Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors

Oleg V. Komogortsev, Denise V. Gobert, Sampath Jayarathna, Do Hyong Koh and Sandeep M. Gowda

***Abstract*— In an effort towards standardization, this paper evaluates the performance of five eye movement classification algorithms in terms of their assessment of oculomotor fixation and saccadic behavior. The results indicate that performance of these five commonly used algorithms vary dramatically even in the case of a simple stimulus evoked task using a single, common threshold value. The important contributions of this paper are: 1) evaluation and comparison of performance of five algorithms to classify specific oculomotor behavior 2) introduction and comparison of new standardized scores to provide more reliable classification performance 3) logic for a reasonable threshold value selection for any eye movement classification algorithm based on the standardized scores and 4) logic for establishing a criterion-based baseline for performance comparison between any eye movement classification algorithms. Proposed techniques enable efficient and objective clinical applications providing means to assure meaningful automated eye movement classification.**

***Index Terms*— eye movement classification, oculomotor behavior, analysis, baseline.**

I. Introduction

Computerized eye tracking technology is increasingly being used for the examination of human visual systems in several medical settings i.e. ophthalmology, cognitive psychology and neuroscience. These systems allow measurement of oculomotor responses to multiple factors such as psychological state, disease, aging and the environment [1]. Two primary eye movements, fixation and saccadic function are essential to these studies of oculomotor behavior. Oculomotor fixation is defined as the ability to suppress ocular drifts while maintaining a steady retinal image of a single target of interest, while saccadic behavior describes eye movements used to produce rapid changes in fixation on different targets within the visual field [2]. Although these measurements are universally known and used, a frustrating concern continues due to the lack of consistency for classification of these oculomotor behaviors across various settings [1, 3].

Although studies of human visual systems are of great value in the study of neurophysical phenomena [4], clinical studies relating to advancements in patient care remain challenging when using computerized eye tracking systems [3]. For instance, oculomotor behavior is used in the differential diagnosis of several disorders including mild traumatic brain injury or mTBI [5], Parkinson's vs. Alzheimer's disease [6], schizophrenia [7], functional deficits relating to macular degeneration [8], attentional deficit disorders [9], and Meniere's Disease [10]. Fixation and saccadic behavior are described in the above-mentioned studies, however there remains disagreement in how to best classify eye movements in terms of various metrics used to characterize "fixation" and "saccadic" behaviors. For instance, Crevits et al [11] used computerized video eye tracking systems to demonstrate that persons with mild traumatic brain injury (mTBI) are able to perform normal antisaccadic control. He therefore advised that this type of measurement does not have any diagnostic capability. However, Heitger et al. [5] recently used a different form of classification algorithm for the same type of eye tracking system to demonstrate that persons with mTBI do in fact demonstrate significant differences in antisaccadic behaviors. Other labs have also supported this finding [12]. Several other discrepancies exist for other disorders which lead one to conclude that eye movement classifications have been very dependent on local measurement technique and subjective interpretation.

Documentation and assessment of fixation or saccadic eye movements provide information about patient impairments and response to medication or improvements in functional tasks during activities of daily living such as reading [8]. Therefore, it is crucial that sensitive and accurate methods be employed with the use of eye tracking systems especially in clinical settings.

Recent interest has increased in the use of mathematical models to standardize classification of components relating to normal eye behavior in response to external stimuli or impairments relating to pathology or aging. However, the challenge continues to exist because analyses techniques used to track oculomotor movements continue to be highly variable and without universal standardization for system identification of specific eye behaviors [7]. This ongoing problem has led to a preference in tedious manual techniques and reluctance to adopt automatic analysis systems with limited capacity for comparisons across settings.

In this article, we will describe a standardized approach to

Manuscript received December 26, 2009. This work was supported in part by the grants from Texas State University-San Marcos and the Sigma Xi Grant in aid of research (GIAR) program grant G200810150639.

Oleg Komogortsev is with Department of Computer Science, Texas State University – San Marcos, TX 78666, phone: 512-245-0349; fax: 512-245-8750; e-mail: ok11@txstate.edu).

Denise Gobert is with Department of Physical Therapy, Texas State University – San Marcos, TX 78666 (e-mail: dg46@txstate.edu).

Sampath Jayarathna is with Department of Computer Science, Texas State University – San Marcos, TX 78666 (e-mail: sampath@txstate.edu).

Do Hyong Koh is with Department of Computer Science, Texas State University – San Marcos, TX 78666 (e-mail: dk1132@txstate.edu).

Sandeep M. Gowda is with Department of Computer Science, Texas State University – San Marcos, TX 78666 (e-mail: sm1499@txstate.edu).

specifically assess fixation and saccadic eye movements. In doing so, we will provide a review and comparison of five of the most popular eye movement classification algorithms.

Specific objectives of this paper are: 1) Provide a review and comparison of five eye movement classification algorithms to automate classification of fixation and saccadic response to a simple standardized stimulus-evoked task; 2) Provide a standardized scoring system to allow an in-depth quantitative and qualitative analysis of oculomotor behavior; 3) Provide logic for a reasonable threshold value selection based on a standardized scoring system; 4) Provide logic for developing a meaningful baseline comparison of classification algorithms' performance in terms of simple and possibly complex oculomotor plant metrics for future studies. A preliminary summary of this work is available [13], however this paper will provide a more detailed and in-depth analysis of automated classification algorithms with standardized scoring for the objectives 1-2 and new material to accomplish objectives 3 and 4.

Standardization of the scoring and baseline selection will benefit researchers outside the medical field and provide tools for meaningful threshold selection. It will also allow validation of classification results via baseline comparisons. Specifically, some other areas which might benefit are research efforts pertaining to human computer interaction [14-19], psychology[20-22], and usability [23-25].

II. AUTOMATIC ANALYSIS OF OCULOMOTOR BEHAVIOR

Both oculomotor fixation and saccades are typically assessed with the use of several clinical tools including manual visual inspection, nystagmography and computerized infrared pupillary tracking devices [7, 26, 27]. Several methods are available to automate the analysis and classification process of oculomotor data including the Velocity Threshold Identification (I-VT), Hidden Markov Model Identification (I-HMM), Minimum Spanning Tree Identification (I-MST) [28], and Kalman Filter Identification (I-KF) [17, 29]. Potentially, the use of these algorithms proves helpful to expedite the analysis process, however little work has been done to compare each method in terms of reliability or robustness of the data analysis process.

A. *Eye Movement Classification Algorithms*

Two groups of the eye-movement classification algorithms are discussed in this paper. The first group is represented by the algorithms that analyze the velocity component of the movement signal. The I-VT, I-HMM, and I-KF belong to this group. The second group contains algorithms that analyze positional properties of the signal. The I-DT and I-MST belong to this group. Fig. 1 illustrates diagrammatical representation of all algorithms. The implementation of the algorithms presented in this paper slightly differs from the previously described versions [28-30] therefore a brief verbal description for each algorithm is provided. All algorithms presented in this paper were designed for the off-line process of eye movement data. Detailed description of the pseudo-code for each algorithm is available [31].

