Behavior Research Methods, Instruments, & Computers
1993, 25 (2), 137-142

Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems

DAVE M. STAMPE
University of Toronto, Toronto, Ontario, Canada

Methods for enhancing the accuracy of fixation and saccade detection and the reliability of calibration in video gaze-tracking systems are discussed. The unique aspects of the present approach include effective low-delay noise reduction prior to the detection of fixation changes, monitoring of gaze position in real time by the operator, identification of saccades as small as 0.5° while eliminating false fixations, and a quick, high-precision, semiautomated calibration procedure.

Eye-position recording has produced important results in fields such as reading research, visual search, and problem solving. Almost all experiments utilizing eye-position measures require reporting the location, length, and order of fixations (significant periods of gaze in which the eye is stationary). Other data, such as pupil size and blink rates, are less often utilized, but should be available if required.

An important consideration in any recording system is that data validity be preserved, with as few artifacts and distortions introduced by the recording process as possible. In the case of eye-position recording equipment, noise can mask small changes in gaze location, and equipment setup problems can cause reported gaze location to be in error by several degrees of visual angle. Therefore, these systems must be designed with the best possible noise reduction and calibration procedures.

The techniques discussed in this paper were developed for a gaze-position recording system implemented on an IBM-compatible 386 computer, with the use of an ISCAN RK-416PC pupil-tracking board and a video camera as input. The system has one monitor for the subject and another for the operator, the latter displaying the same image as the subject's monitor plus a real-time gaze-position cursor. The cursor display requires a low-delay filter to remove noise from the eye-position data in order to reduce jitter. Also, the low delay of the cursor and filter makes possible the implementation of gaze-contingent displays, in which the image on the subject's monitor changes with gaze position. For example, the cursor may be replaced by a mask to continuously block foveal vision. The system itself will not be discussed further, except as it relates to the development of the methods described.

I will first discuss data and noise characteristics of pupil-tracking devices and present the methods developed for the filtering of the position data and for the related issue of fixation and saccade detection, comparing the new methods with the literature. Methods for mapping the eye-tracker-position data into monitor-screen coordinates and their effect on data validity are analyzed. The calibration procedure will then be discussed, and improvements made possible by the real-time feedback system will be described.

The author thanks Eyal Reingold, Elizabeth Bosman, and an anonymous reviewer for their comments on a preliminary version of this paper. A listing of the C code for the mapping function and coefficient calculation is available from the author. Correspondence should be addressed to D. M. Stampe, Department of Psychology, University of Toronto, Toronto, ON, Canada M5S 2Z9 (e-mail: dstampe@psych.toronto.edu).

TRACKER DATA CHARACTERISTICS

Data from video-based pupil tracking devices such as the ISCAN RK-416PC are complicated by the low sample rate and resolution of the present generation of these devices. The sampling rate is set by the video field rate (60 Hz for the NTSC television standard, or 16.7 msec between samples). At this rate, the shortest fixations, 83 msec in length, are covered by only five samples, and short saccades will not be sampled at all, appearing instead as abrupt jumps in position.

Eye-position data from the tracker is quantized in units as large as 1° of visual angle, depending on the exact pupil-tracking method used and the video camera's field of view. Systems that utilize both corneal reflection and pupil tracking to cancel head movement effects (Merchant, 1974) have less than half the resolution of pupil tracking alone. Since noise peaks in the tracker data can be as high as 4 quantization units, the low resolution makes detection of small saccades difficult unless the noise can be removed by filtering.

One solution to the tracker resolution problem is to have the stimulus cover a larger field of view, resulting in larger eye movements and decreasing the relative amount of noise. This should be done with caution, since subjects may adopt different strategies for larger field-of-view presentations than would be the case in more natural viewing conditions. A physical limit of the field of view is occlusion of the pupil edges by the eyelids, which limits the usable eye rotation to ±25° horizontally and ±15° vertically. This range may be extended by increasing environmental lighting to reduce pupil size or by presenting stimuli in black on a white background, which will also reduce

retinal afterimages and may lead to more natural task performance. As a guideline, the current system has a usable field of view of 22° horizontally × 18° vertically, and the tracker data is quantized in steps equivalent to 0.12° horizontally and 0.25° vertically (1.0 mm and 2.0 mm, respectively, at a monitor distance of 450 mm).

