Can We Predict Stressful Technical Interview Settings Through Eye-tracking?

Mahnaz Behroozi
North Carolina State University
Raleigh, NC, USA
mbehroo@ncsu.edu

Chris Parnin
North Carolina State University
Raleigh, NC, USA
cjparnin@ncsu.edu

ABSTRACT

Recently, eye-tracking analysis for finding the cognitive load and stress while problem-solving on the whiteboard during a technical interview is finding its way in software engineering society. However, there is no empirical study on analyzing how much the interview setting characteristics affect the eye-movement measurements. Without knowing that, the results of a research on eye-movement measurements analysis for stress detection will not be reliable. In this paper, we analyzed the eye-movements of 11 participants in two interview settings, one on the whiteboard and the other on the paper, to find out if the characteristics of the interview settings affect the analysis of participants' stress. To this end, we applied 7 Machine Learning classification algorithms on three different labeling strategies of the data to suggest researchers of the domain a useful practice of checking the reliability of the eye-measurements before reporting any results.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*;

KEYWORDS

technical interviews, stress detection, eye-tracking, data mining, machine learning

ACM Reference Format:
Mahnaz Behroozi and Chris Parnin. 2018. Can We Predict Stressful Technical Interview Settings Through Eye-tracking?. In *EMIP '18: Symposium on Eye Movements in Programming, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3216723.3216729

1 INTRODUCTION

Problem-solving on the whiteboard is a common type of a technical interview for software developers. With this interview format, an interviewer is able to observe the thought processes of a candidate and interact with candidate by asking questions or providing hints. Unfortunately, software developers can be uncomfortable with this style of interview, often because they are being asked to a) think-aloud while problem solving, which can cause high cognitive load [Dawson 2003], and b) perform under time pressure and fear of failure, which can cause high stress in candidates [Caviola et al. 2017]. As a result, technical interviews may bias the selection of candidates toward those who perform well under pressure or have had more opportunity to engage in extensive interview preparation. Although research has investigated eye movement of programmers when reading source code [Busjahn et al. 2011], and measured task difficulty [Fritz et al. 2014] using psycho-physiological measures such as electroencephalography (EEG), eye-gaze, electrodermal activity (EDA), limited research has investigated coding in technical interview settings.

Previously, we performed a study [Behroozi et al. 2018] using eye-tracking to investigate if problem-solving in a whiteboard setting contributed to high levels of cognitive load and stress. Eleven participants solved programming problems in two settings: either on a whiteboard in front of an interviewer or privately on a paper. Based on eye movement data, we observed significantly different measures that were previously associated in literature with high cognitive load and stress. Specifically, we observed significantly shorter duration for fixations and higher saccade duration average and saccade velocity average when participants were solving problems on the whiteboard than compared to on paper.

While our study providing some preliminary insights, there were several limitations with our analysis. Because we only focused on observing differences in eye movement measures across interview settings, we had limited ability to distinguish how different features and settings could interact. Eye movement characteristics highly depends on the context of a particular task and the environment during everyday actions [Foulsham 2015]. Thus, it is an open question that how much the interview setting or other factors affects the eye movement measurements and our resulting interpretations.

In this paper, we extend the analysis from our previous study [Behroozi et al. 2018] in several ways. First, we consider in addition to the interview setting, a self-rated ranking of stress by the participant in our analysis. Second, rather than using aggregated measures across the entire session, we also analyze individual eye movement events. Third, to support a better understanding of how the different eye movement measures interact, we build prediction models using various machine learning classification algorithms.

We found that our train classifiers could predict self-reported stress levels of participants reliably when trained for a specific interview setting (i.e. either whiteboard or paper). When attempting to train classifiers that mixed the settings in order to predict stress, their performance declined. Our classifiers could also reliably predict the interview setting. We believe these results complement our previous findings. Future applications can incorporate these techniques for building training/coaching tools for interview applicants or to enhance the analysis of eye movement data.