B. *Description of Eye Movement Classification Algorithms*

Every algorithm presented here can be described in the following general form. The input to an algorithm is provided as a sequence of the eye-gaze position tuples (x_e,y_e,t) where x_e and y_e are horizontal and vertical coordinates of the eye position sample and t is the time when the sample was taken. A threshold value is provided to allow classification of each eye position sample as a fixation or a saccade, according to the classification criteria implemented in the algorithm. Next, the "Merge Function" is employed to perform classification of consecutive eye position points as a part of fixation and then collapsed into a single fixation segment with center coordinates computed as a centroid of the fixation segment. Classified fixations are subsequently merged into larger fixation segments using criteria based on two parameters: length of the time interval between two fixation groups and

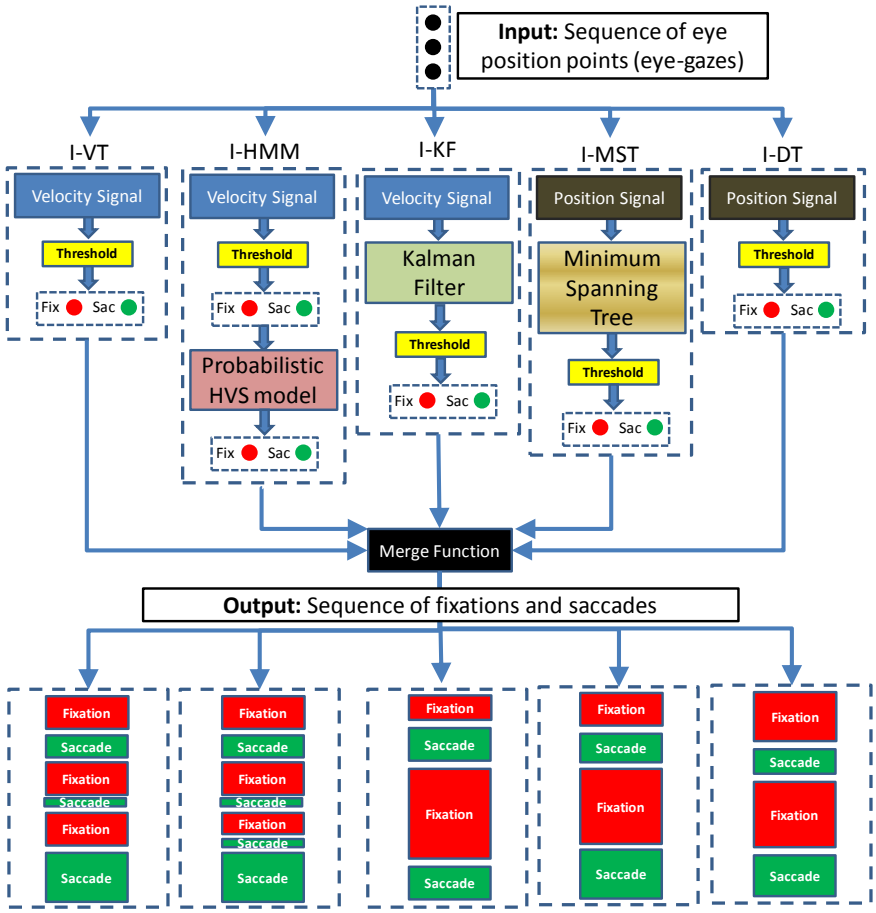

Fig. 1. Illustration of classification flow for five eye movement classification algorithms: I-VT, I-HMM, I-KF, I-DT, I-MST.

the Euclidian distance between those groups. The length of the time interval between two fixation groups serves as a filter for blinks. Evinger et al [32] reported maximum blink duration in the range of 75-425ms,therefore 75 ms. was employed in the merge function. The Euclidian distance between two fixation groups serves as a filter for micro-saccades (saccades with amplitude of less than 0.5º). The center of the merged fixation segment is calculated as a "centroid". The onset of the first fixation group becomes the onset or the beginning of the resulting fixation. The offset of the second fixation group becomes the offset (end point) of the fixation segment. Fixations with duration less than the minimum fixation duration (100ms) are then discarded from the analysis. On another note, consecutive eye position points that are classified as saccades are collapsed into a single saccade with specific onset and offset coordinates. Micro-saccades and saccades that contain eye positions not detected by an eye tracker device as a result of blinks or any other reason are discarded. The "Merge Function" is the same for every algorithm and provides the final output as a sequence of fixations and saccades.

The approach where each individual eye position is: first; classified as a part of a fixation or a saccade and second; processed by the "Merge Function", allows for a more standardized classification behavior, when compared to approaches in which the merging logic is incorporated into the initial classification stage (e.g. I-DT implementation in [28]).

Specific classification criteria that classifies each eye position sample as a fixation or saccade is described in the following sub-sections:

Velocity-Threshold Identification (I-VT) Model

In the I-VT model, the velocity value is computed for every eye position sample. The velocity value is then compared to the threshold. If the sampled velocity is less than the threshold, the corresponding eye position sample is marked as part of a fixation otherwise it is marked as a part of a saccade.

Hidden Markov Model Identification (I-HMM) Model

The Hidden Markov algorithm (I-HMM) is a more sophisticated version of the I-VT model that is augmented by the probabilistic representation of the Human Visual System (HVS). The I-HMM presented in this paper has two states, fixation and saccade. Each state is characterized by a velocity distribution in which the states represent the velocity distributions for saccade and fixation points.

There are three important process stages utilized in the I-HMM. The first stage of the I-HMM is identical to the I-VT, where each eye position sample is classified either as a fixation or a saccade depending on the velocity threshold. The second stage is defined by the Viterbi Sampler [33], where each eye position is re-classified as part of a fixation or saccade, depending on the probabilistic parameters (initial state, state transition and observation probability distributions) of the model. The goal of the Viterbi Sampler is to maximize the probability of the state assignment given probabilistic parameters of the model. The initial probabilistic parameters assigned to the I-HMM are typically not at optimal levels needing further improvement. Therefore, the third and last stage of the I-HMM is defined by the Baum-Welch re-estimation algorithm [34]. This algorithm re-estimates the initial probabilistic parameters and attempts to minimize errors in the state assignments. Parameter re-estimations can be performed by the Baum-Welch multiple times if necessary. In the I-HMM defined in this paper, the number of such re-estimations for optimization was four.

Kalman Filter Identification (I-KF) Model

The I-KF models an eye as a system with two states: position and velocity. The acceleration of the eye is modeled as white noise with fixed maximum acceleration. When applied to the recorded eye position signal, the I-KF generates a predicted eye velocity signal. The values of the measured and predicted eye velocity allow use of the Chi-square test to classify each eye positional sample as a part of a fixation or saccade.

$$\chi^2 = \sum_{i=1}^{p} \frac{(\hat{\theta}_i^- - \dot{\theta}_i)^2}{\delta^2} \quad (1)$$

where $\hat{\theta}_i^-$ is the predicted eye velocity computed by the Kalman filter and $\dot{\theta}_i$ is the observed eye velocity computed with the eye position signal from the eye tracker. δ is standard deviation of the measured eye velocity, respectively, during the sampling interval under consideration while p is the size of the temporal sampling window. Points above the specified χ^2 threshold are classified as part of a saccade while points below the threshold are classified as part of a fixation.