FILTERING

Eye-tracking systems with analog outputs such as EOG (electrooculography) or scleral/limbus reflection devices permit high sampling rates and resolution, although their noise levels are fairly high. For these systems, linear filters are often used to remove noise (Inchingolo & Spanio, 1985). Because of the low sampling rate of pupil-tracking systems, linear filters would smooth position data excessively, making saccade detection difficult or impossible.

Instead of linear filtering, template-matching logical filters may be utilized. These filters compare each data sample with neighboring samples and modify or pass the sample accordingly. They function well at low sample rates and add little or no delay to the data processing. Their template-matching characteristics make them ideal for removing impulse noise and detecting saccades or fixations in the data.

Most of the reported methods of analysis for video gaze-tracking systems rely on the relatively low noise level and attempt to detect fixation changes directly without first removing noise. The delta method, introduced by Mason (1976), computes a running average of all samples in the fixation as the fixation position estimate. When a new sample's distance from the estimated fixation position exceeds the delta value threshold, a new fixation is begun. Unfortunately, this filter requires fairly large delta threshold values (typically 1° or greater) and may produce false fixations caused by noise pulses. It also produces false fixations during long saccades that must be eliminated later. Kliegl and Olson (1981) have developed methods to clean up this filter's output by eliminating or combining short fixations, usually during postprocessing of experimental data; however this would add too much delay to the real-time display system.

Heuristic Filter Design

Rather than detecting fixations directly, a logical filter was developed to remove the noise from the position data before detection of saccades and fixations. This provides clean data for the real-time gaze-position cursor display and prevents false fixation output. Also, the cleaned eye-position data are available for verification of correct operation or for studies in which saccade characteristics are analyzed.

The design of the heuristic filter relies on "rules of thumb" deduced by examining the noise characteristics in the raw data and by studying human analysis methods. These heuristics are similar to those found in expert systems and are stated by a list of goals to be achieved. The rules can be expressed in forms that are implemented in a few lines of code, using comparisons and copies.

Ideally, the data from the eye tracker would consist of periods of little or no motion (fixations) and regions of rapid motion or jumps in position (saccades). See Figure 1 (and Figure 4 for temporal detail) for an example of real horizontal eye-position data recorded during a reading trial, including a long return sweep and smaller word-fixation saccades. Notice the nontrivial noise level that could mask small saccades.

The overall goal is to eliminate the noise content of the tracker output, defined as its difference from a saccade (a monotonically increasing or decreasing feature) and fixation (plateau) model. The noise content is defined as consisting of nonmonotonic features (e.g., an increase followed by a decrease in value) that are too short to be fixations, which are defined as being three samples (50 msec) or greater in duration. The noise also includes ringing or overshoot artifacts following saccades, which can confuse the saccade detector into extending the saccade into the next fixation.

Almost all noise produced by the video tracker is in the form of one-sample spikes, with two-sample pulses occurring less than once a second. A simple means to recognize one-sample noise spikes is to look for an increase in value followed by an immediate decrease in value (or vice versa) by checking each sample against the next and previous samples. The detected noise pulse is replaced by the neighboring sample closest in value to it rather than by the mean of the neighboring samples, since this produces the flattest fixation output (see Figure 2). To make the "next" sample available, the filter introduces a one-sample delay. Output from this stage is largely noise-free and is used for the gaze-position cursor display because of its low delay.