EMIP '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5792-0/18/06…$15.00
https://doi.org/10.1145/3216723.3216729

2 APPROACH

2.1 Data set

In this study we use a data set consisting of the eye movement events of 11 participants, consisting 8 computer science graduate and 3 undergraduate students, while solving two coding problems of the same difficulty once on the whiteboard and once on the paper. The data has been collected using SMI head-mounted eye-tracker with sampling rate of 60Hz. At the end of each problem solving session, the participants has been asked to determine which setting was stressful for them and which was not. 4 participants reported that they were stressed on the paper but not on the whiteboard and the rest reported the reverse.

Eye movement consists of three events: VI, saccade and blink. As Behroozi et al.'s study showed that VI-related and saccade-related measurements were statistically different in two interview settings, we extracted the VI and saccade events to study more and ignored blink events for this study. Table 1 shows the 8 VI-related and 9 saccade-related measurements we used in this study.

In order to find out if the characteristics of the interview setting affects the eye movement measurements or not, we devised three partitioning methods on the data set (see Table 2 for more information):

(1) **Setting and Stress Rating**
Labeling based on user survey (stressed/not stressed) given the interview setting: In this labeling, we split the data into 4 parts: saccade in whiteboard setting, VI in whiteboard setting, saccade in paper setting, VI in paper setting.
(2) **Setting**
Labeling based on the paper and whiteboard setting: Here we partitioned the data into two parts: all the saccade-related measurement in one data set and all the VI-related measurements in another data set.
(3) **Stress Rating**
Labeling based on user survey (stressed/not stressed): Here we partitioned the data into two parts: all the saccade-related measurement in one file and all the VI-related measurements in another file.
By comparing the results from this labeling with Setting and Stress Rating labeling, we will be able to see how robust are the classifiers to predict the stress with or without being provided the interview setting.

2.2 Classification Algorithms

We applied the following classification algorithms on the data, after applying three labeling strategies, using the WEKA data mining software [Hall et al. 2009; Witten et al. 2016]:

We applied seven classification algorithms: Naive Bayes (NB) [John and Langley 1995], Random Forest (RF) [Breiman 2001], Multi-Layer Perceptron (MLP), SVM [Keerthi et al. 2001; Zeng et al. 2008], KNN [Aha et al. 1991], Logistic Regression (LR) [Le Cessie and Van Houwelingen 1992] and Decision Tree (DT) [Salzberg 1994]. Table4 shows the detailed setting information of each classifier in WEKA.

Table 1: List of VI-based and saccade-based measurements

VI-based measurements	Saccade-based measurements
Visual Intake Duration [ms]	Saccade Duration [ms]
Visual Intake Position X [px]	Saccade Start Position X [px]
Visual Intake Position Y [px]	Saccade Start Position Y [px]
Visual Intake Average Pupil Size X [px]	Saccade End Position X [px]
Visual Intake Average Pupil Size Y [px]	Saccade End Position Y [px]
Visual Intake Average Pupil Diameter [mm]	Saccade Amplitude [°]
Visual Intake Dispersion X [px]	Saccade Acceleration Average [$°/s^2$]
Visual Intake Dispersion Y [px]	Saccade Velocity Average [°/s]
	Saccade Peak Velocity at [%]

2.3 Performance Measures and Statistical Methods

2.3.1 ***Model Validation:*** Throughout the study, we used 10-fold-cross-validation technique for validating the generalization of our results obtained from applying classification algorithms on our data [Kohavi et al. 1995].

2.3.2 ***Statistical analysis of binary classification:*** To evaluate the accuracy of the results obtained from classifiers, we reported weighted F-measure along with accuracy, weighted precision and weighted recall of both classes. To illustrate the diagnostic ability of each of the binary classifiers on our data, we also reported the area under receiver operating characteristic curve (ROC Area) [Powers 2011].

3 RESULTS

From Table 3 we can see that classifiers were more successful and accurate on Setting and Stress Rating labeling, specially when it comes to considering VI-related measurements. This shows that VI-related and saccade-related measurements are representative of stress in each of the settings. Comparing the results from Setting labeling and Stress labeling, the classifiers performed better on Setting labeling, except for the saccade in top 3 classifiers where Stress labeling performed slightly better. Hence, we can infer that the properties of the setting affects the eye-movement measurements across the interview settings. Although the effect of the interview settings did not decrease the performance of the classifiers in Stress labeling dramatically, it reveals the importance of checking any potential effect of the setting on eye-movement measurements. The results still confirms the hypothesis that the whiteboard setting bears more stress on participants.

