Indicators of Training Success in Virtual Reality Using Head and Eye Movements

Joy Gisler*
ETH Zurich
Christian Hirt¶
ETH Zurich
Johannes Schneider†
University of Liechtenstein
Wolfgang Fuhl‖
University of Tuebingen
Joshua Handali‡
University of Liechtenstein
Jan vom Brocke**
University of Liechtenstein
Valentin Holzwarth§
University of Liechtenstein
Andreas Kunz††
ETH Zurich

ABSTRACT

An essential aspect in the evaluation of Virtual Training Environments (VTEs) is the assessment of users' training success, preferably in real-time, e.g. to continuously adapt the training or to provide feedback. To achieve this, leveraging users' behavioral data has been shown to be a valid option. Behavioral data include sensor data from eye trackers, head-mounted displays, and hand-held controllers, as well as semantic data like a trainee's focus on objects of interest within a VTE. While prior works investigated the relevance of mostly one and in rare cases two behavioral data sources at a time, we investigate the benefits of the combination of three data sources. We conduct a user study with 48 participants in an industrial training task to find correlations between training success and measures extracted from different behavioral data sources. We show that all individual data sources, i.e. eye gaze position and head movement, as well as duration of objects in focus are related to training success. Moreover, we find that simultaneously considering multiple behavioral data sources allows to better explain training success. Further, we show that training outcomes can already be predicted significantly better than chance by only recording trainees for parts of their training. This could be used for dynamically adapting a VTE's difficulty. Finally, our work further contributes to reaching the long-term goal of substituting traditional evaluation of training success (e.g. through pen-and-paper tests) with an automated approach.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing Methodologies—Machine learning—Machine learning approaches—Feature selection

1 INTRODUCTION

Virtual Training Environments (VTEs) gained popularity due to multiple reasons: i) they allow training personnel for hazardous procedures in a safe environment [3], ii) they allow training under fairly realistic conditions without the presence of trainers [18], iii) they provide a cost-efficient alternative to training in reality, since no real equipment and consumables are required [6], and iv) they allow for better assessment of training success and individualization of the VTE. This is important, since the performance of individuals varies considerably, and accounting for the ability of individual trainees might be helpful, as shown for Virtual Reality (VR) and other contexts such as adjusting explanations to trainees' competencies [26]. In our context, we aim to better understand how to assess training success in a VTE. This would be helpful to dynamically adapt the training and thus support trainees who are underperforming, e.g. by providing more guidance, time, or encouragement. Additionally, a traditional evaluation of training success (e.g. pen-and-paper knowledge tests or practical examination) could be substituted or enhanced by an automated approach. Aside from a few performance related aspects such as task completion time or binary information on training success (pass or failed), little is known without human analysis of trainees. Due to this, interventions during training in a VTE are often considered out-of-scope, i.e. it is difficult to adjust the level of support during training to prevent excellent trainees from getting bored or weak trainees from getting overwhelmed.

VTEs allow collecting a rich set of trainees' behavioral data, e.g. eye movements and velocities or accelerations of head movements. Furthermore, VTEs allow accessing semantic information. That is, detailed information readily available at any point in time for objects a user is looking at. The abundance of information on users is valuable when assessing their behavior. Information on head movements or eye tracking including the duration for which users focus on specific objects, as well as context switches and even other information such as physiological signals (e.g. heart rate) are used in prior work, e.g. for affective states [11], cybersickness [13], biometrics [24], anxiety [30], or stress levels [8]. However, the number of studies is rather limited. Furthermore, investigating the combination of various information sources to assess performance such as training success is limited to two data sources and also lacking combinations of data, e.g. from the VTE's objects in focus and Head-Mounted Display (HMD). Thus, it is interesting to assess the usefulness of combining low-level information extracted from eye and head movements, with high level, semantically rich information from the VTE such as relevant objects in focus. To gather this data, we conducted a user study, wherein 48 participants trained an industrial task in a VTE. We relate training success to eye tracking data, rotational data of the HMD, and the relevant virtual objects within the VTE, i.e. objects the trainees should focus on to understand the pressing procedure. Through statistical testing, we establish relationships between our measures and participants' training success and we also assess their suitability for predicting the final training success using only behavioral data of the first half of the training.