Minimum Spanning Tree Identification (I-MST) Model

Minimum spanning tree is defined as a spanning tree with a Euclidian distance minimum among all spanning trees in a given set of nodes. The I-MST algorithm builds a minimum spanning tree taking a predefined number of eye position points using Prim's algorithm. Eye fixations are characterized by a set of points that are enclosed in a relatively small region. With this in mind the I-MST traverses group of points and classifies each eye position point, into a fixation or a saccade based on point to point distance thresholds. Points below threshold are classified as a part of the fixation and points above the threshold are classified as a part of the saccade. The advantage of using an I-MST is the algorithm's ability to correctly identify fixation points even when a large part of the signal is missing due to noise. For longer eye movement recordings, the I-MST requires a sampling window to build a sequence of non-overlapping MST trees for meaningful classification results. The length of such a window can be equivalent to the duration of the largest saccade expected in the recording. In our experiments, the window size selected was 200ms.

Dispersion-Threshold Identification (I-DT) Model

The fifth and final model, Dispersion Threshold Identification (I-DT) algorithm, takes into account the distribution or spatial proximity of eye position points in the eye movement trace [28, 30]. The algorithm defines a temporal window, which moves one point at a time. The spatial dispersion created by the points within this window is compared against a threshold. If such dispersion is below the threshold, the points within the temporal window are classified as part of a fixation; otherwise, the window is moved by one

sample, and the first sample of the previous window is classified as a saccade. Starting size of the temporal window is held to a minimum fixation duration of 100 ms. The dispersion of the points in the window is computed with the formula D = [max(X) - min(X)] + [max(Y) - min(Y)], with X and Y representing eye position sets within the temporal window.

III. Qualitative and Quantitative Scoring of the Eye Movement Classification Algorithms

To establish a common basis for comparison between the five aforementioned classification algorithms, it was important to define a set of qualitative and quantitative scores for the assessment of classification algorithm performance. Assuming that a classification algorithm classifies the eye position trace into fixation and saccades, the following performance metrics were considered: Average Number of Fixations (ANF), Average Fixation Duration (AFD), Average Number of Saccades (ANS), and Average Saccade Amplitude (ASA). The performance of each classification algorithms could then be assessed by these metrics independent of stimulus activity. These metrics are well-known and have been employed by fields interested in documentation of oculomotor behavior such as usability sciences [23], psychology [22], and rehabilitation sciences [35].

To complement the above metrics, we developed three new metrics to classify quality of measured behavior - Fixation Quantitative Score (FQnS), Fixation Qualitative Score (FQlS), Saccade Quantitative Score (SQnS). Subsequently, these three metrics are identified as behavior scores.

A. Fixation Quantitative Score

The intuitive idea behind the Fixation Quantitative Score (FQnS) is to compare the amount of detected fixational behavior to the amount of fixational behavior encoded in the stimuli.

To calculate the FQnS the fixation stimulus position signal is sampled with the same frequency as the recorded eye position signal. Every resulting coordinate tuple (x_s,y_s,t) of fixation stimulus is then compared to the corresponding tuple (x_e,y_e,t) of the eye position recorded signal. If the recorded eye positional tuple is classified as a fixation with its centroid in a spatial proximity of the stimulus fixation (such proximity is determined by a specified threshold, which was 1/3 of the amplitude of a previous stimulus saccade for our purposes), then the fixation behavior detection counter is incremented by one. The FQnS is calculated by normalizing the resulting fixation behavior detection counter by the total amount of fixation positional points encoded in the stimulus.

$$\mathrm{FQnS} = 100 \cdot \frac{fixation\_detection\_counter}{stimuli\_fixation\_points} \qquad 1$$

According to such design, the FQnS compliments the AFD and ANF metrics, via measuring classified fixational behavior in regard to the temporal and spatial properties of the stimulus signal.

It is important to mention that practically speaking, the FQnS will never reach 100% due to the natural saccadic latency delay in the CNS required to send a neuronal signal to extraocular muscles to execute a saccade [2]. The average delay of 200ms. is reported in healthy humans [2]. In addition, the associated saccade duration approximates to

$$D_{sac\_dur} = (2.2A_{sac\_amp} + 21) \qquad 2$$

where A_{sac_amp} is the saccade's amplitude measured in degrees [36]. With this phenomena in mind, the onset of a fixation will always be delayed by a 200ms. plus the duration time of the saccade. Therefore, the computation of the ideal FQnS can be performed as:

$$\mathrm{Ideal\_FQnS} = 100 \cdot \left(1 - \frac{m \cdot S_l + \sum_{j=1}^{m} D_{sac\_dur_j}}{\sum_{i=1}^{n} D_{stim\_fix\_dur_i}}\right) \qquad 3$$

where n is the number of stimulus fixations, $D_{stim\_fix\_dur_i}$ is duration of the i^{th} stimulus fixation, S_l is saccadic latency, m is the number of the stimulus saccades, and $D_{sac\_dur_j}$ is the expected duration of a saccade in response to the stimulus saccade j.

B. Fixation Qualitative Score

The Fixation Quantitative Score (FQlS) compares the spatial proximity of the classified eye fixation signal to the presented stimulus signal, therefore indicating the positional accuracy or error of the classified fixations.

The FQlS calculation is similar to that of the FQnS, i.e., for every fixation related point (x_s,y_s) of the presented stimulus, the check is made for the data point in the eye position trace (x_e,y_e). If the data point is classified as a fixation the Euclidean distance between the presented fixation coordinates and the centroid of the detected fixation coordinates (x_c,y_c) is computed. The sum of such distances is normalized by the number of data points being compared.

$$\mathrm{FQlS} = \frac{1}{N} \cdot \sum_{i=1}^{N} fixation\_distance_i \qquad 4$$

N is the number of stimulus position points where the stimulus fixation state is matched with each corresponding eye position sample detected as a fixation. $fixation\_distance_i = \sqrt{(x_s^i - x_c^i)^2 + (y_s^i - y_c^i)^2}$ represents the distance between stimulus position and the center of the detected fixation.

Ideally, the FQlS should equal to 0º, which can only happen in the case of absolute accuracy of the eye tracking equipment and assuming that subjects make very accurate saccades to the fixation position. In practice, the accuracy of modern eye trackers remains in the range of <0.5º. Typically, even normal eye movement behavior incorporates undershoots/overshoots when making saccades to fixation targets [2], making the initial segment of exhibited fixation slightly off-target. Additionally each fixation is composed of three sub-movements: tremor, drift, and micro saccades [37] with each sub-movement introducing additional noise. Therefore, we hypothesize that practical values for the FQlS should be at best around 0.5º.

