Department of Computer Science
San Marcos, TX 78666

Report Number TXSTATE-CS-TR-2009-16

Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms

Oleg V. Komogortsev
Sampath Jayarathna
Do Hyong Koh
Sandeep Munikrishne Gowda

2009-09-16

Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms

Oleg V. Komogortsev
Department of Computer Science
Texas State University-San Marcos
ok11@txstate.edu

Sampath Jayarathna
Department of Computer Science
Texas State University-San Marcos
sampath@txstate.edu

Do Hyong Koh
Department of Computer Science
Texas State University-San Marcos
dk1132@txstate.edu

Sandeep Munikrishne Gowda
Department of Computer Science
Texas State University-San Marcos
sm1499@txstate.edu

Abstract

This paper presents a set of qualitative and quantitative scores designed to assess performance of the various eye movement classification algorithms. The scores are designed to provide a foundation for the eye tracking researchers to communicate about the performance validity of various eye movement classification algorithms. The paper concentrates on the five algorithms in particular: Velocity Threshold Identification (I-VT), Dispersion Threshold Identification (I-DT), Minimum Spanning Tree Identification (MST), Hidden Markov Model Identification (I-HMM) and Kalman Filter Identification (I-KF). The paper presents an evaluation of the classification performance of each algorithm in the case when values of the input parameters are varied. Advantages provided by the new scores are discussed. Discussion on what is the "best" classification algorithm is provided for several applications. General recommendations for the selection of the input parameters for each algorithm are provided.

CR Categories: I.6.4 [Simulation and Modeling]: Model Validation and Analysis; J.7 [Computers in Other Systems]: Process control, Real time.

Keywords: Eye movements, classification, algorithm, analysis, scoring, metrics.

1 Introduction

Accurate eye movement classification is a fundamental necessity in the field of eye tracking. Almost every experiment that involves an eye tracker as a measurement or interaction tool requires an eye movement classification algorithm for data reduction and/or analysis. The main role of any eye movement classification algorithm is to break eye position temporal stream into basic eye movement types, as well as provide a set of characteristics about each eye movement type detected. In general, there are six major eye movement types: fixations, saccades, smooth pursuits, optokinetic reflex, vestibule-ocular reflex, and vergence [Leigh& Zee 2006]. Fixations and saccades are the types of most researched eye movements that are employed in human computer interaction (Jacob 1990; Zhai, Morimoto et al. 1999; Sibert and Jacob 2000; Parkhurst and Niebur 2002; Duchowski and Çöltekin 2007; Komogortsev and Khan 2007), psychological studies and reading (Rayner 1998), (Field, Mogg et al. 2004), (Ceballos, Komogortsev et al. 2009), medical studies (Garbutt, Han et al. 2003; Suh, Basu et al. 2006), and usability studies (Poole and Ball 2004), (Ehmke and Wilson 2007; Komogortsev, Mueller et al. 2009). With great simplifications, their roles are described as follows: fixation – eye movement that keeps an eye gaze stable in regard to a stationary target providing visual pictures with highest acuity, saccade – a very rapid eye rotation moving the eye from one fixation point to another, while pursuit stabilizes retina in regard to a moving object of interest (Duchowski 2007).

The development of the eye movement classification algorithms has a long history (McConkie 1980; Widdel 1984; Sauter, J. et al. 1991; A. Varri, B. Kemp et al. 1995; Tigges, Kathmann et al. 1995; Salvucci and Goldberg 2000; Munn, Stefano et al. 2008). Almost every eye movement classification algorithm has a set of input parameters that can significantly impact the result of classification. A large number of the eye tracking studies selects the input parameters for the classification algorithms empirically without a discussion of how the selection of those parameters affects the outcome of the classification. The first goal of this paper is to provide a set of quantitative and qualitative metrics that allow assessment of the performance of any eye movement classification algorithm. The second goal of this paper is to provide an evaluation of the performance of the major classification algorithms employed in the eye tracking field today. This paper also aims to provide a discussion on how the selection of input parameters affects the performance of the algorithm in terms of the proposed metrics. The third goal of this paper is to select the "best" classification algorithm for a specific application.

2 Eye Movement Classification Algorithms

The pseudocodes for the algorithms discussed below are presented in Figure 1. All pseudocodes presented in this paper are designed for an off-line eye position signal classification.