The paper shows that: i) users' eye gazes, head movements, and objects in focus are related to their training success, ii) the combination of these data sources increases the explainability of training success, i.e., the data sources are complementary, iii) using even less than 50% of the data (i.e., half of a training session) allows predicting users' training success significantly better than chance. Our findings provide first steps towards dynamic VTEs that are adjustable to trainees' behavior and learning progress.

2 RELATED WORK

Training in VR and Tracking Data: Evaluation of VTEs and mixed reality training environments in an authentic setting is rela-

*e-mail: gj@ethz.ch
†e-mail: johannes.schneider@uni.li
‡e-mail: joshua.handali@uni.li
§e-mail: valentin.holzwarth@uni.li
¶e-mail: hirtc@ethz.ch
‖e-mail: wolfgang.fuhl@uni-tuebingen.de
**e-mail: jan.vom.brocke@uni.li
††e-mail: kunz@ethz.ch

tively uncommon [9, 10], since such an evaluation is a costly, time-consuming, and a manual activity. Thus, automated evaluation has gained traction recently. In a VTE for troubleshooting surgical robots, Moore et al. [20] use tracking of head movements to train a binary machine learning classifier to predict high and low learning gains. Holzwarth et al. [11] use behavioral data to assess users' affective states during training in VR. Orlosky et al. [23] collect a variety of eye movement metrics to predict a trainee's knowledge obtained from a word recall task in VR. By combining pupil diameter, eye movement, and other metrics, their support vector machine can predict a trainee's knowledge with an accuracy of 62%.

Head Movement Analysis: Head movements are among the most important human behaviors in VR, as they indicate a user's motion within a VTE [31]. This applies in particular to high fidelity VR systems, where users are permitted to navigate freely (e.g. by real walking [22]). In a VR classroom setting, a user's increased head movement indicates higher levels of social anxiety [30]. For a surgery simulator, novice and professional surgeons' head accelerations significantly differ [28], since novices make more sudden and unnecessary movements than experts. This observation is supported by Moore et al. [20], who find that slower and abrupt movements were associated with low learning gains in their VTE.

Eye Tracking Analysis: Eye movements are significant indicators for learning and recalling vocabulary in VR [23]. Even more important is the possibility to use eye tracking to predict the learning progress in serious games [17], in VTEs [25], and in instructional videos [29]. A work in neurology shows that future decisions can be predicted from pupil behavior [5].

Combining Behavioral Data Sources in VR: Hu et al. [12] observe a linear correlation between gaze positions and the head rotation's angular velocities. They implement a real-time gaze position prediction model. Pfeuffer et al. [24] combine eye tracking and body motion as behavioral biometrics, and investigate which behavior is suitable to identify a user. This work is extended by Liebers et al. [19], who show that behavioral biometrics is also possible based solely on user movements. Mu et al. [21] analyze users' eye gaze positions and body movements in VR. They find strong indicators that some of the interpersonal differences in these two metrics are related to users' backgrounds such as personality and related skills.

3 Methodology

3.1 User Study

Participants: The user study included 48 male participants, as there were no female participants available at the vocational school where we conducted the study. All participants were sanitary trainees in their 1st to 3rd year of apprenticeship. Their mean age was 19.65 years with a standard deviation (Std) of 4.84 years. While 18 participants had no prior VR experience, 25 had experienced VR at least once but less than five hours, and five had more than five hours of prior VR experience. Further, six participants never played video games, 11 occasionally played video games (less than one hour on average per week), 19 regularly played video games (between one and seven hours per week), and 12 often played video games (more than seven hours per week). All had normal or corrected to normal vision. All participants had already learned the process of pressing pressfittings with small diameters as part of their education in the vocational school. However, the participants had no prior knowledge regarding the specific task that they trained for in the VTE. All of them were randomly assigned to either use a pressing tool emulated by hand-held controllers (31 participants), or the real pressing tool equipped with HTC Vive trackers 2.0 (17 participants).

Measures: The measures consisted of both self-reported questionnaires as well as behavioral data, which are shown in Table 1.

