *38*, 431–434.

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nystrom, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, *49*, 616–637. https://doi.org/10.3758/s13428-016-0738-9

Angkaw, A. C., Tran, G. Q., & Haaga, D. A. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behaviour Research and Therapy*, *44*, 533–544. https://doi.org/10.1016/j.brat.2005.04.002

Bank, A. L., Macneill, S. E., Hall, E. M., Nadjarian, R. K., Zaccagnini, A. V., & Lichtenberg, P. A. (2002). More than meets the eye: how examiner training affects the reliability of the MacNeill–Lichtenberg decision tree in geriatric rehabilitation patients. *Archives of Physical Medicine and Rehabilitation*, *83*, 405–411.

Beerbaum, P., Barth, P., Kropf, S., Sarikouch, S., Kelter-Kloepping, A., Franke, D., … Kuehne, T. (2009). Cardiac function by MRI in congenital heart disease: Impact of consensus training on interinstitutional variance. *Journal of Magnetic Resonance Imaging*, *30*, 956–966. https://doi.org/10.1002/jmri.21948

Buijze, G. A., Guitton, T. G., van Dijk, C. N., Ring, D., & the Science of Variation Group. (2012). Training improves interobserver reliability for the diagnosis of scaphoid fracture displacement. *Clinical Orthopaedics and Related Research*, *470*, 2029–2034. https://doi.org/10.1007/s11999-012-2260-4

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech Language and Hearing Research*, *45*, 111–126. https://doi.org/10.1044/1092-4388(2002/009)

Chapman, K. L., Baylis, A., Trost-Cardamone, J., Cordero, K. N., Dixon, A., Dobbelsteyn, C., … Sell, D. (2016). The Americleft Speech Project: A training and reliability study. *Cleft Palate–Craniofacial Journal*, *53*, 93–108. https://doi.org/10.1597/14-027

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Mahwah: Erlbaum.

Cooper, H. M., & Hedges, L. V. (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Cusick, A., Vasquez, M., Knowles, L., & Wallen, M. (2005). Effect of rater training on reliability of Melbourne Assessment of Unilateral Upper Limb Function scores. *Developmental Medicine & Child Neurology*, *47*, 39–45.

Degenhardt, B. F., Snider, K. T., Snider, E. J., & Johnson, J. C. (2005). Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: Improvements from consensus training. *Journal of the American Osteopathic Association*, *105*, 465–473.

Foppen, W., van der Schaaf, I. C., Beek, F. J., Verkooijen, H. M., & Fischer, K. (2016). Scoring haemophilic arthropathy on X-rays: Improving inter- and intra-observer reliability and agreement using a consensus atlas. *European Radiology*, *26*, 1963–1970. https://doi.org/10.1007/s00330-015-4013-8

Haj-Ali, R., & Feil, P. (2006). Rater reliability: Short- and long-term effects of calibration training. *Journal of Dental Education*, *70*, 428–433.

Hein, O., & Zangemeister, W. H. (2017). Topology for gaze analyses—Raw data segmentation. *Journal of Eye Movement Research*, *10*, 1: 1–25.

Hooge, I., Holmqvist, K., & Nystrom, M. (2016). The pupil is faster than the corneal reflection (CR): Are video based pupil-CR eye trackers suitable for studying detailed dynamics of eye movements? *Vision Research*, *128*, 6–18. https://doi.org/10.1016/j.visres.2016.09.002

Hooge, I. T. C., Niehorster, D. C., Nystrom, M., Andersson, R., & Hessels, R. S. (2017). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-017-0955-x

Istriana, E., Kurnia, A., Weijers, A., Hidayat, T., Pinxten, L., de Jong, C., & Schellekens, A. (2013). Excellent reliability of the Hamilton Depression Rating Scale (HDRS-21) in Indonesia after training. *Asia-Pacific Psychiatry*, *5*, 141–146. https://doi.org/10.1111/appy.12083

Iwarsson, J., & Reinholt Petersen, N. (2012). Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice*, *26*, 304–312. https://doi.org/10.1016/j.jvoice.2011.06.003

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255–264.