2.1 I-VT

In the I-VT model, velocity value is computed for every eye position sample. The velocity value is compared to the threshold. If the sampled velocity is smaller than the threshold the corresponding eye position sample is marked as part of a saccade, otherwise the eye position sample is assigned to be a part of a fixation (Salvucci and Goldberg 2000). Next, Merge_Function()

performs the following work: consecutive eye position points that are classified as a part of fixation are collapsed into a single fixation segment with center coordinates computed as a centroid of all points in the fixation. Classified fixations are subsequently merged into larger fixations by the criteria based on two parameters: length of the time interval between two fixation groups (<75ms) and the Euclidian distance between those groups (<0.5º). The center of the merged fixation segment is calculated as 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 which have a duration less than the minimum fixation duration (100ms) are discarded from the analysis. On another note, consecutive eye position points that are classified as saccades are collapsed into a single saccade with onset and offset coordinates. Micro saccades with amplitudes of less than 0.5º and saccades that contain eye positions not detected by an eye tracker are discarded from the analysis.

2.2 I-HMM

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. The I-HMM presented in this paper has two states - fixation and saccades and the structure of the model is similar to the model presented by Salvucci and Goldberg (Salvucci and Goldberg 2000). Each state is characterized by a velocity distribution and in which the states represent the velocity distributions for saccade and fixation points.

The first stage of the I-HMM is identical to I-VT, where each eye position sample is classified either as a fixation or a saccade depending on the velocity threshold. Second stage is defined by the Viterbi Sampler (Forney 1973), where each eye position can be re-classified, 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 given to I-HMM are not optimal and can be improved. The third stage of the I-HMM is defined by Baum-Welch re-estimation algorithm (Baum, Petrie et al. 1970). This algorithm re-estimates initial probabilistic parameters and attempts to minimize errors in the state assignments. Parameter re-estimation performed by Baum-Welch can be conducted multiple times. In the I-HMM defined in this paper, the number of such re-estimations is four.

2.3 I-KF

The Kalman filter is a recursive estimator that computes a future estimate of the dynamic system state from a series of incomplete and noisy measurements. A Kalman Filter minimizes the error between the estimation of the system's state and the actual system's state. Only the estimated state from the previous time step and the new measurements are needed to compute the new state estimate.

In our research, we employ a *Two State Kalman Filter* (TSKF) model that is described in detail in an earlier manuscript (Komogortsev and Khan 2009). The TSKF 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 TSKF generates predicted eye velocity signal. The values of the measured and predicted eye velocity allow employing Chi-square test to detect the onset and the offset of a saccade (Sauter, 1991).

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

where $\hat{\theta}_i^-$ is the predicted eye velocity computed by Kalman filter and $\dot{\theta}_i$ is the observed eye velocity computed with eye position signal from the eye tracker. δ is the standard deviation of the measured eye velocity during the sampling interval under consideration. Once a certain threshold of the χ^2 is achieved, a saccade is detected.

2.4 I-MST

Eye fixations are characterized by a set of points that are enclosed in a relatively small region. The I-MST is a dispersion-based identification algorithm that is argued to be a highly flexible and controllable eye movement detection tool (Salvucci and Goldberg 2000). The I-MST algorithm builds a minimum spanning tree taking a predefined number of eye position points using Prim's algorithm. The Minimum Spanning Tree (MST) is defined as a spanning tree with a distance minimum among all spanning trees in this set of nodes. The I-MST traverses the MST and separates the points into the fixations and the thresholds based on the predefined distance thresholds. The advantage of using an I-MST is that the algorithm can correctly identify fixation points even when a large part of the signal is missing due to noise. For long eye movement recordings, the I-MST requires a sampling window to build a sequence of MST trees allowing it to parse a long eye movement recording. The length of such window can be selected to be equivalent to the duration of the largest saccade which is expected to be recorded. In our experiments, the window size is selected to be 200ms.

2.5 I-DT

The Dispersion Threshold Identification (I-DT) algorithm takes into account the close spacial proximity of the eye position points in the eye movement trace (Salvucci and Goldberg 2000). The algorithm defines a temporal window which moves one point at a time, and the spacial dispersion created by the points within this window is compared against the threshold. If such dispersion is below the threshold, the points within the temporal window are classified as a part of 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 tied down 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)], which X and Y represent eye position sets within the temporal window.

3 Qualitative and Quantitative Scoring of the Eye Movement Classification Algorithms

To establish a common ground between eye movement classification algorithms, it is important to define a set of the qualitative and quantitative scores for the assessment of the performance of the classification algorithms. Assuming that a classification algorithm classifies eye position trace into fixation and saccades following performance metrics can be considered Average Number of Saccades (ANS), Average Number of Fixations (ANF), Average Fixation Duration (AFD) and Average Saccade Amplitude (ASA). The performance of the classification algorithms can be assessed by these metrics with or without the

knowledge of the stimuli. The values of these metrics have been previously employed in usability (Poole and Ball 2004), psychology (Ceballos, Komogortsev et al. 2009), and physical therapy (Garbutt, Han et al. 2003).