The best performing classifier for all the labeling strategies was Random Forest. It achieved F-measure of 0.88, 0.77 and 0.78 or better for Setting and Stress Rating labeling, Setting labeling and Stress labeling respectively. The worse performing classifiers were SVM and Naive Bayes with F-measures as worse as 0.43 and not better than 0.80.

We did not apply normalization on the data set since each of the features are meaningful and scaling might hurt the quality of the data. Hence, it was expected that Random Forest performs better in this context and SVM fails to demonstrate its power. Because Random Forest is a tree-based ensemble classifier and, like decision trees, it is a graphical-model based classifier. But SVM tries to maximize the margin and it relies on the concept of distance. Thus, SVM cannot tolerate different scales of the features. Applying SVM with different kernels or finding the best feature scaling and shaping

Table 2: Data set information

	Setting and Stress Rating				Setting		Stress Rating	
	Board Saccade	Board VI	Paper Saccade	Paper VI	Saccade	VI	Saccade	VI
Number of columns (features)	9	8	9	8	9	8	9	8
Number of rows (records)	6694	7451	10440	11723	17134	19174	17134	19174
Ratio of records labeled as stressed to not stressed	5200/1494	5696/1755	4781/5659	5537/6186	-	-	9981/7153	11233/7941
Ratio of records labeled as whiteboard to paper	-	-	-	-	6694/10440	7751/11723	-	-

Table 3: Results from applying classifiers on the labeling based on the participant survey (stressed/not stressed) using 10-fold cross validation. Sorted in descending order of accuracy.

		Setting and Stress Rating				Setting		Stress Rating	
Classifier	Measure	Board Saccade	Board VI	Paper Saccade	Paper VI	Saccade	VI	Saccade	VI
RF	Accuracy%	90.35	95.80	87.92	96.80	77.52	96.45	78.05	94.86
	Precision	0.90	0.96	0.88	0.97	0.77	0.97	0.78	0.95
	Recall	0.90	0.96	0.88	0.97	0.78	0.97	0.78	0.95
	F-measure	0.90	**0.96**	0.88	**0.97**	0.77	**0.97**	0.78	**0.95**
	ROC Area	0.94	0.99	0.95	0.99	0.85	0.99	0.86	0.99
DT	Accuracy%	86.93	93.83	84.89	95.28	72.20	94.63	72.45	92.24
	Precision	0.87	0.94	0.85	0.95	0.72	0.95	0.72	0.92
	Recall	0.87	0.94	0.85	0.95	0.72	0.95	0.73	0.92
	F-measure	0.87	0.94	0.85	0.95	0.72	0.95	0.72	0.92
	ROC Area	0.82	0.94	0.87	0.96	0.76	0.95	0.76	0.94
KNN	Accuracy%	81.10	92.08	80.52	96.71	66.66	93.10	67.16	90.39
	Precision	0.81	0.92	0.81	0.97	0.67	0.93	0.67	0.90
	Recall	0.81	0.92	0.81	0.97	0.67	0.93	0.67	0.90
	F-measure	0.81	0.92	0.81	0.97	0.67	0.93	0.67	0.90
	ROC Area	0.72	0.90	0.80	0.97	0.65	0.93	0.66	0.90
MLP	Accuracy%	81.40	92.08	82.81	93.07	65.44	89.02	60.36	75.02
	Precision	0.80	0.92	0.83	0.93	0.65	0.89	0.61	0.75
	Recall	0.81	0.92	0.83	0.93	0.65	0.89	0.60	0.75
	F-measure	0.78	0.92	0.83	0.93	0.61	0.89	0.52	0.75
	ROC Area	0.79	0.96	0.88	0.97	0.65	0.95	0.63	0.84
LR	Accuracy%	83.25	90.06	79.44	77.57	66.17	75.56	62.55	64.23
	Precision	0.82	0.90	0.80	0.78	0.67	0.75	0.63	0.63
	Recall	0.83	0.90	0.80	0.78	0.66	0.76	0.63	0.64
	F-measure	0.81	0.90	0.80	0.78	0.61	0.75	0.58	0.63
	ROC Area	0.83	0.91	0.84	0.83	0.69	0.79	0.66	0.69
SVM	Accuracy%	78.03	81.01	79.89	77.21	62.61	75.27	58.28	61.40
	Precision	0.80	0.81	0.80	0.78	0.72	0.76	0.63	0.66
	Recall	0.78	0.81	0.80	0.78	0.63	0.75	0.58	0.61
	F-measure	0.69	0.77	0.80	0.78	0.50	0.74	0.43	0.52
	ROC Area	0.51	0.61	0.80	0.78	0.52	0.71	0.50	0.54
NB	Accuracy%	80.83	80.08	56.51	72.77	63.30	76.28	57.93	59.54
	Precision	0.80	0.781	0.65	0.73	0.64	0.77	0.54	0.58
	Recall	0.80	0.80	0.57	0.73	0.63	0.76	0.58	0.60
	F-measure	0.77	0.77	0.45	0.73	0.55	0.75	0.48	0.53
	ROC Area	0.69	0.82	0.75	0.76	0.63	0.79	0.54	0.63