The second filter stage eliminates any two-sample noise events. The first filter stage will convert all two-sample noise events into flat-topped sample pairs, making detection simple. The filter looks for sample pairs with the same value that do not equal either of their neighbors, and

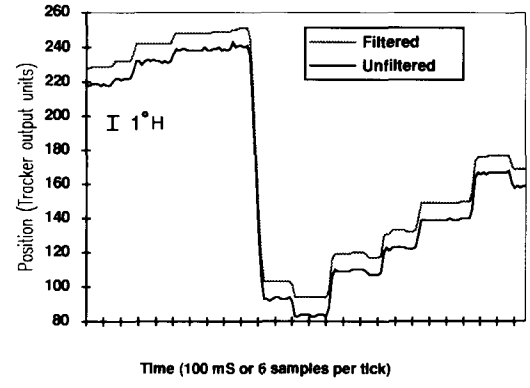

Figure 1. Horizontal eye-position data collected during a reading task, before and after heuristic filtering.

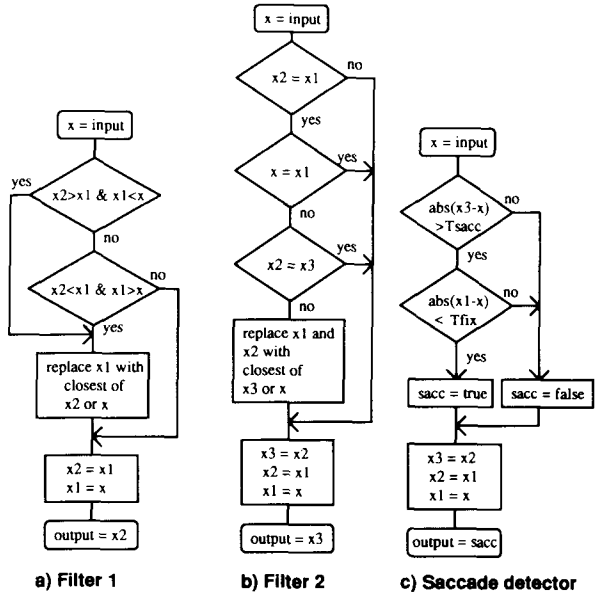

Figure 2. Flowcharts of the heuristic filters and saccade detector. T_{sacc} and T_{fix} are the detector thresholds; x is the current input, $x1$, $x2$, and $x3$ are progressively delayed samples.

replaces both samples by the neighbor closer in value to them. The delay of the second filter is two samples, for a total filter delay of three samples (see Figure 2b). Since this filter's output is used for saccade detection only, the delay is relatively unimportant. The horizontal and vertical position data and the pupil-size data from the eye tracker are filtered separately.

The results of the filter stages on common noise events are shown in Figure 3. In all cases, the noise features are removed completely, revealing the underlying saccade/fixation structure. Performance of the filter on real data from a reading task can be seen in Figure 1, with a sample-by-sample detail from this data shown in Figure 4. Note the preservation of the saccade structure and slow glides (possibly from head movements).

SACCADE DETECTION AND FIXATION PROCESSING

Fixations are defined as being separated by saccades, which may be detected by their rapidly changing location. Where sampling rates are high, a linear highpass filter may be used to detect saccades (Inchingolo & Spanio, 1985). With the low sampling rates of video-based trackers, template-matching filters must be used instead.

Saccades are discriminated by their velocity: A criterion of 30°/sec or higher is common, since this is the limit of pursuit eye movement speed. A simple test for a saccade is to compute the difference between adjacent samples and compare this with the saccade threshold T_{sacc}. The threshold equivalent to 30°/sec is 0.5° per sample at 60 samples/sec, which may be below the limit of resolution for some tracking systems. To improve sensitivity and increase the threshold, we compute the difference of data separated by 2 samples. All nonmonotonic features smaller than 3 samples in size were removed by the heuristic filters, thus the samples between the tested points can be ignored.

To prevent stretching of saccades and erosion of fixations, the detector also requires that the previous sample and the current sample differ by less than the fixation threshold T_{fix}. This forces the saccade detector to turn off as soon as the fixation begins (see Figure 2c for flowchart). The fixation threshold causes quick saccades (less than three samples in duration) to be judged by distance rather than by velocity, which sets the absolute minimum detectable saccade length and rejects smaller steps as noise.