We propose three new metrics the Fixation Quantitative Score, the Fixation Qualitative Score, the Saccade Quantitative Score to evaluate saccade and fixation behavior and complement the metrics mentioned above.

3.1 Fixation Quantitative Score

The intuitive idea behind Fixation Quantitative Score (FQnS) is to compare the amount of the detected fixation behavior to the amount of presented fixation stimuli. The FQnS compliments the AFD and the ANF metrics, because it validates detected fixations in regard to the spacial and temporal properties of the stimuli signal.

To calculate the FQnS, the fixation stimuli position signal is sampled with the same frequency as the recorded eye position signal. Every resulting coordinate tuple (x_s,y_s) inside of the fixation stimuli is compared to the corresponding coordinate tuple (x_e,y_e) in the recorded eye position signal. If the corresponding eye position sample is marked as a fixation with coordinates close to stimuli fixation, then fixation detection counter is increased. The FQnS is calculated by normalizing detection success counter by total amount of the stimuli fixation points.

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

where $fixation\_detection\_counter$ represents the amount of eye position points identified as fixations when corresponding fixation stimuli was present. $stimuli\_fixation\_points$ represents the total amount of stimuli points presented as fixation and sampled at the eye tracker's sampling frequency.

It is important to mention that practically, the FQnS will not reach the 100% mark if the stimuli consists of both fixations and saccades. When a future fixation target appears in the periphery, the brain approximately requires 200ms to calculate and send the neuronal signal to the extraocular muscles to execute a saccade (Leigh and Zee 2006). Additionally, saccade duration approximates to $D_{sac\_dur} = (2.2A_{sac\_amp} + 21)$, where A_{sac_amp} is saccade's amplitude measured in degrees (Carpenter 1977). Due to this phenomena, the onset of the fixation will be always delayed by at least 200ms plus the duration of the saccade.

3.2 Fixation Qualitative Score

The intuitive idea behind the Fixation Quantitative Score (FQlS) is to compare the proximity of the detected fixation to the presented stimuli, therefore providing the information about positional accuracy of the detected fixation.

The FQlS calculation is similar to the FQnS, i.e., for every fixation related point (x_s,y_s) of the presented stimuli, the check is made for the point in the eye position trace (x_e,y_e); if such point is classified as a fixation, the Euclidean distance between 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 amount of points compared.

$$\mathrm{FQlS} = \frac{1}{N} \cdot \sum_{i=1}^{N} fixation\_distance_i \qquad 2$$

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

Ideally, the FQlS should equal 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 stimuli. In practice, the accuracy of modern eye trackers remains in the <0.5º range. In addition, subjects very frequently experience undershoots or overshoots when making saccades (Leigh and Zee 2006), therefore placing detected fixations slightly off-target. As a result, we hypothesize that practical values for the FQlS will be around 0.5º or larger.

3.3 Saccade Quantitative Score

The intuitive idea behind the Saccade Quantitative Score (SQnS) is to compare the amount of the detected saccades given the properties of the saccadic behavior of the presented stimuli. The SQnS adds to the ASA and the ANS metrics because it quantifies the correct saccade behavior even in cases when subjects experience large numbers of express saccades, overshoots or undershoots (Leigh and Zee 2006).

To calculate SQnS, two separate quantities are computed, one measures the amount of the saccade invoking behavior present in the stimuli, and the second one computes the total amplitude of the detected saccades. To calculate stimuli related metric, each jump in the location of the fixation target is considered to be a stimuli saccade, and the absolute distances difference between targets are added to the *total_stimuli_saccade_amplitude*. Similarly, the quantity called *total_detected_saccade_amplitude* represents the sum of the absolute values of the saccade amplitudes detected by a given classification algorithm.

$$\mathrm{SQnS} = 100 \cdot \frac{total\_detected\_saccade\_amplitude}{total\_stimuli\_saccade\_amplitude} \qquad 3$$

The SQnS of 100% indicates that the amount of the detected saccades equals the amount of the saccades invoked by the presented stimuli. The SQnS can be larger than 100%, which essentially means two things: abnormal saccadic behavior of the subject or classification algorithm that amplifies saccadic behavior, i.e., some of the fixations are classified as saccades. An example of the abnormal saccadic behavior can be a subject with 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 (Leigh and Zee 2006)). The amplification of the saccadic behavior by a classification algorithm can be caused by the erroneous selection of the threshold classification parameter. The SQnS can be smaller than 100% in cases of hypometric saccadic behavior (target undershoots) or damping behavior of the classification algorithm.