C. Saccade Quantitative Score

The Saccade Quantitative Score (SQnS) represents the amount of classified saccadic behavior given the amount of saccadic behavior encoded in the stimuli. The SQnS is an important addition to the ASA and the ANS metrics, because it correctly quantifies saccadic behavior even in situations where complex oculomotor events such as

undershoots/overshoots, dynamic saccades, express saccades, compound saccades are present [2]. Such oculomotor events can skew resulting numbers for the ASA and ANS, however does not directly interfere with computations for the SQnS.

To calculate the SQnS, two separate quantities are computed. One represents the amount of stimulus saccadic behavior and the second represents the amount of classified saccadic behavior. To calculate the stimulus related metric, each jump of the fixation target to a new location is considered as a stimulus saccade and the distance difference between targets indicates the stimulus saccade amplitude. The absolute values of the amplitudes of all stimulus saccades are summed together to produce $total\_stimuli\_saccade\_amplitude$. Similarly, absolute values of all response saccade amplitudes detected by a given classification algorithm are summed together to represent the accumulative amplitude for the classified saccadic behavior $total\_detected\_saccade\_amplitude$. The following formula presents the computation of the ratio score:

$$\mathrm{SQnS} = 100 \cdot \frac{total\_detected\_saccade\_amplitude}{total\_stimuli\_saccade\_amplitude} \quad (5)$$

The SQnS of 100% indicates that the integral sum of detected eye saccade amplitudes equals that of the presented stimuli. The SQnS can be larger than 100%, which essentially can occur due to two things: abnormal saccadic behavior of the sample or the classification algorithm has incorrectly amplified saccadic behavior, i.e., some fixations classified as saccades. An example of abnormal saccadic behavior could occur when a sample contains a large number of hypermetric saccades (target overshoots) followed by glissades (post saccadic drifts) and possibly saccadic intrusions or oscillations (inappropriate movements that take the eye away from the target during attempted fixation [2]). In addition, the amplification of the saccadic behavior would be caused by the inappropriate selection of a threshold classification parameter. The SQnS would be smaller than 100% in cases of hypometric saccadic behavior (target undershoots) or damping behavior of the classification algorithm.

In view of the foregoing, we were now able to employ seven different assessment metrics ANF, AFD, ANS, ASA, SQnS, FQnS, FQlS, SQnS to provide a performance comparison of the five eye movement classification models of interest.

IV. Methodology

A. Procedure

Oculomotor behaviors were recorded using the Tobii x120 eye tracker [38], which includes a standalone unit connected to a 24-inch flat panel screen with a resolution of 1980 x 1200 pixels. The eye tracker performed binocular tracking with the following characteristics: accuracy 0.5°, spatial resolution 0.2°, drift 0.3° with eye position sampling frequency of 120Hz. A chin rest was used to stabilize the head for higher accuracy and stability in eye tracings.

B. Participant Data Samples

A total of 22 participants (9 males/ 13 females), ages 18 – 25 years with an average age of 21.2 (+/-3.12), volunteered for the project from the Texas State University campus. Participants were chosen from a larger data pool from a larger study using the following inclusion criteria to ensure high quality data: positional accuracy during calibration better than 1.70º and invalid data percentage of the data less than 20%. The resulting data pool had a calibration error mean of 1.01º ± 0.41 with only a resulting mean of 3.23% ± 2.26 for invalid data.

C. Fixation & Saccade Invocation Task

A stimulus was presented as a white, single 'jumping point' on a black background with vertical coordinates fixed to the middle region of the screen. The size of the point was approximately 1º of the visual angle with the center marked as a black dot. The first point was presented in the center of the screen, then subsequent 14 points moved to the left and right of center with a spatial amplitude of 10-20º. Therefore, the jumping sequence consisted of 15 total fixation points including the original point in the center. The first saccade of 10º and the 13 subsequent saccades of 20º, resulted in an overall average saccade amplitude of 19.3º to represent all 14 saccades. After each subsequent jump, the fixation point remained stationary for 1s before the next jump.

Considering the simple stimulus behavior and the normal subject pool, the following metrics introduced in Section III were set up as ideal metric performance: AFN=15 fixations, AFD=1s, ANS=14 saccades, ASA=19.3º, FQlS=0º, FQnS=82.3%, SQnS=100%.

D. Threshold Range Selection

It was important to test the performance of each classification algorithm over a sensible range of the threshold values. Such range was selected based on the research literature recommendations and physiological eye movement properties.

The range of the threshold values for the velocity based models (I-VT, I-HMM) was set from 5º/s [2] to 300º/s [28]. For the dispersion/position based models (I-MST, I-DT) the threshold range was set from 0.033º to 2º. Such range was selected to test the performance between two extremities: 0.033º is the value that represents a minimum common amplitude of involuntary saccades exhibited during a fixation [39] and 2º is the value often suggested by the eye tracking vendors [38].Various recommendations exist for the threshold values for the I-KF method: 5 [40], 20.5 [29], 50 [41]. We selected the range from 1 to 60 to include suggested values and to investigate the performance of the I-KF on the boundaries of the suggested range.

To assess the performance of the selected classification algorithms on the same scale a range coefficient concept is introduced. Given the value of the range coefficient, the threshold value for a classification algorithm can be found as:

$$\mathrm{T} = RC \cdot Inc + C \quad (6)$$

where T is the threshold, RC is the range coefficient, C is the initial value of the threshold, Inc is the threshold's increment. For the I-VT and I-HMM C=5º/s, Inc=5º/s. For the I-MST and I-DT C =0.033º, Inc =0.033º, for the I-KF C =1, Inc =1. Range coefficient from 0 to 59, allows for testing the algorithms over threshold ranges discussed above with sufficient granularity without introducing too many data points.