The saccade detector sees filtered eye-tracker-position data before it has been mapped into screen coordinates. Thus, the thresholds are specified in eye-tracker data units

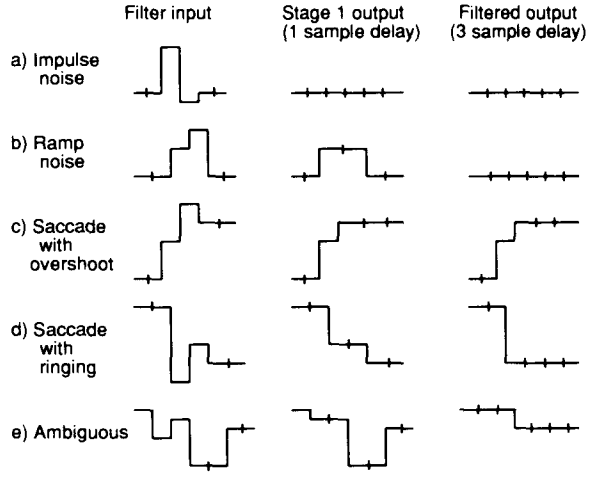

Figure 3. Heuristic filter processing of common noise types in video tracker data. Ticks or steps indicate divisions between data samples. Examples (a) and (b) are noise reduction within a fixation, (c) and (d) are saccades, and (e) is ambiguous data interpreted as a small saccade.

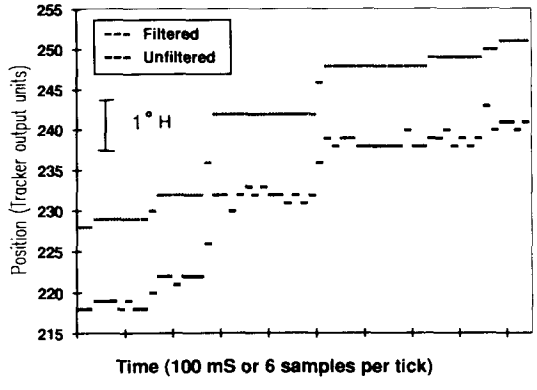

Figure 4. Detail of heuristic filtering on data (vertical scale expanded) from the top left of Figure 1. Each line segment is a separate data sample. Note the preservation of saccade structure and the slow drift at right caused by head movement.

and may not correspond to a constant visual angle over the entire screen. This results in the best noise rejection by the detector, assuming the tracker noise level does not vary with the eye position. T_{fix} is determined by the residual tracker noise level after filtering and cannot be arbitrarily reduced if an eye-tracker setup with lower resolution is used, whereas T_{sacc} scales approximately with tracker resolution, to a minimum of 2 units. For the horizontal saccade detector of the implemented system, T_{sacc} is 5 (tracker data units) and T_{fix} is 1, which allows detection of saccades of 0.5° or greater.

Tests on real data show that the saccade detector works better than the single-sample difference detector both in reading tasks with small, fast saccades, and on visual search tasks that may have slower saccades. Figure 5 shows its action on filtered data from Figure 1, with the saccade detector response indicated by gray bars. The bars have been shifted to compensate for the one-sample delay introduced by the detector. The detector always marks the last sample of a fixation as part of the following saccade, which is easily compensated for in the fixation integration processing. A separate saccade detector is used for horizontal and vertical data.

All data between saccades are part of one fixation, including the first sample marked as part of the saccade. The position for the fixation is computed by averaging all of its samples' horizontal and vertical positions. Pupil diameter is integrated in the same fashion. After mapping of the averaged position data to pixel coordinates on the subject's monitor (as described later), these data are written to the output file. Blinks are detected by sudden drops in pupil size and are also recorded. The flow of data through the eye-position recorder is shown in Figure 6. A postprocessing program processes the output file to reject short fixations and blinks. The postprocessor provides the flexibility required to implement a variety of analysis methods (e.g., fixation cluster analysis or lumping of fixations separated by blinks).

