FELiX: Fixation-based Eye Fatigue Load Index A Multi-factor Measure for Gaze-based Interactions

Mohsen Parisay
Dept. of Computer Science
Concordia University
Montreal, Canada
m_parisa@encs.concordia.ca

Charalambos Poullis
Dept. of Computer Science
Concordia University
Montreal, Canada
charalambos@poullis.org

Marta Kersten-Oertel
PERFORM Center
Dept. of Computer Science
Concordia University
Montreal, Canada
marta@ap-lab.ca

***Abstract*—Eye fatigue is a common challenge in eye tracking applications caused by physical and/or mental triggers. Its impact should be analyzed in eye tracking applications, especially for the dwell-time method. As emerging interaction techniques become more sophisticated, their impacts should be analyzed based on various aspects. We propose a novel compound measure for gaze-based interaction techniques that integrates subjective NASA TLX scores with objective measurements of eye movement fixation points. The measure includes two variations depending on the importance of (a) performance, and (b) accuracy, for measuring potential eye fatigue for eye tracking interactions. These variations enable researchers to compare eye tracking techniques on different criteria. We evaluated our measure in two user studies with 33 participants and report on the results of comparing dwell-time and gaze-based selection using voice recognition techniques.**

***Index Terms*—Eye tracking, eye fatigue, dwell-time, voice recognition, cognitive workload, NASA TLX, gaze-based interaction**

I. Introduction

A. Cognitive Workload

Cognitive workload is defined as the amount of mental effort of a person performing a task or in the process of problem-solving. It is related to a person's working memory which has a limited capacity [1], [2]. It is important to measure the amount of cognitive workload related to performing a task given a specific interface in order to compare the usability of different systems. The NASA Task Load Index (TLX) questionnaire is a well-known multidimensional method used to measure subjects' perceived workload in user studies [3], [4]. The TLX questionnaire, which has been shown to be a valid tool to measure workload [5], comprises of six scales: (1) physical demand, (2) mental demand, (3) temporal demand, (4) effort, (5) performance and (6) frustration, each on a 100-point range with 5-point steps [6]. Each scale can be weighted based on its importance and used to calculate the average value - known as the *overall workload*. The overall workload serves as a measure of the efficacy of the interaction technique and can be used for comparing different methods based on their workload. However, the results of the NASA TLX are subjective and suffer from several limitations. One such limitation is that subjects often confound task performance with the perceived mental effort. Furthermore, as the results are obtained after a task is completed so as not to interrupt the task, the NASA TLX is not ideally suited for real-time scenarios [7]. For these reasons, more robust and accurate methods should be applied for measuring cognitive load, such as the use of physiological data [8]. Researchers are thus beginning to investigate the use of physiological signals, for example by measuring brain activity. Techniques for measuring brain activity include: (1) Electroencephalography (EEG) which detects brain waves [9], (2) Magnetoencephalography (MEG) that records magnetic fields of electrical activities in brain [10], and (3) Near-infrared spectroscopy (NIRS) which is a spectroscopic method that uses wavelengths in the near-infrared range to measure blood flow changes in the frontal cortex [11]. These methods although accurate in detecting brain activity require specialized and sometimes cumbersome equipment. In addition, these techniques are intrusive for users and therefore are restricted to controlled environments such as laboratories [7].

B. Eye Fatigue

According to Vasiljevas *et al.*, fatigue is the increase of tiredness of a subject under load [12] and can be grouped into physical, e.g. lack of sleep, and mental related causes such as stress [13]. According to Marcora *et al.*, mental fatigue is the result of high cognitive activity [14]. Visual fatigue defined as "eyestrain or asthenopia, which can be caused by both two-dimensional and stereoscopic moving images" [15] and which can cause motion sickness [16], occurs when focusing on near objects. The visual function of the eyes may cause visual fatigue, especially in long-time periods. Other symptoms of visual fatigue include: tiredness, headaches, and irritation of the eyes [17]. In this paper, we propose an integrated measure to detect task load and visual fatigue during gaze-based interactions. Our focus is on visual fatigue as it is a common issue among computer users due to the prolonged periods of time they spend working in front of a monitor [12]. We believe *that a comprehensive measure that combines the quantitative aspects of eye tracking fixation points with the qualitative aspects of the NASA TLX scores could provide an effective means to distinguish task load and fatigue in different gaze-based interactions*. The developed measure is an alternative to using sensory devices in situations where the application of biological sensors are either not possible, or cumbersome to participants for user studies.

The contribution of this paper is twofold. Firstly, we introduce FELiX: Fixation-based Eye Fatigue Load Index, an

978-1-7281-7392-4/20/$31.00 ©2020 IEEE

integrated measure of task workload and visual fatigue. The term *eye fatigue load* is defined as a combined measure of task workload and visual fatigue. The FELiX measure combines the accuracy of the objective eye tracking data (quantitative inputs) with the subjectivity of user's experience as calculated by the NASA TLX scores (qualitative inputs) during gaze-based interactions. Secondly, we investigate the ability of FELiX to measure eye fatigue load by conducting two user studies comparing two gaze-based interaction techniques: dwell time and voice recognition. The results of our studies show that FELiX is able to distinguish between different gaze-based interaction methods.