techniques [Chapelle and Keerthi 2011; Forman et al. 2009] might help SVM to perform better but this is not the concern of this study.

Another interesting result is that VI-related measurements were more representative of both the interview setting and the stress of the participants. In other words, values of VI-related measurements are more distinctive than saccade-related measurements. When the top three classifiers do classification based on VI-related measurements, they show F-measures 0.92-0.97, 0.93-0.97 and 0.90-0.95 for Setting and Stress Rating labeling, Setting labeling and Stress labeling, respectively. However, when they use saccade-related measurements, they cannot achieve better F-measures than 0.81-0.90, 0.67-0.77 and 0.67-0.78 for Setting and Stress Rating labeling, Setting labeling and Stress labeling, respectively.

Results obtained from applying classifiers on three different labeling of the data showed that the eye-movement measurements are representative of stress if they are analyzed in an interview

Table 4: Detailed parameter settings of the classifiers in WEKA

Classifier	Parameter setting
RF	-P 100 -I 100 -num-slots 1 -K 0 -M 1 -V 0.001 -S 1
DT	-C 0.25 -M 2
KNN	-K 1 -W 0 -A "wka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclidianDistance -R first-last"
MLP	-L 0.3 -M 02 -N 500 -V 0 -S 0 -E 20 -H a
LR	I 0 -M 500 -H 50 -W 0
SVM	-C 1 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"
NB	-

setting without any mixture of the records from another setting. They are also representative of the properties of the interview setting. This means that eye-movement shows the stress differently from setting to setting. This is the reason why classifiers performed best when applied on Setting and Stress Rating labeling.

4 DISCUSSION

4.1 Prediction Approaches

In this study we separated VI-related and saccade-related measurements for the following reasons. First, concentrate on each of these two eye-movement events to see their prediction ability in absence of another one. Because there is a probability that each eye-movement event be affected by the characteristic of the interview setting differently. Second, the eye-movement events take place at each moment exclusively. Also, the number of the features for VI events and saccade events are not the same.

But it is still interesting to find an strategy to combine the VI-related and saccade-related measurement, such as fitting them in time windows, to see if they can enhance the performance of the classifiers.

4.2 Other Measures

There are other measurements which we can take into account in future studies such as blink information. We can also investigate AOI, revisits, information seeking patterns (linearity), etc. Lallé et al. in [Lallé et al. 2016] tried to reveal confusion through eye tracking. In their study they found that pupil size and the distance of the participant's head from the screen can show confusion. It is worth studying whether it is also representative of stress or not.

4.3 Applications

Another interesting application is automated detection of stress towards coaching and training. For example, in this application, we can break a coaching/training process to several sessions and analyze the stress level of a participant and customize coaching/training for each person based on their stress. As another example, suppose that we have a number of professional interview settings or even different IDEs for the candidate to take the interview. If the candidate is not sure about which setting to choose, we can help them with giving them a chance to examine different settings and the classifiers can do real-time decision making on suggesting the setting in which the candidate was less stressed. It is worth noting that for the last application, we need sufficient labeled eye-movement records in each of the settings in order to train the classifiers.