4 METHODOLOGY

4.1 Apparatus

The experiments were conducted with a Tobii x120 eye tracker, which is represented by a standalone unit connected to a 24-inch flat panel screen with resolution of 1980x1200. The eye tracker performs binocular tracking with the following characteristics:

accuracy 0.5°, spatial resolution 0.2°, drift 0.3° with eye position sampling frequency of 120Hz. The Tobii x120 model allows 300x220x300 mm freedom of head movement. Nevertheless, a chin rest was employed in our experiments to provide higher accuracy and stability.

4.2 Procedure

4.2.1 Accuracy test

The accuracy test was employed prior to the experiment providing us with average calibration error and invalid data percentage for each subject. The accuracy test is described in more detail in (Koh, Gowda et al. 2009).

4.2.2 Fixation & Saccade Invocation Task

The stimulus was presented as a 'jumping point' with a vertical coordinate fixed to the middle of the screen. The first point was presented in the middle of the screen, the subsequent points moved to the left and to the right of the center of the screen with a spacial amplitude of 20º, therefore providing average stimuli amplitude of approximately 19.3º. The jumping sequence consisted of 15 points, including the original point in the center, therefore providing 14 stimuli saccades. After each subsequent jump, the point remained stationary for 1.5s before the next jump. The size of the point was approximately 1º of the visual angle with the center marked as a black dot. The point was presented with white color with peripheral background colored in black.

4.3 Participants

The test data consisted of a heterogeneous subject pool, age 18-25, with normal or corrected-to-normal vision. A total of 77 participants volunteered for the evaluation test. None of the participants had prior experience with eye tracking. Advanced accuracy test procedures were used to control the data collection by employing two parameters, first with the average calibration error eye and second with the invalid data percentage. The data analyzer was instructed to discard recordings from subjects with a calibration error of >1.70º (mean 1º) and invalid data percentage of >20% (mean 3.23%). Only 22 subject records passed these criteria.

4.4 Results

Figure 2 presents the results, where each models' behavior is given for a range of the threshold values. The I-VT and the I-HMM models were tested for the velocity threshold range of 5º/s to 300º/s, the I-MST and the I-DT were tested for the distance/dispersion threshold range of 0.033º to 2º, and the I-KF was tested for the threshold range of 1 to 60. The x-axis of the graphs presented by Figure 2 depicts the range coefficient value that allows mapping of the specific threshold range of each model into a unifying range coefficient space. Threshold values for each algorithm can be represented by the input threshold function Th=RC*Inc+C. Where *Th* is the resulting value of the threshold, *RC* is a range coefficient changing from 0 to 59, *C* is the initial threshold value for every model, and Inc is the threshold increment value for each model. For the I-VT and the I-HMM, the *C* value is 5º/s; for the I-MST and the I-DT, this value is 0.033º; and for the I-KF, this value is 1. For the I-VT and the I-HMM Inc, the value is 5º/s; for I-MST and I-DT, this value is 0.033º; and for the I-KF, this value is 1. The input threshold function allows for comparison of performance of the classification models in the same range coefficient dimensions.

4.5 Fixation Qualitative Score (FQlS)

The performance of the 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 providing a slightly lower score, therefore indicating higher accuracy in terms of the coordinates of the detected fixation. Our previous study provided similar results in an online comparison of a real-time eye-gaze-guided system, showing 10% improvement in accuracy when the I-KF was compared to the I-VT (Koh, Gowda et al. 2009). The I-HMM was an outlier and provided the FQlS score that was essentially 33% higher than other algorithms, indicating a much lower accuracy in fixation coordinate detection.

4.6 Fixation Quantitative Score (FQnS)

The FQnS was monotonically growing for all classification algorithms. For all algorithms except the I-DT, there was an immediate jump in the score; and after a certain threshold value, there was a point of saturation where the increased threshold value did not produce an increased amount of the eye position points classified as fixations. All algorithms merged into the FQnS score of 74-77% which is agreeable with physiological latencies discussed in Section 3.1. The outlier from the rest of the group was the I-MST algorithm providing the saturated FQnS of 57% which was approximately 23% lower than the FQnS provided by other algorithms.