II. Related Work

Researchers proposed various measures to measure eye fatigue based on either eye movement analysis or biological sensor inputs. Zheng *et al.* investigated the correlation between eye blinks and mental workload among surgeons. They found that shorter blink duration and frequency indicate an increase of the mental workload [18]. Additionally, Borghini *et al.* studied brain activity and heart rate of car drivers and found the same results regarding the eye blink rates with mental workload [19]. Lanthier *et al.* studied the correlation between fixations and eye fatigue during visual search tasks and found that fixation duration increases with fatigue [20]. Abdulin *et al.* showed that the distance drift of fixation points in response to a stimuli can reveal physical eye fatigue [21] and calculated this using the fixation qualitative score (FQlS) [22]. Vasiljevas *et al.* examined an analytical model of muscle fatigue proposed to measure athletes fatigue [23] and adopted it to assess eye fatigue in gaze-based tasks [12]. In studying the impact of learning on fatigue, they found that the required break time for gaze-based interactions can be measured. Researchers have also applied self-evaluation questionnaires to evaluate eye fatigue in user studies for gaze-based applications [24]. There are saccades-based approaches to measure eye fatigue [25]–[27]. However, according to Abdulin *et al.*, analysis of saccades raw data requires expensive eye trackers, and these approaches are not applicable on budget-friendly devices [21]. Building on the previous works, we propose a fixation-based approach which can be applied on most eye trackers. Although, previous measures can be used to measure eye fatigue with high probability, they rely solely on eye movements or sensor inputs. To the best of our knowledge, there are currently no measures that integrate NASA TLX scores with the measurements of eye movements using eye tracking to assess eye fatigue. To take advantage of both physiological data and user perceptions, we integrate eye movements (fixation points) as an objective measure, and NASA TLX scores as a subjective measure in FELiX. By combining workload and eye fatigue in one measure, FELiX is ideally suited to compare different interaction techniques in gaze-based interaction user studies.

III. Eye Fatigue Load Index (FELiX)

We propose two variations of FELiX, both of which integrate the beneficial features (simplicity and direct ratings by users) of the NASA TLX with gaze fixations to measure eye fatigue load. Out of the six TLX questionnaire scales, we only employ scores for the following three scales: *physical demand* (PD), *mental demand* (MD), and *performance* (P). This choice is based on the fact that the *physical* and *mental* demands best describe the concept of workload to users, whereas *performance* is best interpreted by the users as the overall performance of the method. In contrast, the other three scales (e.g. temporal demand, frustration, effort) focus on usability and user satisfaction. The first variation of the proposed measure $FELiX_{per}$ incorporates fixations recorded (*x*, *y*, *timestamp*) during a gaze-based test as well as the error rates of target selections and can be used in experiments where performance is of high importance. On the other hand, if accuracy is of higher importance, the second variation $FELiX_{acc}$ can be used, which incorporates the Euclidean distance to the target as well as the number of fixations. The proposed measures, *performance-based* and *accuracy-based*, measure the eye fatigue load for any gaze-based interactions relying on eye movement measurements.

A. Cognitive and Eye-Tracking Coefficients

FELiX involves two coefficients, namely *cognitive* and *eye-tracking* coefficients. The *cognitive coefficient* is a qualitative factor which is calculated based on the users' rating scores of the NASA TLX questionnaire for the scales PD, MD, P (rated on a scale of 1-100). The *eye tracking coefficient* is a quantitative factor which is calculated from the eye-tracking data recorded during the test session. Since the recorded values used in calculating the eye-tracking coefficient can vary depending on the test conditions, we use the logarithmic function to scale down to a lower range the potentially large index values. Furthermore, to avoid cases where one coefficient diminishes the effect of the other (e.g. eye-tracking coefficient or cognitive coefficient is close to zero), we offset the coefficients by 1 and 9 respectively, such that the lowest value is ≥ 1 as explained below. We applied similar parameters introduced by previous research based on saccades [25]–[27], and fixation analysis such as *average fixation duration time* (AFD), and *average number of fixations* (ANF), as proposed by Komogortsev *et al.* [22].

B. Performance-based FELiX ($FELiX_{per}$)

Equation 2 shows the formula for the first variation of the measure, $FELiX_{per}$. This measure can be used for calculating the eye fatigue load index for interaction techniques and is dependent on the following parameters:

- average fixation duration time (AFD),
- error rate (ER) which is the total number of error selections divided by the total number of targets,
- average number of fixations (ANF), and
- NASA TLX questionnaire (3 scores: PD, MD, P)

The conditions and range of each of the parameters are given by,

1) $\forall a \in \{PD, MD, P\} : a \in \mathbb{Z} \wedge 1 \leq a \leq 100$
   TLX scores are integers in range of 1 to 100.
2) $\forall b \in ER : b \in \mathbb{R} \wedge 0 \leq b \leq 1$
   The error rate is a real number from 0 to 1.

3) $\forall c \in \{ER \times P\} : c \in \mathbb{R} \wedge 0 \leq c \leq 100$
The product of error rate and performance score is a real number from 0 to 100.
4) $\forall d \in CC_{per}, d \in \mathbb{R} \wedge 1 \leq d \leq 200$
The cognitive coefficient CC_{per} (equation 1) is the average of TLX scores (PD, MD) added to the product of error rate and performance score (P) which results in a real number from 1 to 200. This coefficient reflects the increase of task workload by multiplying the error rate factor. In the case of an error-free condition, the performance factor is removed to lower the cognitive coefficient.