Lou, X., Lee, R., Feins, R. H., Enter, D., Hicks, G. L., Jr., Verrier, E. D., & Fann, J. I. (2014). Training less-experienced faculty improves reliability of skills assessment in cardiac surgery. *Journal of Thoracic and Cardiovascular Surgery*, *148*, 2491–2496. https://doi.org/10.1016/j.jtcvs.2014.09.017

Lundh, A., Kowalski, J., Sundberg, C. J., & Landen, M. (2012). A comparison of seminar and computer based training on the accuracy and reliability of raters using the Children's Global Assessment Scale (CGAS). *Administration and Policy in Mental Health*, *39*, 458–465. https://doi.org/10.1007/s10488-011-0369-5

Magnan, M. A., & Maklebust, J. (2009). The effect of Web-based Braden Scale training on the reliability of Braden subscale ratings. *Journal of Wound Ostomy & Continence Nursing*, *36*, 51–59. https://doi.org/10.1097/WON.0b013e3181919b8d

Meade, M. O., Cook, R. J., Guyatt, G. H., Groll, R., Kachura, J. R., Bedard, M., … Stewart, T. E. (2000). Interobserver variation in interpreting chest radiographs for the diagnosis of acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine*, *161*, 85–90. https://doi.org/10.1164/ajrccm.161.1.9809003

Mist, S., Ritenbaugh, C., & Aickin, M. (2009). Effects of questionnaire-based diagnosis and training on inter-rater reliability among practitioners of traditional Chinese medicine. *Journal of Alternative and Complementary Medicine*, *15*, 703–709. https://doi.org/10.1089/acm.2008.0488

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*, 188–204. https://doi.org/10.3758/brm.42.1.188

Nystrom, M., Hooge, I., & Holmqvist, K. (2013). Post-saccadic oscillations in eye movement data recorded with pupil-based eye trackers reflect motion of the pupil inside the iris. *Vision Research*, *92*, 59–66. https://doi.org/10.1016/j.visres.2013.09.009

Rigas, I., Komogortsev, O., & Shadmehr, R. (2016). Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transactions on Applied Perception*, *13*, 6. https://doi.org/10.1145/2842614

Rosen, J., Mulsant, B. H., Marino, P., Groening, C., Young, R. C., & Fox, D. (2008). Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. *Psychiatry Research*, *161*, 126–130. https://doi.org/10.1016/j.psychres.2008.03.001

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye tracking protocols. Paper presented at the Eye Tracking Research & Applications Symposium, New York.

Sattler, D. N., McKnight, P. E., Naney, L., & Mathis, R. (2015). Grant peer review: Improving inter-rater reliability with training. *PLoS ONE*, *10*, e0130450. https://doi.org/10.1371/journal.pone.0130450

Schredl, M., Burchert, N., & Gabatin, Y. (2004). The effect of training on interrater reliability in dream content analysis. *Sleep and Hypnosis*, *6*, 139–144.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd). New York: McGraw-Hill.

Solah, V. A., Meng, X., Wood, S., Gahler, R. J., Kerr, D. A., James, A. P., … Johnson, S. K. (2015). Effect of training on the reliability of satiety evaluation and use of trained panellists to determine the satiety effect of dietary fibre: a randomised controlled trial. *PLoS ONE*, *10*, e0126202. https://doi.org/10.1371/journal.pone.0126202

Staelens, A. S., Tomsin, K., Oben, J., Mesens, T., Grieten, L., & Gyselaers, W. (2014). Improving the reliability of venous Doppler flow measurements: Relevance of combined ECG, training and repeated measures. *Ultrasound in Medicine & Biology*, *40*, 1722–1728. https://doi.org/10.1016/j.ultrasmedbio.2014.01.014

Stampe, D. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods*, *25*, 137–142. https://doi.org/10.3758/bf03204486

Store-Valen, J., Ryum, T., Pedersen, G. A., Pripp, A. H., Jose, P. E., & Karterud, S. (2015). Does a web-based feedback training program result in improved reliability in clinicians' ratings of the Global Assessment of Functioning (GAF) Scale? *Psychological Assessment*, *27*, 865–873. https://doi.org/10.1037/pas0000086

Taninishi, H., Pearlstein, M., Sheng, H., Izutsu, M., Chaparro, R. E., Goldstein, L. B., & Warner, D. S. (2016). Video training and certification program improves reliability of postischemic neurologic deficit measurement in the rat. *Journal of Cerebral Blood Flow & Metabolism*, *36*, 2203–2210. https://doi.org/10.1177/0271678X15616980

Weinstock, M. A., Bingham, S. F., Cole, G. W., Eilers, D., Naylor, M. F., Kalivas, J., … DiGiovanna, J. J. (2001). Reliability of counting actinic keratoses before and after brief consensus discussion: the VA topical tretinoin chemoprevention (VATTC) trial. *Archives of Dermatology*, *137*, 1055–1058.

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, *50*, 160–181. https://doi.org/10.3758/s13428-017-0860-3