MAPPING GAZE POSITION TO MONITOR SCREEN COORDINATES

All filtering and fixation integration is performed in eye-tracker-position coordinates. To be useful, the eye-tracker-position data must be converted to locations on the subject's monitor screen. By expressing gaze locations as display pixel coordinates, a standard format for drawing images and analyzing the subject's fixation positions is created. In the present system, the standard pixel coordinate system is the 640×480 pixel VGA display mode.

Conversion from eye-tracker data to screen coordinates is performed by a mapping function, the selection of which determines how distortions between screen and tracker data are corrected. The coefficients of the mapping function are derived by the process of calibration, in which a set of targets in known positions are displayed to the subject, and the eye-tracker-position data is recorded. Given several of these position correspondences, the mapping function's coefficients can be computed.

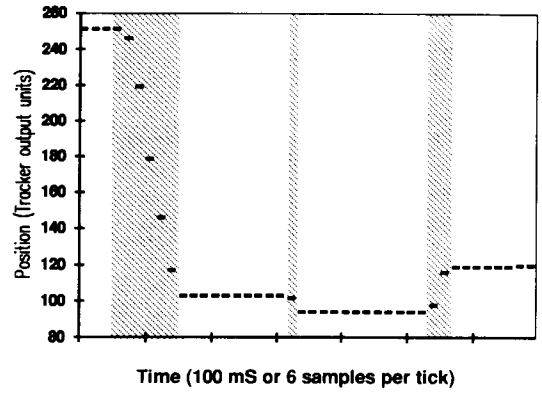

Figure 5. Saccade detector output (gray bars) for filtered data from reading task.

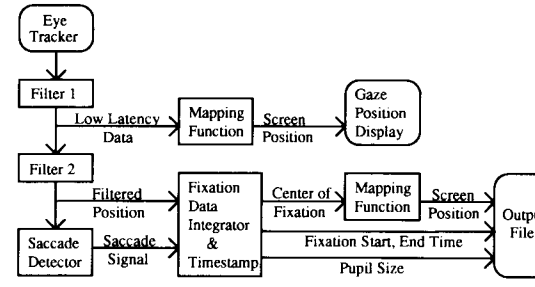

Figure 6. Data flow for the eye-position recording system with heuristic filters and real-time gaze-position display.

Mapping Function and Coefficient Solution

Choice of a mapping function sets the number of calibration points that must be presented. Calibration schemes reported in the literature use anywhere from 3 points (for a nonlinear one-dimensional calibration) to 25 points (for an extreme example of piecewise linear calibration). The average gaze position mapping error is a U-shaped function of the number of points gathered for calibration (e.g., Karpala & Jernigan, 1980). A low number of points forces the use of a simplistic mapping function that may be unable to correct all distortions. As more target positions are gathered, the spatial noise caused by inexact fixation of targets by the subject increases. McConkie (1981) has suggested that each calibration target be presented several times and that the mean position be used, but this can result in subject habituation, and the longer calibration time increases the likelihood of head movement.

Mapping functions for two-dimensional data reported in the literature are of two types, piecewise and nonlinear. The piecewise method divides the screen into a grid of cells, and a target point is presented at each grid junction. The data gathered from the tracker then defines a grid of quadrilaterals, each of which is separately mapped back onto its original rectangular grid cell (Kliegl & Olson, 1981; McConkie, 1981). The shortcoming of this method is that

abrupt changes in scaling and distortion can occur at the boundaries between grid cells. Adding more cells to the grid reduces these changes but requires a large increase in the number of data points to be collected.