$$CC_{per} = (\frac{PD + MD}{2}) + (ER \times P) \quad (1)$$

5) $\forall e \in \left(\frac{ANF}{AFD}\right), e \in \mathbb{R}_{>0}$
The eye tracking coefficient is comprised of the average number of fixations (ANF) divided by average fixation duration time (AFD) which results in a positive real number greater than 0. This measure reflects the duration of fixation points on average.
6) $\forall f \in FELiX_{per}, f \in \mathbb{R} \wedge f \geq 1$
$FELiX_{per}$ (equation 2) is the product of (a) logarithm of cognitive coefficient CC_{per} with the fixed constant value 9 in base 10, and (b) the eye tracking coefficient $\frac{ANF}{AFD}$ with the fixed constant value 1 which results in a real number greater or equal than 1.

$$FELiX_{per} = \log_{10}(9 + \underbrace{CC_{per}}_{\text{cog. coeff.}}) \times (1 + \underbrace{\frac{ANF}{AFD}}_{\text{eye-track. coeff.}}) \quad (2)$$

C. Accuracy-based FELiX ($FELiX_{acc}$)

Equation 4 shows the formula for the second variation of the measure, $FELiX_{acc}$. The measure can be used to calculate the eye fatigue load index for interaction techniques where accuracy is of utmost importance i.e. distance to target, such as in target selection tasks. $FELiX_{acc}$ is dependent on the parameters:

- average number of fixations (ANF),
- average Euclidean distance to the target (ADT), and
- NASA TLX questionnaire (2 scores: PD, MD) as described above.

The distance (ADT) is measured as the difference between the 2D coordinates of the center of a target and the coordinates of the corresponding fixation point. The conditions and ranges of each of the parameters are defined as,

1) $\forall a \in \{PD, MD\} : a \in \mathbb{Z} \wedge 1 \leq a \leq 100$
TLX scores are integers in range of 1 to 100.
2) $\forall b \in \left(\frac{ANF}{ADT}\right), b \in \mathbb{R}_{>0}$
The eye tracking coefficient is comprised of the average number of fixations (ANF) divided by average Euclidean distance to the target (ADT) which results in a positive real number greater than 0. This measure reflects the distance of fixation points to the target on average.
3) $\forall c \in CC_{acc}, c \in \mathbb{R} \wedge 1 \leq c \leq 100$
The cognitive coefficient CC_{acc} (equation 3) is the average of TLX scores PD and MD which results in a positive real number between 1 and 100.

$$CC_{acc} = \frac{PD + MD}{2} \quad (3)$$

4) $\forall d \in FELiX_{acc}, d \in \mathbb{R} \wedge d \geq 1$
$FELiX_{acc}$ (equation 4) is the product of (a) logarithm of cognitive coefficient CC_{acc} with the fixed constant value 9 in base 10, and (b) the eye tracking coefficient $\frac{ANF}{ADT}$ with the fixed constant value 1 which results in a real number greater or equal than 1.

$$FELiX_{acc} = \log_{10}(9 + \underbrace{CC_{acc}}_{\text{cog. coeff.}}) \times (1 + \underbrace{\frac{ANF}{ADT}}_{\text{eye-track. coeff.}}) \quad (4)$$

D. Discussion: Rational of FELiX

We employed quantitative parameters typically recorded in eye tracking applications in our measure since they reflect technical workflow of an interaction technique. These technical parameters are bound to test applications and equipment. Additionally, we applied workload parameters obtained from the NASA TLX scores to include direct ratings of participants who were involved in the practical aspects of an interaction technique. The proposed measure should result in a single value based on both technical and empirical parameters regarding the available measures. The purpose of multiplication of both coefficients (quantitative and qualitative) is to control the influence of both coefficients. In fact, the proposed measure should be balanced in the way that no aspects of an interaction technique (technical or empirical) can undermine the impact of the other.

IV. Methodology

To evaluate the effectiveness of the proposed measures we calculated the FELiX measure based on two gaze-based interaction studies with 33 participants (13 female, from 22 to 35 years old, $SD = 2.96$). All subjects partook in both experiments. The equipment is illustrated in Figure 1a.

A. Interaction Methods

1) Dwell-time: The dwell-time method integrates both pointing and selection phases using the eye tracker only. The range of dwell-time has been between 300-1100 milliseconds for target selection in the literature [28]. We defined the target activation threshold to 500 milliseconds, since it showed the best performance in [29] and participants prefer dwell-times of around 500 ms [28]. In other words, the target was considered as selected when a subject focused on it for 0.5 seconds; if the subject moved their gaze away from the target prior to the 0.5 seconds the selection process would restart.

2) Eye Tracking with Voice Recognition: For voice recognition, eye tracking was used for pointing and voice for selection. The selection phase for the voice recognition technique is triggered by a voice command which in our case was the word 'select' that was interpreted as a mouse click. The voice command is captured by a headset microphone (Logitech H370). An artificial ambient noise was introduced in the background through stereo desktop speakers at a volume of 50 dB

to simulate a typical work environment. The method was developed using the built-in Windows 10 speech recognition capabilities available in the .NET framework. We implemented a C# application to respond to the activation keyword 'select' to trigger a mouse click.

B. Hypotheses

Based on the previous literature, which has demonstrated dwell-time to be one of the most effective gaze-based interaction techniques [30], but one which can suffer from issues related to Midas touch [31], we hypothesized that:

1) The accuracy-based FELiX ($FELiX_{acc}$) will be lower for dwell-time than voice recognition because dwell-time should have lower fixation distances to target ($\frac{ANF}{ADT}$), as well as, lower physical demand (PD) and mental demand (MD).
2) The performance-based FELiX ($FELiX_{per}$) will be higher for dwell-time than voice recognition because dwell-time tends to result in more errors due to Midas touch and should have higher duration of fixation points ($\frac{ANF}{AFD}$).
3) The analysis of both FELiX variations will allow us to distinguish dwell-time and a multi-modal interaction technique.