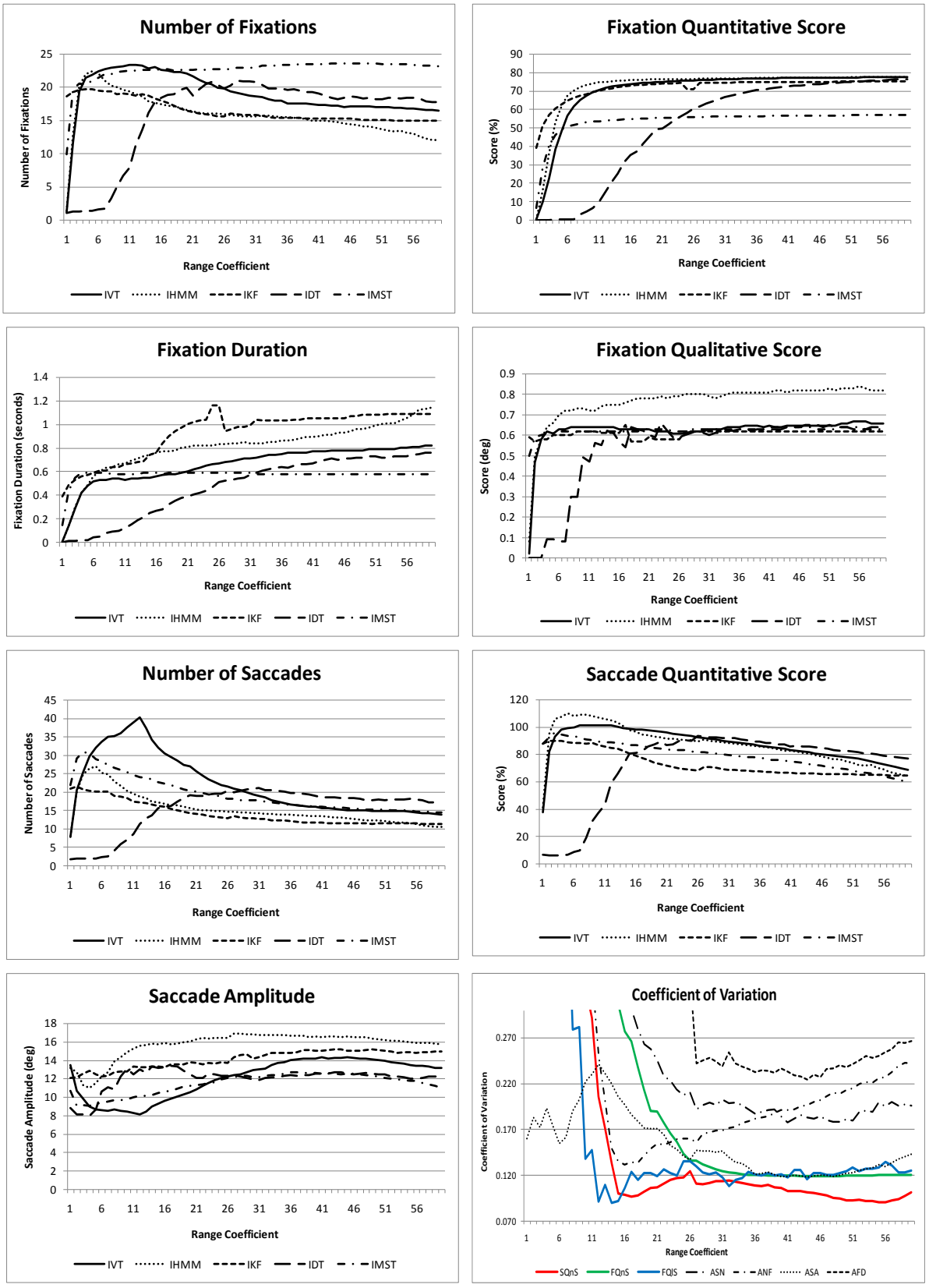

Fig. 2. Results indicating the aggregate of each oculomotor behavior for 22 human subjects using the five automated classification algorithms. a. Average Number of Fixations (ANF), b. Average Fixation Duration (AFD), c. Average Number of Saccades (ANS), d. Average Saccade Amplitude (ASA), e. Fixation Quantitative Score (FQnS), f. Fixation Qualitative Score (FQlS), g. Saccade Quantitative Score (SQnS), h. Coefficient of Variation for all metrics/scores.

V. Results

Fig. 2 presents classification results.

A. Average Number of Fixations (ANF)

The trend for each classification algorithm was a low ANF score at small threshold values with very rapid increase to a certain threshold level. Afterwards, there was a reverse trend in ANF numbers down to a specific threshold value where performance of all algorithms stabilized. The I-MST method provided the highest number of fixations (24) and the I-HMM the lowest (12); therefore there was a twofold difference between these two methods. All methods were able to reach the number of fixations presented by the stimulus signal (15), while some algorithms were able to reach this value twice.

B. Average Fixation Duration (AFD)

Each classification algorithm trend yielded a low AFD at small threshold values and then rapidly increased in AFD values up to a certain threshold. After this threshold was reached, the AFD value increase leveled off or saturated. The difference between algorithms was significant even when AFD values stabilized. The I-KF provided maximized AFD values while the I-MST provided values on average reduced by 50%. Only two algorithms, the I-KF and the I-HMM were able to achieve oculomotor fixation duration values encoded in the presented stimulus (1s).

C. Average Number of Saccades (ANS)

The trend for each classification algorithm indicated a low ANS with the use of a small threshold values, and then exhibited peak performance at a specific plateau with a specific saturation point at the high threshold values, which was close to the stimulus signal (14 saccades). For the highest threshold values, the I-DT provided the highest ANS of 17 while the I-HMM provided the lowest ANS value of 10. (See Fig. 2.c).

D. Average Saccade Amplitude (ASA)

None of the methods indicated average saccade amplitude at levels presented by the stimulus signal (19.3º). The highest value of 17º was reached using the I-HMM model, with the lowest value of 8º provided by both the I-VT and I-DT models. The saturated ASA value produced by peak threshold values yielded a difference of more than 5º. (See Fig. 2.d)

E. Fixation Quantitative Score (FQnS)

The FQnS monotonically increased for all classification algorithms. For all algorithms except the I-DT there was an immediate jump in values, however, there was a point of saturation after a certain threshold value or no increase in detected fixational behavior. All algorithms peaked with a FQnS score of 74-77% which is agreeable with physiological latencies discussed in Section III. The I-MST algorithm was the outlier providing the saturated FQnS score of 57% which was approximately 23% lower than the other algorithm FQnS percentages (Fig. 2.e).

F. Fixation Qualitative Score (FQlS)

The performance of all four (I-VT, I-KF, I-DT, I-MST) algorithms was very similar in terms of the positional accuracy of the detected fixation, with the I-KF model providing a slightly lower score, thus showing higher accuracy in terms of coordinates of the detected fixation location (previous study supports this fact showing 10% improvement in accuracy when the I-KF was compared to the I-VT in a real-time eye-gaze guided system [41]. The I-HMM provided the score that was essentially 33% higher than the other models indicating significantly less accuracy in fixation coordinate detection (See Fig. 2.f).

G. Saccade Quantitative Score (SQnS)

Each algorithm had a point of maximum SQlS performance after which the score values monotonically decreased. After the SQlS score peak is reached, the amount of saccadic behavior goes down because a lesser amount of eye position samples are classified as saccades. (See Fig. 2.g). This peak value was highest for the I-HMM algorithm with a value of approximately 110% and lowest for the I-KF with a value of 90%. The performance of the I-MST and I-DT was slightly higher than the performance of the I-KF. For the upper values for performance thresholds of the I-VT, I-DT and the I-HMM were very close to each other. The I-KF provided the most damping behavior in terms of detected saccades. The difference in performance between each individual algorithm did not exceed 22%.

VI. Discussion

A. Advantages of the new Qualitative and Quantitative Behavior Assessment Scores

The Fixation Qualitative Score (FQlS) was found to be extremely useful in measuring the positional accuracy of the classified fixations given a threshold value (See Fig. 2.f).

