To be published in In Proceedings of ACM Eye Tracking Research & Applications Symposium, Austin, TX, 2010

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

Oleg V. Komogortsev Sampath Jayarathna Do Hyong Koh Sandeep Munikrishne Gowda
Department of Computer Science Texas State University-San Marcos {ok11,sampath,dk1132,sm1499@txstate.edu}

## Abstract

This paper presents a set of qualitative and quantitative scores designed to assess performance of any eye movement classification algorithm. 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, vestibulo-ocular reflex, and vergence [Leigh and Zee 2006]. Fixations and saccades are the types of most researched eye movements that are employed in human computer interaction, psychological studies and reading, medical studies, and usability studies [Ceballos et al. 2009; Duchowski et al. 2009; Garbutt et al. 2003]

The development of the eye movement classification algorithms has a long history [McConkie 1980; Munn et al. 2008; Salvucci and Goldberg 2000]. 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 Qualitative and Quantitative Scoring

The description and pseudocodes for the Velocity Threshold Identification (I-VT), Dispersion Threshold Identification (I-DT), Minimum Spanning Tree Identification (MST), Hidden Markov Model Identification (I-HMM), and Kalman Filter employed in this paper can be found in [Komogortsev et al. 2009].

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, the 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 [Duchowski 2007], psychology [Ceballos et al. 2009], and physical therapy [Garbutt 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.

### 2.1 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 [Leigh and Zee 2006]. Due to this phenomena, the onset of the fixation will be always delayed by at least 200ms plus the duration of the saccade.

## 2.2 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.

## 2.3 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.

# 3 METHODOLOGY

**Apparatus:** The experiments were conducted with a Tobii x120 eye tracker (sampling rate 120Hz), which is represented by a standalone unit connected to a 24-inch flat panel screen with resolution of 1980x1200. Chin rest was employed to provide additional head stability. **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. **Participants & Data Quality:** The test data consisted of a heterogeneous subject pool, age 18-25, with normal or corrected-to-normal vision. 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 [Koh et al. 2009]. The data analyzer was instructed to discard recordings from subjects with a calibration error of >1.70º and invalid data percentage of >20%. Only 22 out of 77 subjects' records passed these criteria. The remaining records had a mean accuracy of 1º and a mean invalid data percentage of 3.23%.

# 4 Results & Discussion

Figure 1 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 Chi-square test threshold range of 1 to 60. The

range values came as suggestions from the research literature [Duchowski 2007; Koh et al. 2009; Leigh and Zee 2006; Salvucci and Goldberg 2000].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.

**Performance Metrics:** ANS, ANF, AFD, and ASA behavior varied greatly depending on the values of the threshold values. Such difference in classification performance between algorithms frequently reached 100% mark or higher. Based on the results it is possible to distinguish trends in classification performance depending on the threshold values, but such trendlines continue to be extremely jittery.