C. Apparatus

In our user study, the mouse pointer position is captured using the Tobii 4C eye tracker[1]. All test applications were developed and the user studies were run on a commodity computer system: 64-bit Windows 10 PC with Intel i7 2.67GHz CPU, 12 GB RAM, 1 TB hard disk and NVIDIA GeForce GTX 770 graphics card. Figure 1a shows the required equipment of both interaction techniques.

1) Eye Tracking: Pointing Phase: The Tobii SDK (TobiiEyeXSdk-Cpp-1.8.498) supports different events related to eye tracking activities such as the location of the current eye gaze, positions of both eyes, fixation points, and user presence in front of the eye tracker. We employed the eye gaze library (API) to obtain users' gaze locations. These locations show the current gaze position on the screen in pixel coordinates. The SDK supports eye movements in a 3D coordinate system (horizontal, vertical, depth). However, we applied a 2D coordinate system (*x*,*y*) combined with a unique timestamp corresponding to the recorded location such that the mouse cursor was synchronized with the gaze positions to control the mouse pointer on the screen. Eye-tracking for both user studies was developed in C++ and integrated as a new plug-in into the Tobii SDK. The samples were recorded in distance of 60 cm (23.6 in) to the eye tracker with the sampling rate of 90 Hz.

2) Voice Processing: Selection Phase: To simulate a click on the item to be selected a headset microphone listens to the user while suppressing the background ambient sounds/noise in real-time. The Windows 10 Speech Recognition engine (available in the .NET framework) was selected to parse the received commands and a C# program was developed to trigger a left mouse click.

[1]https://tobiigaming.com/product/tobii-eye-tracker-4c/

D. Experimental Design

Prior to running the studies, subjects were informed about the purpose of the study, trained on each of the methods to be tested, and participated in a pre-test questionnaire inquiring on their background in the fields of eye tracking, voice recognition technologies and their preferred kind of interaction. After the pre-test questionnaire the Tobii calibration software was used to calibrate the system for each participant before starting the study. During the study each user partook in two experiments with different stimuli: (1) matrix-based and (2) dart-based. Overall, the studies took 8 minutes on average for each participants, 6 minutes for the matrix-based, and 2 minutes for the dart-based test.

E. User Study 1: Matrix-based Test

In the first experiment, a matrix of buttons (targets), were randomly distributed across the screen. The task of the subjects was to point and click on buttons shown on the screen in increasing numerical order for various levels of difficulty from 1 (easy) to 5 (hard), described in detail below. The level of difficultly was presented in ascending order. Further, the transition from lower levels to higher levels was done automatically, thus the whole test session for each participant was continuous.

1) Stimulus: The stimulus consisted of 77 buttons (11 columns $\times$ 7 rows) in size of 110 $\times$ 80 pixels, some labeled with numbers and others not, which covered the entire screen at a resolution of 1920 $\times$ 1080 pixels on a Dell P2411Hb monitor. Two marginal columns (far left, far right) and two rows (top, bottom) were removed from the active selection due to the high difficulty to be selected by users during the pilot-test. Buttons that were not labeled are considered as *barriers* or *distractions*. To provide feedback to the subject, labeled buttons change color after the user has successfully pointed and selected on the correct button. Wrongly selected barriers (buttons with no label) are highlighted in red. The level of difficulty of the stimulus was also increased across subject trials. This was done by increasing the number of targets that had to be selected by the subject. Five levels of difficulty were used for each interaction method: level 1 (4 targets), level 2 (6 targets), level 3 (8 targets), level 4 (10 targets) and level 5 (12 targets). Targets were randomly distributed over the entire screen for each level. Figure 1b shows the matrix-based test during difficulty level 5.

2) Measures: The following variables were recorded: *fixation duration time*, *number of fixations*, *error rates*, and *subjective ratings* (based on the NASA TLX scores). An internal logging module recorded subjects' actions, fixation duration times, wrongly selected targets, as well as the number of fixations per each method.

F. User Study 2: Dart-based Test

In this experiment the subject was to select, as accurately as possible, the bull's-eye of a dart target using each interaction method. In order to take into consideration the fact that eye

tracking has different accuracy in different regions of the monitor [32], we computed an average value based on five trials for each interaction method where the stimulus was shown at different areas of the screen near the center of the screen randomly. Each new randomly chosen trial began two seconds after selection of the previous target, allowing users time to change their gaze and to focus on the new target. For the dwell-time method, a countdown (5 to 0) representing remaining 100 milliseconds was displayed during the selection phase and users needed to focus on the dart shape before this time was up.

1) Stimulus: The stimulus for this experiment consisted of a dart-like target with three circles: green (0 to 30 pixels radius), blue (30 to 60 pixels radius) and red (60 to 90 pixels radius) as in Figure 1c. Points within the center area i.e. green have the lowest range of distances to the bulls-eye; each other co-centric circle has a larger range of distance values. Any point lying outside the three co-centric circular areas is considered as having a fixed maximum distance of 90 pixels. For this experiment, a cross-hair icon was used.