The Fixation Quantitative Score (FQnS) depicted fixational behavior that was much less "noisy" than that of the data provided by the Average Fixation Duration (AFD) and Average Number of Fixations (ANF). This was observed for the I-VT, I-DT, I-HMM and I-KF models that produced varying behavior in terms AFD and ANF values, however all essentially converged in the FQnS as a summary score. With this in mind, the FQnS ensures the temporal validity of the sampled fixations by matching them with the stimulus signal. In so doing, the FQnS is set up to pick out classification disadvantages of a particular algorithm, e.g., I-MST's spurious fixations due to possible overlapping data.

The Saccade Quantitative Score (SQnS) proved helpful identifying specific input parameters (thresholds) that allowed detection of similar saccadic behavior as presented by the stimulus. This was not entirely possible with the Average Number of Saccades (ANS) and Average Saccade Amplitude (ASA) metrics. In cases when subjects required multiple saccades to reach target's position, the ANS values were smaller or larger than their ideal values. As a result, selection of the input threshold based on the ANS and the ASA alone

was quite difficult, but the SQnS actually resolved the ambiguity by indicating the amount of saccadic behavior given a specific threshold value.

Additionally, the SQnS and FQnS values demonstrated much less variability than computed by the Coefficient of Variation (CV) formula: CV=δ/μ. Where μ is the mean value between the values of a metric/score calculated by all classification methods computed for a fixed Range Coefficient (RC), δ is the standard deviation. Fig. 2.h represents the CV results computed for all metrics and thresholds. It is possible to see that after a certain value (RC=26), the CV for SQnS and FQnS were substantially less than values for the ASN, ANF, ASA and AFD metrics. Such behavior indicated that the SQnS and FQnS were more stable and therefore more suitable in establishing the baseline comparisons between all five of the classification methods.

B. Meaningful Threshold Selection via Behavior Scores

Behavior scores allow "calibration" of the performance of any eye-movement classification algorithm given the stimulus has pre-set characteristics, e.g., stimulus type defined in Section IV.C. We use the term "calibration" in the sense of a reasonable threshold selection for a specific classification algorithm and subject/experiment setup. All commercial eye tracking equipment requires a calibration procedure in a manner of step stimulus depicted as a sequence of two or more jumping points to be able to compute the location of the eye gaze during actual recording. We suggest use of the recorded eye positional data from this already established procedure to "calibrate" the performance of classification algorithm.

To obtain a reasonable threshold, the meaningful initial threshold range should be selected. Range selection can be based on the physiological considerations of the eye movement classification method, e.g., logic presented in Section IV.D. In case physiological considerations are unavailable, the selected threshold range should result in a wide range of meaningful SQnS values e.g. 0-150%. For the same threshold range, the FQnS should also result in a range of meaningful values, e.g., 0-100%. Healthy human subjects are expected to have similar saccadic behavior as encoded in the stimulus, therefore the threshold value that yields selection of an ideal SQnS value (or closest) can be considered as meaningful for a specific algorithm. At the same time corresponding FQnS and FQlS values must be meaningful without variance too far from their ideal values. If the ideal SQnS value is achieved more than once, a threshold must be selected that provides an FQnS value closer to the ideal value. If several thresholds result in the same FQnS value, a threshold must be selected that provides the best FQlS value.

Following this recommendation and considering that ideal SQnS=100% and ideal FQnS=73.4%, optimal thresholds for two classification algorithms were identified in our experimental setup: I-VT - threshold of 70°/s (SQnS =100%, FQnS =73%, FQlS =0.64°), I-HMM - threshold of 70°/s (SQnS =101%, FQnS =75%, FQlS =0.75°).

Selection of the threshold for the I-KF, I-DT, I-MST is more challenging, due to the dampening effect of these algorithms on the classified saccadic performance - none were able to achieve a SQnS value of 100%. Additionally, maximum SQnS values achieved did correspond to the FQnS that is 22-42% lower than the ideal value. In cases when accurate saccadic classification is not possible it makes sense to stabilize fixation behavior by considering first the FQnS that is not too far from the ideal value (the maximum difference of up to 15% is reasonable) and matching the threshold with corresponding SQnS values that are also not too far from the ideal value (difference of up to 20% is reasonable for saccades). There might be a case when for all thresholds, within the selected range, the SQnS and the FQnS values are quite far from their ideal values. In such scenarios it is important to select a different threshold range/input parameters or/and classification algorithm to ensure meaningful classification performance. The above outlined logic allows for the threshold selections for the I-KF - threshold of 15 (FQnS = 72%, SQnS = 80%, FQnl = 0.61) and the I-DT - threshold of 1.35° (FQnS = 72%, SQnS = 86%, FQnl = 0.63). The I-MST presents the case where the algorithm almost fails to provide meaningful classification results, i.e., the FQnS is more than 15% smaller than the ideal FQnS score of 73.4% for the whole threshold range. If the best threshold has to be selected for the I-MST, one strategy would be to identify the onset of the FQnS saturation behavior with the associated SQnS value close to ideal, resulting in the threshold of 0.6° (SQnS = 85.3%, FQnS = 55.2%, FQlS = 0.62).

C. Criterion-based Baseline for the Assessment of the Oculomotor Behavior

We define the baseline as a fixed set of thresholds that allow comparison of classification performance between various eye movement classification algorithms in a meaningful way. Conventional metrics such as ANF, AFD, ANS, and ASA do not provide the means to achieve this goal, e.g., if one of these metrics is fixed to an ideal value the amount of the variability is high in the remaining metrics. To illustrate this argument, we selected a RC range of 14-60 with relatively stable behavior across all metrics (CV<0.5) and fixed the Average Number of Saccades (ANS) to the ideal number ANS=14. Table I presents the results.

Table I. Coefficient of Variation with ANS=14°

Classification Algorithm	Trheshold	RC	SQnS	FQnS	FQlF	ASA	ANF	AFD
I-VT	300	60	68.8	77.5	0.7	13.2	16.5	0.8
I-HMM	160	32	87.7	76.7	0.8	16.7	15.7	1.0
I-KF	26	26	71.8	73.9	0.6	13.7	16.2	1.2
I-DT	0.462	14	72.5	25.4	0.6	13.4	16.1	0.2
I-MST	1.98	60	59.2	57.1	0.6	11.0	23.1	0.6
Coefficient of Variation	***N/A***	***N/A***	***0.143***	***0.357***	***0.115***	***0.150***	***0.180***	***0.494***

The CV exhibited high amount of variability ranging from 0.115 to 0.494.

We propose a heuristic to compute a comparison baseline following similar logic to that used for the threshold selection outlined in the previous sub-section. Saccadic behavior represents the amount of movement (variability) of the eye, therefore the common ideal SQnS value between all algorithms must be selected first. To reduce possible variability of classification results the threshold range that yields a coefficient of variation of SQnS of 0.2 between classification algorithms is suggested (Our results indicated an RC of 14 to 60 for this range). As it stated previously, not all eye movement classification algorithms can achieve an ideal SQnS value. Therefore, we suggest selection of the

largest SQnS value achieved by all the classification algorithms. The largest SQnS should not exceed 100% for normal subjects. When the largest SQnS value is selected with corresponding FQnS and FQlS values must also be meaningful, i.e., not too far from their ideal values. If the algorithm achieves a maximum SQnS value more than once select the threshold that provides the FQnS closer to the ideal value. If several thresholds result in an ideal FQnS value, selection of the threshold is based on the best achieved FQlS value. Resulting fixed set of thresholds that yield above mentioned oculomotor behavior serve as the baseline for each corresponding algorithm. Note, when algorithms selected for the baseline analysis employ different threshold units the concept of the Range Coefficient introduced in the Section IV.D can be employed to bring them to the same scale and to investigate the coefficient of variability.