4.7 Saccade Quantitative Score (SQnS)

Each algorithm had a point of the maximum SQlS performance after which the score values monotonically decreased. This peak value was highest for the I-HMM algorithm with a value of approximately 110% and lowest for the I-KF with the value of 90%. The SQlS performance of the I-MST and the I-DT was slightly higher than the performance of the I-KF. For the high threshold values, the SQlS performance of the I-VT, I-DT and the I-HMM was quite similar. The I-KF provided the most damping behavior in terms of the amount of the detected saccades. The difference in performance between each individual algorithm did not exceed 22% after the Range Coefficient (RC) of 30 was reached. Prior to that RC value, the I-DT algorithm presented itself as an outlier with very low SQnS score.

We hypothesize that the initial score increase was due to the Merge_Function(). This function tends to make saccades longer when a large amount of points is classified as part of a saccade. With this type of behavior, the actual detected number of the saccades tend to be smaller with smaller amplitudes, resulting in the overall reduced saccadic behavior in terms of the SQlS score. After the SQlS peak is reached, the amount of saccadic behavior goes down because a lesser amount of eye position samples are classified as saccades.

4.8 Average Fixation Duration (AFD)

The trend for each classification algorithm was a low AFD following with very fast AFD growth up to a certain threshold. After this threshold the growth of the AFD was very slow or saturated. The difference between algorithms was substantial even when the AFD performance of each individual algorithm was saturated. For example, the I-KF provided maximized AFD while

the AFD value provided by the I-MST was, on average, twice as small. None of the classification algorithms were able to provide fixation duration at the level of the presented stimuli (1.5s).

The I-KF provided a very interesting spike in performance, providing the maximum AFD of 1.16s at the threshold of 25. This AFD was the closest to the stimuli fixation duration (1.5s) than for any other algorithm.

4.9 Average Number of Fixations (ANF)

The trend for each classification algorithm was a low ANF at the small threshold values and very fast ANF growth until a peak performance was reached. After the peak, there was a reverse trend where the ANF numbers went down. At the high threshold values, the performance of all algorithms stabilized. The I-MST method provided the highest amount of fixations (24) and the I-HMM the lowest (12), therefore showing a twofold difference between these two methods. The ANF of the remaining algorithms stayed between those two methods after the Range Coefficient reached the value of 16.

We hypothesize that the rapid decrease in the ANF values for the high threshold values was due to the Merge_Function() which creates smaller number of fixations with low duration in the signal where a significant amount of eye position points is classified as a part of a fixation.

It is important to mention that all methods were able to reach the number of fixations presented by the stimuli signal (15), but for some algorithms this value was reached twice.

4.10 Average Saccade Amplitude (ASA)

None of the methods were able to reach average saccade amplitude presented by the stimuli signal (19.3º). In general, each classification algorithm started with a higher ASA number following with the decrease in the ASA until a local minimum was reached. After such minimum, the ASA increased up to a certain threshold, reaching a local maximum that was algorithm specific. After the local maximum, the ASA numbers slowly decreased. The highest value of 17º was reached by the I-HMM model, with the lowest value of 8º provided by the I-VT and the I-DT models. During the saturated ASA behavior for high threshold values for each algorithm, the difference between reported ASA values was more than 5º, indicating the difference in number of the detected saccades of more than 50%.

4.11 Average Number of Saccades (ANS)

The trend for each classification algorithm was a low ANS for the low threshold values, following a peak performance at a certain threshold and a decreased/saturated performance at the high threshold values, with ANS values close to the stimuli signal behavior (stimulus saccade amplitude 15º). For the low threshold values represented by the Range Coefficient from 0 to 16, the I-DT and the I-VT were outliers with the I-DT detecting very few saccades and the I-VT detecting a very large saccade number. For the high threshold values, the I-DT provided the highest ANS of 17 while the I-HMM provided the lowest ANS of 10.

5 Discussion

Input parameter selection affects the performance of the classification algorithms quite significantly. Specifically, almost all performance metrics discussed in this paper produced very different values based on the algorithm and the threshold values selected. Depending on the algorithm and the value of the threshold parameter such difference reached 100% in certain cases. Nevertheless, the general performance trend in terms of the already established metrics and the new proposed metrics was similar.

5.1 Advantages provided by quantitative and qualitative scores

The Fixation Qualitative Score (FQlS) proved to be extremely useful in being able to distinguish the accuracy of the eye movement detection method given the threshold value or any other input parameters.

The Fixation Quantitative Score (FQnS) was able to provide an overall picture for the fixation detection behavior that was much less "noisier" than the data provided by the Average Fixation Duration (AFD) and the Average Number of Fixations (ANF) metrics. This can be observed for the I-VT, I-DT, I-HMM and the I-KF models that provide varying behavior in terms of the AFD and the ANF but essentially converge in terms of the FQnS. The important feature of the FQnS is that it ensures the temporal validity of the presented fixations by matching them with the spacial and temporal characteristics of the stimuli signal. The FQnS is able to pick out classification disadvantages of an algorithm, such as I-the MST algorithm where spurious fixations can be detected due to the overlapping data.