2) Measures: The purpose of this test was to measure the selected point's distance on the dart target to the center of the core circle (in green), thus the accuracy is measured in pixels. The distance between the selected location and the center of the stimulus is calculated based on the Euclidean distance. Since the measured trials are chosen randomly, the average is calculated to compare the two different methods based on accurate selection. In addition, the number of fixation points for each method was recorded.

G. Test Workflow

The order of interaction methods was randomly selected for each participant. At the end of the two studies subjects filled out a post-test questionnaire, which among other questions consisted of the NASA TLX questionnaire [6].

V. Results

We analyzed the results of our experiments using an analysis of variance (ANOVA) followed by Bonferroni posthoc tests with the JASP 0.11.1 software[2].

A. User Study 1: Matrix-based Test

A one-way repeated measure ANOVA was performed to examine the effect of interaction type on (1) number of fixations, (2) fixation duration time, (3) error rate, and (4) eye fatigue load index. Since we calculate average values on the entire test session for each participant, we can ignore the difficulty level factor in the analysis and take the total number of targets (40) into account.

1) Number of fixations: We found a significant effect of interaction method on average number of fixations (F(1,32)=7.79, $p < .05$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 262.97\ fixations$, $SE = 34.06\ fixations$) and voice recognition ($M = 425.84\ fixations$, $SE = 68.75\ fixations$).

[2]https://jasp-stats.org/

2) Fixation duration time: We found a significant effect of interaction method on average fixation duration (F(1,32)=32.93, $p < .001$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 16.52\ sec$, $SE = 1.32\ sec$) and voice recognition ($M = 39.77\ sec$, $SE = 3.97\ sec$).

3) Error rate: We found a significant effect of interaction method on error rate (F(1,32)=5.26, $p < .05$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 0.12\ errors$, $SE = 0.03\ errors$) and voice recognition ($M = 0.05\ errors$, $SE = 0.01\ errors$). Table I summarizes test results of the Matrix-based test.

4) Error locations on screen: Previous research has shown that the right side of a monitor has lower precision for eye tracking applications [32]. We studied the regions of the screen in regard to errors. We divided the screen size into nine equally-sized squares and counted the number of errors occurring in each location. In our study, errors are defined as wrongly selected targets (depicted in red in Figure 1b). Errors on the borders were counted for all adjacent regions. For instance, errors which occur in two regions are counted as occurring in both regions. Figure 1d illustrates the total number of errors for all participants for both interaction techniques.

5) Eye fatigue load index (performance-based): We found a significant effect of interaction method on our eye fatigue load index (F(1,32)=24.09, $p < .001$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 17.24$, $SE = 1.2$) and voice recognition ($M = 11.85$, $SE = 0.94$). Figure 3a illustrates the calculated performance-based eye fatigue load index for the Matrix test. This confirms our second hypothesis that $FELiX_{per}$ is higher for dwell-time than voice recognition.

	Dwell-Time	Voice Recog.	Sig.
Mean number of fixations	262.97	425.84	$p < .05$
Mean fixation duration (sec.)	16.52	39.77	$p < .001$
Error rate	0.12	0.05	$p < .05$

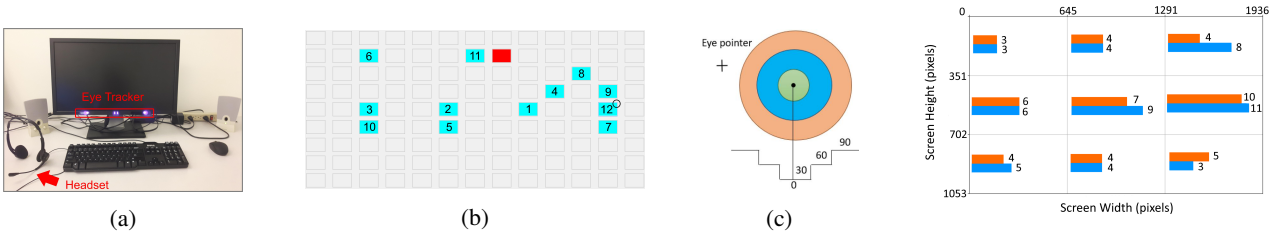
TABLE I: Test results of the Matrix-based test. Dwell-Time caused significantly more errors as expected.

B. User Study 2: Dart-based Test

A one-way repeated measure ANOVA was performed to examine the effect of interaction type on (1) number of fixations, (2) average distance to target, and (3) eye fatigue load index.

1) Number of fixations: We found a significant effect of interaction method on average number of fixations (F(1,32)=26.38, $p < .001$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 455.52\ fixations$, $SE = 1.71\ fixations$) and voice recognition ($M = 1379.66\ fixations$, $SE = 179.17\ fixations$).

2) Average distance to target: We found a significant effect of interaction method on average distance to target (F(1,32)=8.33, $p < .05$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 35.30\ pixels$, $SE = 2.11\ pixels$) and voice recognition



(a) (b) (c) (d)

Fig. 1: (a) shows test setting and equipment for both user studies. (b) shows the matrix-based test. The red button represents an error selection. The circle on number 12 represents the eye pointer. (c) shows the Dart-based test stimuli, and (d) shows error locations on screen. Orange bars represent total number of errors for voice recognition, and blue bars for dwell-time method.

($M = 29.27\ pixels$, $SE = 2.07\ pixels$). Since our accuracy-based FELiX (see equation 4) calculates the average distance to target in its eye tracking coefficient ($\frac{ANF}{ADT}$), it is similar with the FQlS measure [22] in measuring distance to target. In comparing the two measures, we found that $FELiX_{acc}$ decreases when the distance to target increases. In other words, higher distance to the target (lower accuracy) is associated with lower eye fatigue (Figure 2a). On the contrary, $FELiX_{per}$ increases with distance to target (Figure 2b). Table II summarizes the test results of the Dart-based test.