Proposed heuristic produces meaningful behavior in terms of the classification performance and yields low variability across various metrics as illustrated by Table II.

In case of our experiment the amount of variability in the remaining metrics goes down to 0.108-0.149 when the SQnS is fixed to 84% (maximum SQnS value achievable by all classification algorithms).

Table II Coefficient of Variation with SQnS=84%

Method Name	Trheshold	RC	FQnS	FQlS	ANS	ASA	ANF	AFD
I-VT	200	40	76.9	0.6	16.0	14.1	17.4	0.8
I-HMM	195	39	76.9	0.8	13.5	16.5	15.2	0.9
I-KF	13	13	71.3	0.7	17.0	13.3	19.0	0.7
I-DT	1.551	47	74.1	0.7	18.1	12.6	18.5	0.7
I-MST	0.66	20	55.4	0.6	20.3	11.2	22.6	0.6
Coefficient of Variation	***N/A***	***N/A***	***0.127***	***0.108***	***0.149***	***0.145***	***0.146***	***0.146***

Once common baseline is derived it is possible to measure the classification accuracy of the algorithms based on the absolute difference between the ideal and the actual metric values. Table III presents the results, bold numbers highlight smallest differences.

Table III Absolute difference between classified and ideal metric values at the baseline

Method Name	Trheshold	RC	FQnS	FQlS	ANS	ASA	ANF	AFD
I-VT	200	40	5.38	0.64	1.95	5.19	2.41	0.24
I-HMM	195	39	**5.37**	0.81	**0.50**	**2.82**	**0.18**	**0.12**
I-KF	13	13	10.98	0.68	3.00	6.04	3.95	0.31
I-DT	1.551	47	8.24	0.65	4.09	6.70	3.50	0.29
I-MST	0.66	20	26.89	**0.63**	6.32	8.09	7.59	0.41

Table IV. Absolute difference between classified and ideal metric values at the baseline

The I-HMM algorithm provided classification results that were the closest in terms of the ideal behavior encoded in the stimulus, therefore posing itself as the most behaviorally accurate at this baseline. At the same time, the I-HMM algorithm had highest positional error between the classified and presented fixation stimulus, but such error was smaller than the average calibration error of 1.01º reported in the Section IV.B.

It is possible to imagine a case when one algorithm will achieve minimum differences in just one category of metrics with remaining algorithms achieving minimums in the remaining categories. In such cases the definition of the "best" algorithm should be defined by the researcher with importance of the resulting behavior assessed via the scope of the specified task and the goals of the experiment.

D. Automated vs Manual Classification

Manual techniques are frequently employed to classify eye movement behavior [4]. However, this type of classification technique is susceptible to human error and can be open for biased interpretation with limited generalizability. Additionally, it becomes extremely tedious and time consuming to analyze large quantities of data.

One might conclude that reliable automation of classification of eye movements is impossible based on the high variability in eye behavior even with presentation of a simple stimulus and variation of a single metric threshold [7]. However, behavioral classification for eye movements as introduced in this paper provides a more stable point of reference based on stimulus behavior and therefore uses a standardized criterion-based or meaningful threshold selection to support automated classification methods.

E. Limitations

1) Sampling Frequency

It can be argued that the sampling frequency of 120Hz (8.3 ms. per eye sample) employed in our studies can be considered as low for detection of saccadic behavior by some researchers. A sampling frequency of 120Hz translates to approximately 3 data points for 1º and 10 data points for 30º saccades (2). Average saccade's amplitude encoded in the stimulus was 19.3º providing approximately 8 data points for each recorded saccade. Our previous research indicates that 120Hz sampling frequency is sufficient to classify basic and complex saccadic eye movement behaviors such as express saccades, dynamic overshoots, simple/corrected undershoots/overshoots, and compound saccades [42]. However, it is still understood that a low sampling frequency would prevent a reliable analysis of extremely small saccades with amplitudes of less than 0.2, and sub-movements during fixations such as drift and tremor.

2) Blink Detection

Simple criteria for blink removal presented in Section II.A worked well for 120Hz eye movement data which was obtained with x120 Tobii eye tracker. The employment of the higher sampling frequency equipment or the use of a different eye tracker might require an introduction of a more sophisticated blink removal algorithm to ensure meaningful classification behavior.

VII. Conclusion

This paper provided a comprehensive overview of five major eye-movement classification algorithms in terms of the assessment of oculomotor saccadic and fixation behavior. The results indicate that even in case of fixed stimulus behavior and alternation of a single threshold parameter the classification results differ dramatically with differences up to 100% or greater.

The paper analyzed actual data to introduce a set of standardized qualitative and quantitative scores that provided more stable algorithm performance in terms of fixation and saccade classification. Standardized scores allowed for a proposed logic for developing a criterion-based baseline for comparison between any classification algorithms. In addition, we provided logic for meaningful threshold selection for any

eye-movement classification algorithm. Such logic would be extremely beneficial for the eye tracking practitioners and eye tracking vendors as a tool for the selection of the input parameters, including thresholds, that would allow to assure reasonable classification behavior given specific equipment, software, experiment setup, and subject.

In view of the numerous advantages of standardized automatic analysis systems, future studies are necessary to explore the feasibility of this scoring system for clinical applications.

VIII. References

[1] M. Köllensperger, F. Geser, K. Seppi *et al.*, "Red flags for multiple system atrophy," *Movement Disorders*, vol. 23, no. 8, pp. 1093-1099, 2008.

[2] R. J. Leigh, and D. S. Zee, *The Neurology of Eye Movements*: Oxford University Press, 2006.

[3] N. Smyrnis, "Metric issues in the study of eye movements in psychiatry. ," *Brain Cogn. 2008* vol. 3, no. Dec;68, pp. 341-58, , 2008.

[4] D. C. Richardson, & Spivey, M. J. , *Eye Tracking: Characteristics and Methods. Encyclopedia of Biomaterials and Biomedical Engineering*: Marcel Dekker, Inc., 2004.

[5] J. R. Heitger MH, Macleod AD, Snell DL, Frampton CM, Anderson TJ., "Impaired eye movements in post-concussion syndrome indicate suboptimal brain function beyond the influence of depression, malingering or intellectual ability.," *Brain*, no. Oct;132(Pt 10), pp. 2850-70. , Epub 2009 Jul 16., 2009.

[6] U. P. Mosimann, R. M. Muri, D. J. Burn *et al.*, "Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies," *Brain*, vol. 128, no. 6, pp. 1267-1276, June 1, 2005, 2005.