In some cases, stress handling might be critical for recruiting a candidate. As having stress is not always devastating but might be a means of pushing someone to better finish a task, it might be interesting for some recruiters to see how a stressed candidate benefits from controlling his/her stress to finish a task in a timely manor or in a better shape.

Finally, our approach may be considered in the future for other contexts with dynamic environments, such as measuring eye-tracking while driving.

4.4 Limitations and Future Work

Stress level varies throughout an interview session. Hence, it is better to consider the stress finding problem, as a multi-class classification problem rather than binary class classification. Also, labeling the whole interview session can obscure the results of the analysis as the stress level is not constant during the session. As a result, it might be better to break the session into specific time frames and find a way to ask the candidates to survey what is their stress level at that moment. Besides eye-tracking measurements, integrating the eye-movement analysis with other bio-metrics such as heart rate can enhance the predictability of the stress.

In classification problems, imbalance data can affects the results negatively. Although the data we used in this study is not severely imbalance, still devising oversampling and undersampling techniques to balance the class distribution can enhance the results. The most common technique for oversampling is Synthetic Minority Over-sampling Technique (SMOTE) [Chawla et al. 2002]. Cluster centroids [Yen and Lee 2009] and Tomek links [Tomek 1976] are the examples of undersampling techniques.

Feature scaling is another concern in the data we used. Since some of the eye-movement measurements such as saccade velocity peak have large ranges, they can affect the accuracy of the classifiers, specially those that do distance based classification such as SVM. But at the same time, we might lose data information by feature scaling.

To generalize our findings, we need a larger data set. Although each experiment has thousands of records, still we have to investigate more experiments to be confident about generalization of the results.

5 CONCLUSION

Comprehending the stress level of the participants during technical interviews helps toward refining the interview process. Interview setting characteristic can affect eye movement measurements. Thus, it is important to check if the eye tracking measurements are representative of the cognitive load and stress or it is diluted by the interview setting characteristics. With the proposed approach we can make sure that a set of eye movement measurements are dependable and can be used toward stress and cognitive load study.

REFERENCES

David W Aha, Dennis Kibler, and Marc K Albert. 1991. Instance-based learning algorithms. *Machine learning* 6, 1 (1991), 37–66.

Mahnaz Behroozi, Alison Lui, Ian Moore, Denae Ford, and Chris Parnin. 2018. Dazed: Measuring the Cognitive Load of Solving Technical Interview Problems at the Whiteboard. In *40th International Conference on Software Engineering, NIER Track (ICSE'18)*.

Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

Teresa Busjahn, Carsten Schulte, and Andreas Busjahn. 2011. Analysis of Code Reading to Gain More Insight in Program Comprehension. In *Proceedings of the 11th Koli Calling International Conference on Computing Education Research (Koli Calling '11)*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/2094131.2094133

Sara Caviola, Emma Carey, Irene C Mammarella, and Denes Szucs. 2017. Stress, Time Pressure, Strategy Selection and Math Anxiety in Mathematics: A Review of the Literature. *Frontiers in psychology* 8 (2017), 1488.

Olivier Chapelle and S Sathiya Keerthi. 2011. Multi-class feature selection with support vector machines.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

Jane Dawson. 2003. Reflectivity, creativity, and the space for silence. *Reflective Practice* 4, 1 (2003), 33–39.

George Forman, Martin Scholz, and Shyamsundar Rajaram. 2009. Feature shaping for linear SVM classifiers. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 299–308.

Tom Foulsham. 2015. Eye movements and their functions in everyday tasks. *Eye* 29, 2 (2015), 196.

Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 402–413.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

George H John and Pat Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 338–345.

S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation* 13, 3 (2001), 637–649.

Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection.

Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data.

Saskia Le Cessie and Johannes C Van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied statistics* (1992), 191–201.

David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

Steven L Salzberg. 1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning* 16, 3 (1994), 235–240.

Ivan Tomek. 1976. Two modifications of CNN. 6 (11 1976).

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Show-Jane Yen and Yue-Shi Lee. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* 36, 3 (2009), 5718–5727.

Zhi-Qiang Zeng, Hong-Bin Yu, Hua-Rong Xu, Yan-Qi Xie, and Ji Gao. 2008. Fast training support vector machines using parallel sequential minimal optimization. In *Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on*, Vol. 1. IEEE, 997–1001.