Nonlinear mapping functions are capable of smooth changes in mapping across the screen. The most common nonlinear mapping function is the biquadratic, introduced by Sheena and Borah (1981). This function requires nine calibration points to compute all coefficients (close to the optimum for minimal mapping error) and can be evaluated in real time (60 times/sec) for display of the gaze-position cursor. The form chosen for the mapping function is the following:

$$x_1 = a + bx + cy + dx^2 + ey$$

$$y_1 = f + gx + hy + ix^2 + jy^2$$

$$X = x_1 + m[q]x_1y_1$$

$$Y = y_1 + n[q]x_1y_1,$$

in which x, y are the tracker data coordinates, X, Y are the monitor screen coordinates, a, b, c, d, e, f, g, h, i, j are coefficients determined by solving a 5×5 matrix for each equation, $m[q]$, $n[q]$ are correction coefficients for each of the four data quadrants, and q is the quadrant into which data mapped by the first two equations fall.

The nonlinear terms allow curved distortions to be corrected and can change scaling smoothly across the screen (unlike piecewise mapping functions). However, the nonlinear characteristics of the function can produce problems as well. The squared terms in the equation become very large as the gaze position approaches the edges of the screen, and small errors in calibration-point fixation or head movements can result in large errors in screen gaze position there. For this reason, it is suggested that the area of the screen outside the calibration "box" in Figure 7 not be used. Figure 7 also shows the positions of the nine calibration targets and a set of typical eye-tracker-position correspondences as displayed during calibration on the operator's monitor.

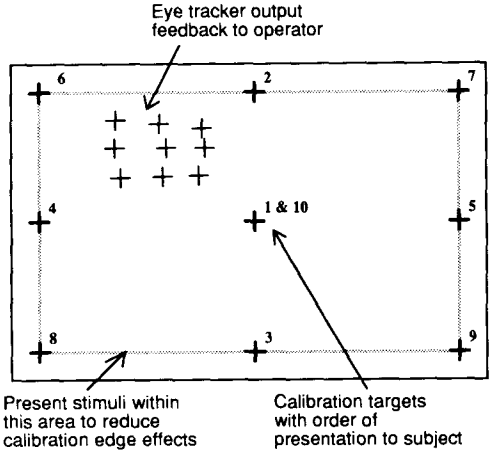

Figure 7. Calibration screen showing target positions, presentation order, and eye-tracker-output feedback.

Given the nine eye-tracker/screen-position correspondences, the coefficients of the mapping equation can be computed. Five simultaneous equations for each of the biquadratic equations are solved, using the position of Points 1–5. The solution is reduced to a 4×4 matrix by translating Point 1 to the coordinate (0,0) in both screen and tracker coordinate systems, which also provides coefficients a and e. The matrix is then solved with Cramer's rule. A full matrix solution is used instead of an incremental procedure (Sheena & Borah, 1981), because the full solution improves the calibration when the mapping function must perform rotation. The solution takes less than 500 μsec on a 486/25 PC (with internal math coprocessor) and less than 50 msec on a 386/33 PC (without coprocessor). The quadrant correction coefficients are determined by using Points 6, 7, 8, and 9 after these points are translated by the biquadratic equations. The C code for the mapping function and the coefficient calculation is available from the author.

Calibration Procedure

To perform the system calibration, the first nine calibration points are displayed in the order indicated in Figure 7. The eye-tracker-position data are displayed in real time by a cursor on the operator's monitor, along with markers showing the eye-tracker positions for previous points in the calibration sequence. This presentation helps the operator decide if the subject is properly fixating the target. Once the subject's eye position is correct and stable, the operator presses the spacebar to record the calibration point, and the next target is presented. Because the operator monitors the eye-tracker data and controls the presentation of the targets, subjects quickly learn to fixate the points properly and not to make spurious eye movements.

During the calibration-point collection, both a lowpass (smoothing) filter with a time constant of 1 sec and the heuristic filtering are applied to the tracker data to remove any microsaccades and noise. Because blinks or saccades may occur while the operator is pressing the spacebar, resulting in collection of bad data, a motion detector monitors the eye-tracker output and disables the collection of calibration data for 300 msec after such an event is detected. If the subject blinks, the operator sees that the calibration did not proceed to the next point and repeats the keypress a second or so later. Operators quickly learn the best "rhythm" for each subject, allowing quick calibration of otherwise unusable subjects, some of whom can display blink rates as high as twice a second.