The Saccade Quantitative Score (SQnS) is able to identify specific values for the input parameters (thresholds) that allow detection of the same amount of saccadic behavior as presented by the stimuli. This was not entirely possible with the Average Number of Saccades (ANS) and the Average Saccade Amplitude (ASA) metrics, due to some subjects making multiple saccades to reach a target. This produced large ANS with small ASA and lead to an erroneous conclusion that the algorithm provides incorrect classification.

5.2 Best eye movement detection algorithm

It is difficult to select "best" eye movement classification algorithm or to set a "golden standard" in terms of the eye movement classification scores/metrics. The most accurate classification algorithm would be the algorithm that achieves the minimum value (0º) for the Fixation Qualitative Score, maximum value for the Fixation Quantitative Score (100%) and the Saccade Quantitative Score value of approximately 100% with values from the remaining eye movement metrics in sync with the stimuli behavior. The selection of the "best" eye movement detection algorithm will also depend on the actual application. For a real-time eye-gaze-based interaction where dwell-time is the primary mode of selection the I-KF can be considered as the best performer for the following reasons: high accuracy (lowest FQlS), FQnS was at an acceptable level of 70%, saccadic performance was dampened (signal jumps are smoothed) SQnS=68.5%, number of fixations and saccades was very close to the number present in the stimuli signal, detected fixation duration was closest to the value presented in the stimuli among all classification methods, and the detected saccade amplitude was second closest to the stimuli. For the studies related to sciences that investigate saccadic behavior, e.g, Physical Therapy, Psychiatry, the accurate detection of saccadic behavior is of paramount importance. Traditionally, the I-VT is a model of choice in this domain. From

the results presented in this paper, we can validate this choice by looking at the FQnS behavior which indicates the same amount of saccades in the classified signal as in the stimuli signal for the velocity threshold range of 30-70°/s. There is large number of saccades (ANS) detected by the I-VT in this threshold range, and those saccades have smaller amplitudes (ASA). We hypothesize that such behavior provides an opportunity to properly detect eye movement artifacts such as overshoots, undershoots, express saccades, corrective saccades and dynamic overshoots. Additionally, this window (30-70°/s) in the threshold range provides an opportunity to fine-tune the performance of the I-VT model. This can be done in terms of the fixation related metrics; e.g., by selecting a higher velocity threshold, it is possible to provide a more stable I-VT performance in terms of the detected fixations.

For practical selection of the input parameters (threshold values) for any classification algorithm, we suggest creating a controlled stimuli presentation and selection of appropriate input parameters based on the score values that we have discussed in this paper.

6 Conclusion

In this paper, we have discussed a set of scores that allows one to assess the implementation of any eye movement classification algorithm by providing the qualitative and the quantitative information about the classification performance. The performance of the five most usable classification algorithms was discussed in terms of the proposed scores. The results indicate that the classification performance differs significantly based on the algorithm and the selected threshold values. This result suggests that the description of the eye movement detection algorithms, and their parameters, in the research papers is of paramount importance. Specifically, we suggest that the performance of each classification algorithm should be reported in terms of qualitative and quantitative metrics discussed in this paper due to the fact that these metrics provide a more complete and accurate information about classification behavior.

The choice of the "best" algorithm in terms of eye movement classification proves to be challenging. We provide the argument that among the five classification algorithms we considered in this paper, Kalman filter shows the most benefits for implementation for the real-time eye-gaze-guided systems. The Velocity Threshold algorithm proves to be the better choice for the systems measuring saccadic performance.

A. Varri, B. Kemp, et al. (1995). "Multi-centre comparison of five eye movement detection algorithms." Journal of Sleep Research **4**(2): 119-130.

Baum, L. E., T. Petrie, et al. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." The Annals of Mathematical Statistics **41**(1): 164-171.

Carpenter, R. H. S. (1977). Movements of the Eyes. London, Pion.

Ceballos, N., O. Komogortsev, et al. (2009). "Ocular Imaging of Attentional Bias Among College Students: Automatic and Controlled Processing of Alcohol- Related Scenes." Journal of Studies on Alcohol and Drugs, September: 1-8.

Duchowski, A. (2007). Eye Tracking Methodology: Theory and Practice, Springer.

Duchowski, A., T. and A. Çöltekin (2007). "Foveated gaze-contingent displays for peripheral LOD management, 3D visualization, and stereo imaging." ACM Transactions on Multimedia Computing, Communications, and Applications **3**(4).