	Dwell-Time	Voice Recog.	Sig.
Mean number of fixations	455.52	1379.66	$p < .001$
Average distance to target	35.30	29.27	$p < .05$

TABLE II: Test results of the Dart-based test. Dwell-Time reached significantly lower number of fixations as expected.

3) Eye fatigue load index (accuracy-based): We found a significant effect of interaction method on eye fatigue load index (F(1,32)=31.74, $p < .001$). A posthoc Bonferroni comparison test showed a significant difference between dwell-time ($M = 4.28$, $SE = 0.26$) and voice recognition ($M = 12.96$, $SE = 1.53$). Figure 3b illustrates the calculated accuracy-based eye fatigue load index for the Dart test. This result confirms our first hypothesis that dwell-time has a lower $FELiX_{accc}$ score than voice recognition.

C. Bi-variate Comparison

We proposed two variations on different criteria (performance and accuracy). Each interaction technique can be analyzed on both measures. A one-way repeated measure ANOVA was performed to examine the effect of interaction type on the mean of both FELiX variations. We found no significant effect of interaction method on bivariate eye fatigue load index (F(1,32)=3.77, $p > .05$). A posthoc Bonferroni comparison test showed no significant difference between dwell-time ($M = 10.76$, $SE = 0.53$) and voice recognition ($M = 12.40$, $SE = 0.86$). Figure 3c illustrates the calculated bivariate (performance-, and accuracy-based) average of eye fatigue load index. Table III summarizes calculated FELiX values on both criteria.

	Dwell-Time	Voice Recog.	Sig.
$FELiX_{per}$	17.24	11.85	$p < .001$
$FELiX_{acc}$	4.28	12.96	$p < .001$
Bi-variate FELiX	10.76	12.40	$p > .05$

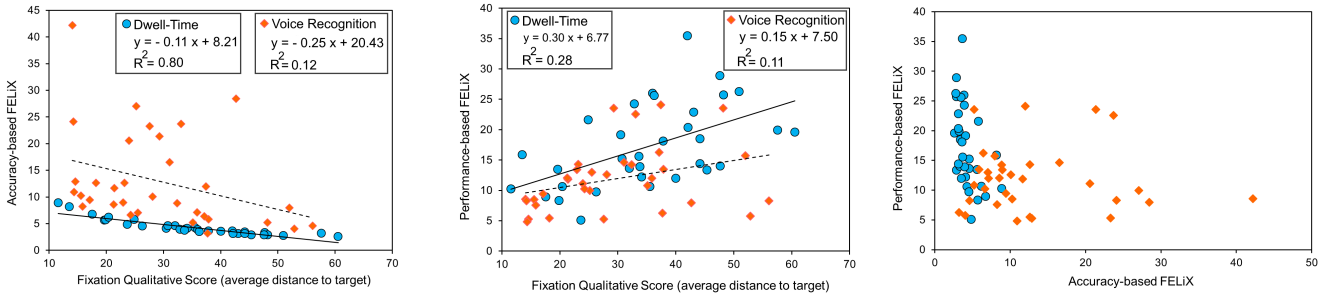
TABLE III: Test results of FELiX calculations. Dwell-Time caused significantly higher eye fatigue based on performance and lower eye fatigue based on accuracy as expected.

D. NASA TLX Scores

Figure 3d shows the required NASA TLX scores by FELiX variations from the post-test questionnaire.