The data are now processed to compute the mapping function coefficients. Finally, the center point is presented again, and its new position is used to correct for any drift in head- or eye-resting position. The new center-point position is subtracted from the original position of the center point, and the resulting correction applied to the eye-tracker-position data before mapping to screen coordinates. Assuming that head movements cause the same

change in tracker-position data for all points on the screen, this cancels the effects of head motion without the need for recalibration. This process of recentering is quick and effective.

Immediately after system calibration, a test screen is presented to the subject. This is a 5×5 grid of letters, each about 1.2° in size. The subject is asked to fixate the center letter, then each of the corner letters by name. Because the gaze position is displayed in real time, the calibration error can be estimated and the calibration repeated immediately if required. The gaze-position cursor size is the equivalent of 1° of visual angle and is used as a reference measure of the position error. An error in gaze position mapping of 0.5° on the center letter and 1° on the corner letters is considered acceptable, although most calibrations show less error. Typically, less than 1 calibration in 10 needs to be repeated with untrained subjects, and even less often with experienced subjects.

The recentering target is usually presented before each trial screen, and the corresponding eye-tracker position is used to correct for head movement. The target may take the place of, or be combined with, the fixation point that is usually presented before each trial screen. The target need not be at the center of the screen: A reference point at the desired location may be collected during calibration and used to compute the change in position of a recentering target presented before later trials. With the use of recentering, calibration need only be performed at the start of every block of trials, or at the start of every recording session. As head movement is a fact of life in most gaze-position recording systems, this unobtrusive correction technique is essential.

RESULTS AND CONCLUSIONS

Video-based eye-tracking systems pose special problems of noise reduction, and of saccade and fixation analysis, due to their relatively low spatial resolution and sampling rates. By using a heuristically derived cascade of two template-matching filters, noise can be removed before saccades are detected, resulting in a significant reduction in false fixations compared with direct fixation detection methods. The filtered data can also be used directly for real-time display of gaze position and for implementation of gaze-contingent displays.

The on-line calibration procedure takes advantage of real-time feedback of gaze position to improve the quality of the data collected and easily handles subjects with high blink rates or unstable fixation patterns who otherwise could not be tested. Relying on the operator's implicit feedback through control of the calibration process, one or two training calibrations for new subjects are sufficient to achieve good calibration results. The use of recentering screens lets full calibration be performed less often, reducing subject fatigue and speeding data collection. The semiautomated calibration procedure helps train subjects and is not affected by anticipatory saccades or blinks. Even if automated calibration is available, manual collection should remain a calibration option for handling difficult subjects who cannot otherwise be tested.

REFERENCES

INCHINGOLO, P., & SPANIO, M. (1985). On the identification and analysis of saccadic eye movements: A quantitative study of the processing procedures. *IEEE Transactions on Biomedical Engineering*, **32**, 683-693.

KARPALA, F., & JERNIGAN, M. E. (1980). Compensation for distortion in eye-movement monitors. *IEEE Transactions on Biomedical Engineering*, **27**, 113-119.

KLIEGL, R., & OLSON, R. K. (1981). Reduction and calibration of eye monitor data. *Behavior Research Methods & Instrumentation*, **13**, 107-111.

MASON, R. L. (1976). Digital computer estimation of eye fixations. *Behavior Research Methods & Instrumentation*, **8**, 185-188.

MCCONKIE, G. W. (1981). Evaluating and reporting data quality in eye movement research. *Behavior Research Methods & Instrumentation*, **13**, 97-106.

MERCHANT, J. (1974). Remote measurement of eye direction allowing subject motion over one cubic foot of space. *IEEE Transactions on Biomedical Engineering*, **21**, 309-317.

SHEENA, D., & BORAH, B. (1981). Compensation for some second-order effects to improve eye position measurements. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception*, (pp. 257-268). Hillsdale, NJ: Erlbaum.