Ehmke, C. and S. Wilson (2007). Identifying web usability problems from eye-tracking data. 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, British Computer Society.

Field, M., K. Mogg, et al. (2004). "Eye movements to smoking-related cues: effects of nicotine deprivation." Psychopharmacology **173**(1): 116-123.

Forney, G. D., Jr. (1973). "The viterbi algorithm." Proceedings of the IEEE **61**(3): 268-278.

Garbutt, S., Y. Han, et al. (2003). "Vertical Optokinetic Nystagmus and Saccades in Normal Human Subjects." Invest. Ophthalmol. Vis. Sci. **44**(9): 3833-3841.

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

Koh, D. H., S. A. M. Gowda, et al. (2009). Input evaluation of an eye-gaze-guided interface: kalman filter vs. velocity threshold eye movement identification. Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems. Pittsburgh, PA, USA, ACM**:** 197-202.

Komogortsev, O., C. Mueller, et al. (2009). An Effort Based Model of Software Usability. International Conference on Software Engineering Theory and Practice (ISETP).

Komogortsev, O. V. and J. Khan (2007). Kalman Filtering in the Design of Eye-Gaze-Guided Computer Interfaces. 12th International Conference on Human-Computer Interaction (HCI 2007), Beijing, China.

Komogortsev, O. V. and J. Khan (2009). "Eye Movement Prediction by Oculomotor Plant Kalman Filter with Brainstem Control." Journal of Control Theory and Applications **7**(1): 14-22.

Leigh, R. J. and D. S. Zee (2006). The Neurology of Eye Movements, Oxford University Press.

McConkie, G., W. (1980). Evaluating and reporting data quality in eye movement research, University of Illinois.

Munn, S. M., L. Stefano, et al. (2008). Fixation-identification in dynamic scenes: comparing an automated algorithm to manual coding. Proceedings of the 5th symposium on Applied perception in graphics and visualization. Los Angeles, California, ACM.

Parkhurst, D., J. and E. Niebur (2002). "Variable resolution displays: A theoretical, practical, and behavioral evaluation." Human Factors **44**(4): 611-629.

Poole, A. and L. J. Ball (2004). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. Encyclopedia of Human Computer Interaction, Idea Group.

Rayner, K. (1998). "Eye Movements in Reading and Information Processing: 20 Years of Research." Psychological Bulletin **124**(3): 372-422.

Salvucci, D. D. and J. H. Goldberg (2000). Identifying fixations and saccades in eye tracking protocols. Eye Tracking

Research and Applications Symposium, New York, ACM Press.

Sauter, D., M. B. J., et al. (1991). "Analysis of eye tracking movements using innovations generated by a Kalman filter." Med. Biol. Eng. Comput.: 63–69.

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

Suh, M., S. Basu, et al. (2006). "Increased oculomotor deficits during target blanking as an indicator of mild traumatic brain injury." Neuroscience Letters **410**(3): 203-207.

Tigges, P. K., N. Kathmann, et al. (1995). Semiautomated extraction of decision relevant features from a raw data based artificial neural network demonstrated by the problem of saccade detection in EOG recordings of smooth pursuit eye movements. Neural Networks for Signal Processing [1995] V. Proceedings of the 1995 IEEE Workshop.

Widdel, H. (1984). Operational problems in analysing eye movements. Theoretical and Applied Aspects of Eye Movement Research. A. Gale and F. Johnson. New York, Elsevier**:** 21-29.

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

Algorithm: *I-VT*
Input: array of eye position points, velocity threshold
Output: array of fixations and saccades

Calculate point-to-point velocities for each point in the eye position array

Mark points below velocity threshold as fixations and the points above the threshold as saccades

Merge Function(array of pre classified fixation and saccades)

> *Merge every group of consecutive saccade points into a saccade with an onset, offset, amplitude, and duration*
>
> *Remove micro and corrupted saccades*
>
> *Merge every group of fixation points into a fixation with center coordinates, onset and duration*
>
> *Remove fixations that are below minimum fixation duration threshold*

Return saccades and fixations

Algorithm: *I-HMM*
Input: array of eye position points, velocity threshold, initial, transitional, observation probabilities
Output: array of fixations and saccades

Calculate point-to-point velocities for each point in the eye position array

Mark points below velocity threshold as fixations and the points above the threshold as saccades

Define Viterbi sampler of the HMM and re-estimate fixation, saccade assignment for every eye position point

Use Baum-welch algorithm to re-estimate initial, transition, and observation probabilities for the defined sampler

Merge Function(array of pre classified fixation and saccades)