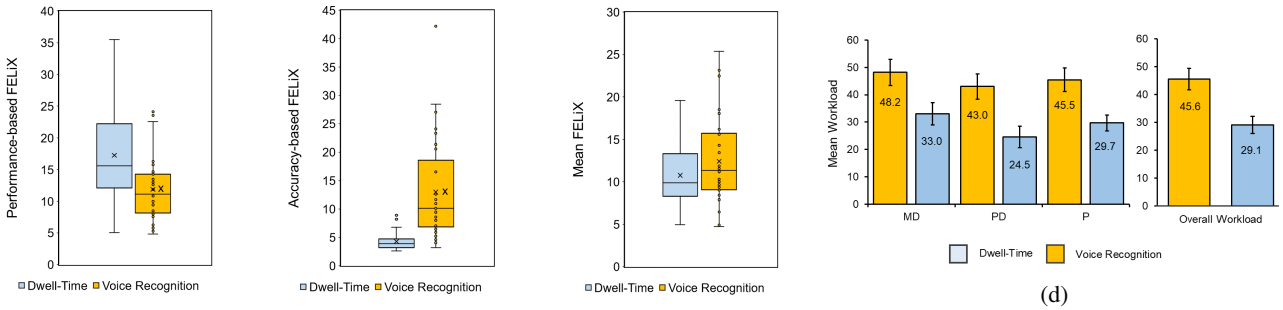
VI. Discussion

The results indicate that the developed multi-factor simple-to-calculate measure, which is solely dependent on the recorded data of a user study, can be used to accurately assess the amount of eye fatigue on participants based on available measures and NASA TLX scores. Further, we showed how to compare our measures with the available FQlS measure and illustrated the correlations between them (Figures 2a, 2b). Although we only studied voice recognition as a multimodal gaze-based interaction technique, the dwell-time results confirmed our assumptions that it results in a lower number of fixations and lower fixation duration time compared to a multimodal interaction technique. Although dwell-time showed lower accuracy (higher distance to the target) than voice recognition (see Table II), it reached significantly lower eye fatigue based on accuracy (see Table III and Figure 3b) confirming our first hypothesis that $FELiX_{acc}$ is lower for dwell-time. This is due to a significantly lower number of fixations (Table II) and lower TLX scores (Figure 3d) for dwell-time. The higher distance to the target for dwell-time is due to the activation threshold which bounds a user's decision time into a limited time window to respond to target movements. The results of the performance-based FELiX depicted in Figure 3a shows higher eye fatigue for the dwell-time technique. This is due to higher error rate and higher duration of fixation points ($\frac{ANF}{AFD}$) of dwell-time as expected (see table I). This confirms our second hypothesis that $FELiX_{per}$ is higher for dwell-time than voice recognition. Although the bivariate comparison of both FELiX variations (Figure 3c) shows relatively lower eye



(a) (b) (c)

Fig. 2: Correlations of the accuracy-based (a) and performance-based (b) FELiX with fixation qualitative score (FQlS), for 33 participants. Dashed lines represent regression through voice recognition and solid lines through dwell-time. (c) shows eye fatigue load index on both variations. Voice recognition technique shows sparse values on both variations.



(a) (b) (c) (d)

Fig. 3: (a) shows performance-based eye fatigue load index for the Matrix test ($p < .001$), and (b) shows accuracy-based eye fatigue load index for the Dart test ($p < .001$). (c) shows the calculated mean of both variations ($p > .05$). The cross symbols show mean, and the horizontal lines show median points. (d) shows NASA TLX scores. Error bars represent standard error.

fatigue for dwell-time, the difference is statistically not significant (see Table III). Additionally, Figure 2c shows distinctive clusters of dwell-time and voice recognition techniques based on FELiX variations and reflects the potential of FELiX measure to analyze similar eye tracking techniques based on their eye fatigue values, and therefore our third hypothesis that dwell-time can be distinguished from a multi-modal interaction technique based on FELiX variations is confirmed. We believe that these results would generalize, and that FELiX is an effective means of determining eye fatigue load and can differentiate different gaze-based interaction methods based on their tendencies to cause the user more discomfort in terms of visual fatigue and task load. We also studied the role of target locations on screen and their relation with error rate and eye fatigue. As illustrated in Figure 1d, the middle row of the screen, towards the right side, has higher eye fatigue potential according to the performance-based FELiX as these regions produced higher errors. Since we applied no biological sensor devices in our user studies, we could not compare the results to study the correlations between our proposed measure and physiological data. We leave this for future work. We did, however, demonstrate that FELiX is an alternative measure to be used in user studies with no access to electronic sensors. Although the eye tracking parameters involved in FELiX measure can be analyzed individually, the emerging interaction devices offer a variety of quantitative parameters. Therefore, the application of different parameters may be difficult to compare different techniques. The analysis of our results indicates the potential of our multi-aspect evaluation measure on two similar interaction techniques. This experiment provides new insight into the feasibility of multi-factor compound evaluation measures for gaze-based interactions.

VII. Conclusion and Future Work

As emerging interaction techniques become more sophisticated and multi-dimensional, the need for more complex and multi-factor measures is necessary. Therefore, we propose *fixation-based eye fatigue load index* (FELiX), a compound evaluation measure for gaze-based interactions based on the NASA TLX scores and recorded eye tracking data. Our measure combines the quantitative (technical) and qualitative (empirical) aspects of interaction techniques in a simple-to-calculate measure. Since NASA TLX scores are very common in user studies, we can take benefit of its simplicity to assess cognitive workload of different interaction techniques on the same tasks. FELiX includes two variations to measure visual eye fatigue based on (a) performance, and (b) accuracy. These measures enable researchers to compare different eye tracking techniques, specially dwell-time and multi-modal techniques,

based on eye fatigue load index on different criteria. The *performance-based* measure can be applied when the duration of the entire fixation sequences and the error rates of target selection are recorded, and the *accuracy-based* measure is applicable for case scenarios where distance to target (selection accuracy) is available in the analysis process and can be measured in user studies. Both measures take benefit of three scores from the NASA TLX, (a) physical demand, (b) mental demand, and (c) performance. The application of the proposed measures can be regarded as a feasible alternative to biological sensor inputs or to adopt gaze-based applications for children, users with disabilities or elderly users to assess the amount of eye fatigue in user studies before final release of eye tracking applications. In addition, we presented an in-depth analysis of the dwell-time method as the most common gaze-based interaction technique with different approaches. As well as developing measures for eye fatigue load, we proposed two test applications to analyze eye tracking applications. In future work, we plan on applying the proposed eye fatigue measures on VR headsets with integrated eye trackers to study motion sickness in VR applications.

VIII. Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada Grants DG-N01670 and DG-N06722.

References

[1] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway, *Robust multimodal cognitive load measurement*. Springer, 2016.

[2] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

[3] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[4] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[5] J. F. Ruiz-Rabelo, E. Navarro-Rodriguez, L. L. Di-Stasi, N. Diaz-Jimenez, J. Cabrera-Bermon, C. Diaz-Iglesias, M. Gomez-Alvarez, and J. Briceño-Delgado, "Validation of the nasa-tlx score in ongoing assessment of mental workload during a laparoscopic learning curve in bariatric surgery," *Obesity surgery*, vol. 25, no. 12, pp. 2451–2456, 2015.