**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 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.

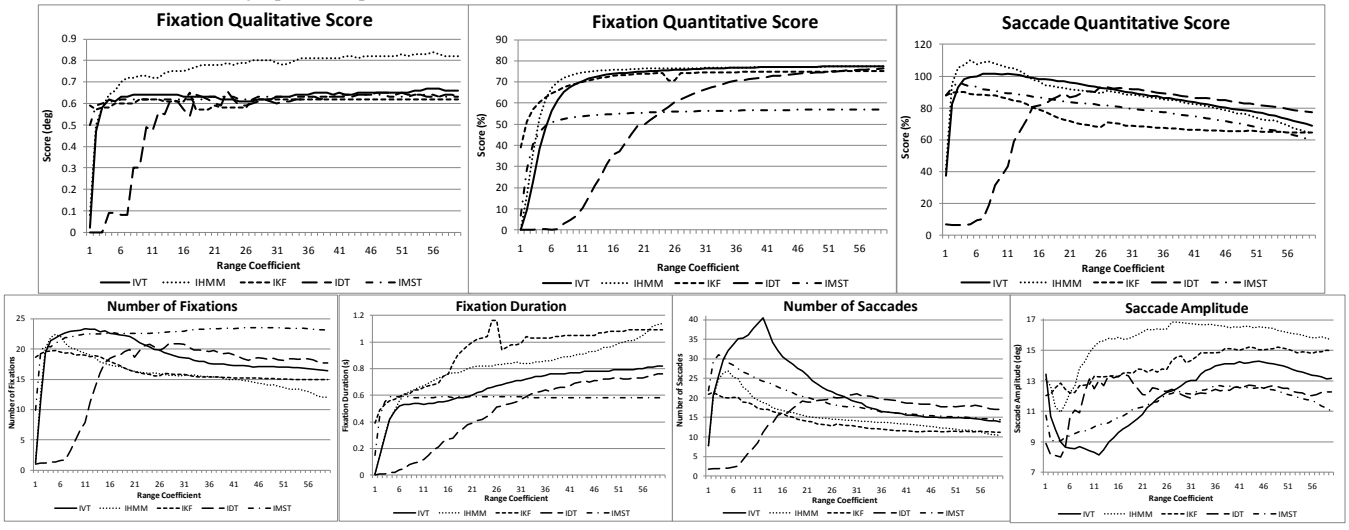
**Fixation Quantitative Score (FQnS):** The FQnS was monotonically growing for all classification algorithms. For all algorithms except the I-DT, there was an immediate jump in the score; and after a certain threshold value, there was a point of saturation where the increased threshold value did not 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 2.1. 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.

**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.

**Advantages of Quantitative/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.

**Limitations:** 1) Practical use of the FQlS, FQnS, and SQnS metrics will require a presentation of a controlled stimulus prior to the experiment for the selection of the thresholds values. While we agree that this can be considered as an extra step in the calibration process, the outcome of such calibration will allow to have a much better, performance-based values for the input thresholds for the actual experiment. 2) This paper varies just single input parameter for each classification algorithm. The change in other input parameters or/and eye-tracker's sampling rate, noise in the eye tracking signal or/and random amplitude of the ramp stimulus would definitely affect the performance of the scores. Therefore, the amount of variability in our evaluation setup was minimized to show that classification performance is greatly affected just by a single parameter.

**Figure 1. Quantitative/Qualitative Scores and Performance Metrics.**

**Best eye movement classification 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). This behavior provides an opportunity to properly detect eye movement artifacts such as overshoots, undershoots, express saccades, corrective saccades and dynamic overshoots. Additionally, the velocity threshold 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 fine tuning the fixation related metrics by selecting a higher velocity threshold.

## 5 Conclusion

In this paper, we have discussed a set of scores that allows one to assess an implementation of any eye movement classification algorithm by providing the qualitative and the quantitative information about the classification performance. Such information allows to provide a point of reference offering a capability to validate the results of an experiment involving an eye tracker. 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.

## 6 References

CEBALLOS, N., KOMOGORTSEV, O., AND TURNER, G. M., 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*, 2nd edition.(Springer).

DUCHOWSKI, A. T., BATE, D., STRINGFELLOW, P., THAKUR, K., MELLOY, B. J., AND GRAMOPADHYE, A. K., 2009. On spatiochromatic visual sensitivity and peripheral color LOD management, *ACM Trans. Appl. Percept.* 6, 1-18.

GARBUTT, S., HAN, Y., KUMAR, A. N., HARWOOD, M., HARRIS, C. M., AND LEIGH, R. J., 2003. Vertical Optokinetic Nystagmus and Saccades in Normal Human Subjects, *Invest. Ophthalmol. Vis. Sci.* 44, 3833-3841.

KOH, D. H., GOWDA, S. A. M., AND KOMOGORTSEV, O. V., 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*, 197-202.

KOMOGORTSEV, O. V., JAYARATHNA, U. K. S., KOH, D. H., AND GOWDA, S. M., 2009. *Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms*(Texas State University - San Marcos, San Marcos, ).

LEIGH, R. J., AND ZEE, D. S., 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., Stefano, L., AND Pelz, J. B., 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*.

Salvucci, D. D., AND Goldberg, J. H., 2000. Identifying fixations and saccades in eye tracking protocols, *Eye Tracking Research and Applications Symposium*, 71-78.