Return saccades and fixations

Algorithm: *I-KF*
Input: array of eye position points, chi-square threshold, initialization parameters for Kalman filter
Output: array of fixations and saccades

Calculate point-to-point velocities for each point in the eye position array

Use Kalman filter to generate predicted point-to-point velocities for each point in the eye position array

Employ Chi-square test between the actual and the predicted point-to-point velocities

Mark points with Chi-square test value below the threshold as fixations and above the threshold as saccades

Merge Function(array of pre classified fixation and saccades)

Return saccades and fixations

Algorithm: *I-MST*
Input: array of eye position points, distance threshold, temporal window size
Output: array of fixations and saccades

Move temporal window until the end of the eye position trace is reached. / previous and next window positions do not overlap*/*

For each point within the window

Compute the Euclidean distance between each eye position point

Construct the minimum spanning tree using Prim's algorithm

Test the weight (length) of each edge in the MST

For each edge with its weight below the distance threshold, mark its end points as fixation points.

For each edge with its weight above the distance threshold, mark its end points as saccade points.

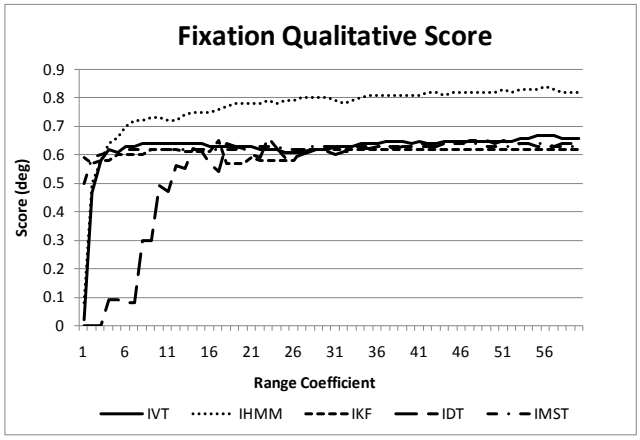
Merge Function(array of pre classified fixation and saccades)

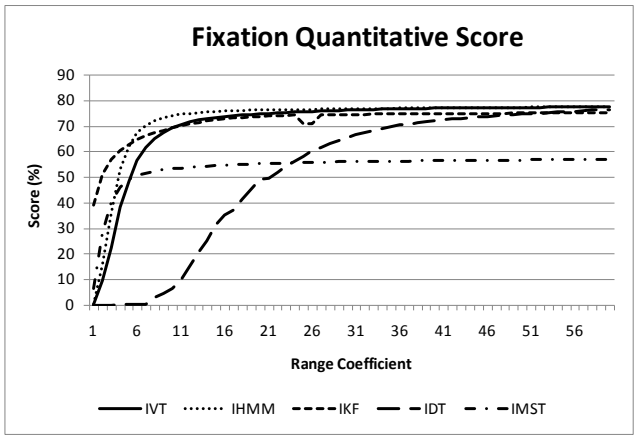
Return saccades and fixations

Algorithm: *I-DT*
Input: array of eye position points, dispersion threshold, temporal window size
Output: array of fixations and saccades

Initialize temporal window over first points in the remaining eye movement trace
Calculate dispersion of points in window
If (dispersion < dispersion threshold)
  While dispersion < dispersion threshold
    Add one more point to window
    Calculate dispersion of points in window
  End while
  Mark the points inside of the window as fixations
  Clear window
Else
  Remove first point from window
  Mark first point as a saccade
End if
Merge Function(array of pre classified fixation and saccades)
Return saccades and fixations

Figure 1. Pseudocode for each classification algorithm





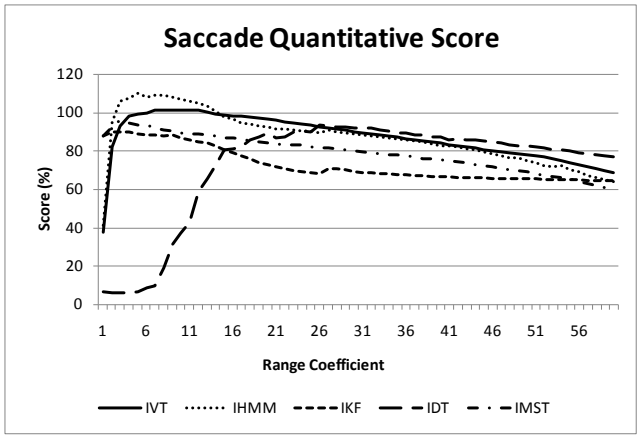


Figure 2. Metric and score values for each classification algorithm