[6] N. H. P. R. Group, "Nasa task load index (tlx) paper and pencil package," 1986. [Online]. Available: https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf

[7] J. Zagermann, U. Pfeil, and H. Reiterer, "Measuring cognitive load using eye tracking technology in visual computing," in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, ser. BELIV '16. New York, NY, USA: ACM, 2016, pp. 78–85. [Online]. Available: http://doi.acm.org/10.1145/2993901.2993908

[8] J. Annett, "Subjective rating scales: science or art?" *Ergonomics*, vol. 45, no. 14, pp. 966–987, 2002.

[9] M.-W. O. Dictionary, "electroencephalography." [Online]. Available: https://www.merriam-webster.com/dictionary/electroencephalography

[10] M.-W. Dictionary, "magnetoencephalography." [Online]. Available: https://www.merriam-webster.com/dictionary/magnetoencephalography

[11] Y. Ishii, H. Ogata, H. Takano, H. Ohnishi, T. Mukai, and T. Yagi, "Study on mental stress using near-infrared spectroscopy, electroencephalography, and peripheral arterial tonometry," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 4992–4995.

[12] M. Vasiljevas, T. Gedminas, A. Ševčenko, M. Jančiukas, T. Blažauskas, and R. Damaševičius, "Modelling eye fatigue in gaze spelling task," in *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2016, pp. 95–102.

[13] J. Hawley and T. Reilly, "Fatigue revisited." *Journal of sports sciences*, vol. 15, no. 3, p. 245, 1997.

[14] S. M. Marcora, W. Staiano, and V. Manning, "Mental fatigue impairs physical performance in humans," *Journal of applied physiology*, vol. 106, no. 3, pp. 857–864, 2009.

[15] I. IWA, "ISO, image safety: Reducing the incidence of undesirable biomedical effects caused by visual image sequence," 2005.

[16] J. Kuze and K. Ukai, "Subjective evaluation of visual fatigue caused by motion images," *Displays*, vol. 29, no. 2, pp. 159–166, 2008.

[17] K. Ukai and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, no. 2, pp. 106–116, 2008.

[18] B. Zheng, X. Jiang, G. Tien, A. Meneghetti, O. N. M. Panton, and M. S. Atkins, "Workload assessment of surgeons: correlation between nasa tlx and blinks," *Surgical endoscopy*, vol. 26, no. 10, pp. 2746–2750, 2012.

[19] G. Borghini, G. Vecchiato, J. Toppi, L. Astolfi, A. Maglione, R. Isabella, C. Caltagirone, W. Kong, D. Wei, Z. Zhou *et al.*, "Assessment of mental fatigue during car driving by using high resolution eeg activity and neurophysiologic indices," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 6442–6445.

[20] S. Lanthier, E. Risko, D. Smilek, and A. Kingstone, "Measuring the separate effects of practice and fatigue on eye movements during visual search," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, no. 35, 2013.

[21] E. Abdulin and O. Komogortsev, "User eye fatigue detection via eye movement behavior," in *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*. ACM, 2015, pp. 1265–1270.

[22] O. V. Komogortsev, D. V. Gobert, S. Jayarathna, D. H. Koh, and S. M. Gowda, "Standardization of automated analyses of oculomotor fixation and saccadic behaviors," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 11, pp. 2635–2645, 2010.

[23] T. W. Calvert, E. W. Banister, M. V. Savage, and T. Bach, "A systems model of the effects of training on physical performance," *IEEE Transactions on systems, man, and cybernetics*, no. 2, pp. 94–102, 1976.

[24] P. Majaranta, U.-K. Ahola, and O. Špakov, "Fast gaze typing with an adjustable dwell time," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 357–360.

[25] A. T. Bahill and L. Stark, "Overlapping saccades and glissades are produced by fatigue in the saccadic eye movement system," *Experimental neurology*, vol. 48, no. 1, pp. 95–106, 1975.

[26] T. Megaw and T. Sen, "Visual fatigue and saccadic eye movement parameters," in *Proceedings of the Human Factors Society Annual Meeting*, vol. 27, no. 8. Sage Publications Sage CA: Los Angeles, CA, 1983, pp. 728–732.

[27] L. L. Di Stasi, M. Marchitto, A. Antolí, and J. J. Cañas, "Saccadic peak velocity as an alternative index of operator attention: A short review," *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, vol. 63, no. 6, pp. 335–343, 2013.

[28] O. Špakov and D. Miniotas, "On-line adjustment of dwell time for target selection by gaze," in *Proceedings of the third Nordic conference on Human-computer interaction*. ACM, 2004, pp. 203–206.

[29] I. S. MacKenzie, "Evaluating eye tracking systems for computer input," in *Gaze interaction and applications of eye tracking: Advances in assistive technologies*. IGI Global, 2012, pp. 205–225.

[30] V. Sundstedt, "Gazing at games: An introduction to eye tracking control," *Synthesis Lectures on Computer Graphics and Animation*, vol. 5, no. 1, pp. 1–113, 2012.

[31] H. Istance, R. Bates, A. Hyrskykari, and S. Vickers, "Snap clutch, a moded approach to solving the midas touch problem," in *Proceedings of the 2008 symposium on Eye tracking research & applications*, 2008, pp. 221–228.

[32] A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, and M. R. Morris, "Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design," in *Proceedings of the 2017 Chi conference on human factors in computing systems*. ACM, 2017, pp. 1118–1130.