[7] N. Smyrnis, "Metric issues in the study of eye movements in psychiatry," *Brain and Cognition*, vol. 68, no. 3, pp. 341-358, 2008.

[8] X. Radvay, S. Duhoux, F. Koenig-Supiot *et al.*, "Balance training and visual rehabilitation of age-related macular degeneration patients," *Journal of Vestibular Research*, vol. 17, no. 4, pp. 183-193, 2007.

[9] D. P. Munoz, I. T. Armstrong, K. A. Hampton *et al.*, "Altered Control of Visual Fixation and Saccadic Eye Movements in Attention-Deficit Hyperactivity Disorder," *J Neurophysiol*, vol. 90, no. 1, pp. 503-514, July 1, 2003, 2003.

[10] E. Isotalo, A. Heikki, and P. Ilmari, "Oculomotor findings mimicking a cerebellar disorder and postural control in severe Meniere's disease," *Auris Nasus Larynx*, vol. 36, no. 1, pp. 36-41, 2009.

[11] L. Crevits, M. C. Hanse, P. Tummers *et al.*, "Antisaccades and remembered saccades in mild traumatic brain injury," *Journal of Neurology*, vol. 247, no. 3, pp. 179-182, 2000.

[12] G. Cockerham, and E. D. W. Gregory L. Goodrich, James C. Orcutt, Joseph F. Rizzo, Kraig S. Bower, Ronald A. Schuchard, "Eye and visual function in traumatic brain injury," *Journal of Rehabilitation Research & Development*, vol. 46, no. 6, pp. 811 - 818, 2009.

[13] O. V. Komogortsev, U. K. S. Jayarathna, D. H. Koh *et al.*, "Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms." pp. 1-4.

[14] R. J. K. Jacob, "What you look at is what you get: eye movement-based interaction techniques," in Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people, Seattle, Washington, United States, 1990.

[15] S. Zhai, C. Morimoto, and S. Ihde, "Manual and gaze input cascaded (MAGIC) pointing," in Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, Pittsburgh, Pennsylvania, United States, 1999.

[16] L. E. Sibert, and R. J. K. Jacob, "Evaluation of eye gaze interaction," in Proceedings of the SIGCHI conference on Human factors in computing systems, The Hague, The Netherlands, 2000.

[17] O. V. Komogortsev, and J. Khan, "Kalman Filtering in the Design of Eye-Gaze-Guided Computer Interfaces." pp. 1-10.

[18] D. Parkhurst, J., and E. Niebur, "Variable resolution displays: A theoretical, practical, and behavioral evaluation," *Human Factors*, vol. 44, no. 4, pp. 611-629, 2002.

[19] A. Duchowski, T., and A. Çöltekin, "Foveated gaze-contingent displays for peripheral LOD management, 3D visualization, and stereo imaging.," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 4, 2007.

[20] K. Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372-422, 1998.

[21] M. Field, K. Mogg, and B. Bradley, "Eye movements to smoking-related cues: effects of nicotine deprivation," *Psychopharmacology*, vol. 173, no. 1, pp. 116-123, 2004.

[22] N. Ceballos, O. Komogortsev, and G. M. Turner, "Ocular Imaging of Attentional Bias Among College Students: Automatic and Controlled Processing of Alcohol- Related Scenes," *Journal of Studies on Alcohol and Drugs, September*, pp. 1-8, September 2009, 2009.

[23] A. Poole, and L. J. Ball, "Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects," *Encyclopedia of Human Computer Interaction*: Idea Group, 2004.

[24] O. Komogortsev, C. Mueller, D. Tamir *et al.*, "An Effort Based Model of Software Usability," in International Conference on Software Engineering Theory and Practice (SETP), Orlando, FL, 2009, pp. 1-8.

[25] C. Ehmke, and S. Wilson, "Identifying web usability problems from eye-tracking data," in Proceedings of the 21st British CHI Group Annual Conference on HCI 2007: People and Computers XXI: HCI...but not as we know it - Volume 1, University of Lancaster, United Kingdom, 2007.

[26] S. P. Newman-Toker DE, Chowdhury M, Clemons TM, Zee DS, Della Santina CC., "Penlight-cover test: a new bedside method to unmask nystagmus.," *J Neurol Neurosurg Psychiatry*, vol. 80(8), no. Aug, pp.:900-3. , 2009.

[27] Y. H. Ramey NA, Irsch K, Müllenbroich MC, Vaswani R, Guyton DL, "A novel haploscopic viewing apparatus with a three-axis eye tracker.," *J AAPOS*, no. Oct;12(5):, pp. 498-503. , 2008.

[28] D. D. Salvucci, and J. H. Goldberg, "Identifying fixations and saccades in eye tracking protocols." pp. 71-78.

[29] D. Sauter, M. B. J., N. Di Renzo *et al.*, "Analysis of eye tracking movements using innovations generated by a Kalman filter," *Med. Biol. Eng. Comput.*, pp. 63–69, 1991.

[30] A. Duchowski, *Eye Tracking Methodology: Theory and Practice*, 2nd ed.: Springer, 2007.

[31] O. V. Komogortsev, U. K. S. Jayarathna, D. H. Koh *et al.*, *Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms*, Texas State University - San Marcos, San Marcos, , 2009.

[32] C. Evinger, K. Manning, and P. Sibony, "Eyelid movements. Mechanisms and normal data," *Invest. Ophthalmol. Vis. Sci.*, vol. 32, no. 2, pp. 387-400, February 1, 1991, 1991.

[33] G. D. Forney, Jr., "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268-278, 1973.

[34] L. E. Baum, T. Petrie, G. Soules *et al.*, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171, 1970.

[35] S. Garbutt, Y. Han, A. N. Kumar *et al.*, "Vertical Optokinetic Nystagmus and Saccades in Normal Human Subjects," *Invest. Ophthalmol. Vis. Sci.*, vol. 44, no. 9, pp. 3833-3841, September 1, 2003, 2003.

[36] R. H. S. Carpenter, *Movements of the Eyes*, London: Pion, 1977.

[37] L. Yarbus, *Eye Movements and Vision*, Moscow: Institute for Problems of Information Transmission Academy of Sciences of the USSR, 1967.

[38] Tobii. "Tobii technology," http://www.tobii.com.

[39] L. Yarbus, "Eye Movements and Vision," Moscow: Institute for Problems of Information Transmission Academy of Sciences of the USSR, 1967.

[40] T. Grindinger, "Eye Movement Analysis & Prediction with the Kalman Filter," Computer
Science, Clemson University, Clemson, 2006.

[41] D. H. Koh, S. A. M. Gowda, and O. V. Komogortsev, "Input evaluation of an eye-gaze-guided interface: kalman filter vs. velocity threshold eye movement identification." pp. 197-202.

[42] O. V. Komogortsev, D. Gobert, and Z. Dai, *2010*, Texas State University - San Marcos, San Marcos, Classification Algorithm for Saccadic Oculomotor Behavior. Technical Report. http://ecommons.txstate.edu/cscitrep/18/.