The behavioral data consisted of head movements, eye movements, and objects in focus. Data was recorded at a sampling rate of 40 Hertz using a Unity script, which is comparable to other literature [4]. Each sample consisted of the frame's timestamp, the rotation of the HMD, the combined eye gaze direction of both eyes, and the currently focused object. The rotation of the HMD was described by the rotation angles around the axes of the global coordinate system, shown in the lower right corner of Fig. 1. The eye gaze direction was described by a three-dimensional unit vector in the local reference frame of the HMD as provided by the HTC SRanipal SDK. Since eye gaze direction was described in the HMD's local reference frame, and thus head movements and eye gaze could be measured independently. The focused object was given by a string containing its name. The VTE was implemented such that only the five objects "Demo Collar", "Instruction Screen", "Pressing Collar on Pressfitting", "Pressing Tool", and "Buzzer" were recognized as focused. These five objects were selected because they are essential for learning the task at hand either by providing information on the task (e.g. the instruction screen), or by being an integral part of the task itself (e.g. the pressing tool). These five objects are marked in red in Fig. 1.

The self-reported measures consisted of standardized questionnaires, i.e. the Simulator Sickness Questionnaire (SSQ) [16], SUS Presence [27], and two items of the adapted Learner Satisfaction Questionnaire [15]. These two items were "Participants' satisfaction with the content of the training system" as well as "Participants' confidence that their answers in the knowledge test are correct". The assessment of training success was conducted through a knowledge test consisting of eight single-choice questions with three response options for each question (i.e. participants had to choose A, B, or C). To ensure external validity, the knowledge test was developed in collaboration with the pressing equipment manufacturer, as well as the division "Building Technologies" at the vocational school.

Study Procedure: The user study was conducted in single-user sessions. A session consisted of three phases and lasted 45 minutes on average. In the first phase, the participant was welcomed and informed about the study, and signed a consent form. He then completed a demographic questionnaire and the SSQ pre-questionnaire.

In the second phase, the participant was introduced to the VR system and the eye tracker was calibrated using the standard HTC Vive Pro Eye calibration procedure. The tutorial scene was started, familiarizing the participant with the VTE and how to interact with objects (e.g. grasping). The participant proceeded with the actual VTE, which will be more closely described in Sect. 3.1. Two videos containing the steps of the tutorial scene as well as the actual VTE can be found in the supplementary material. While undergoing training in the VTE, the participant's body movements, eye gaze and focused objects were recorded as described in Sect. 3.1. In the third phase, the participant was asked to fill out another set of questionnaires (SSQ post-questionnaire, SUS Presence, Learner Satisfaction). Finally, the knowledge test was administered.

Virtual Training Environment: The task used in the VTE of this study was the pressing of steel pressfittings using a hand-held pressing tool. The pressing procedure had to be conducted in a predefined sequence of four steps while utilizing special equipment. The equipment consisted of a pressing tool with a stage 1 adapter for pre-pressing collar, a stage 2 adapter for post-pressing collar, and a pressing collar. All objects and the pressing process were provided in the VTE (see Fig. 1). The VTE had an instruction screen where assignments and instructions were presented to the participant in written form, and a buzzer to be triggered in order to proceed to the next step of the training. Further objects were a "Demo Collar", which is an enlarged version of the pressing collar used to clarify procedural steps, and the pressfitting to be pressed.

Each VR training session starts with the participant standing in front of the instruction screen reading instructions followed by executing a work step and, finally, triggering the buzzer in order to proceed to the next step. The four steps of the task were as follows: 1) Positioning the pressing collar on the pressfitting segment to be

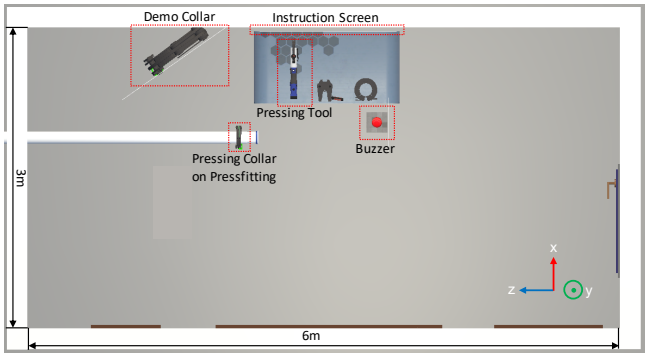

Figure 1: Overview on the VTE: Instruction screen, demo collar, buzzer, task specific components (pressing tool, stage 2 adapter for pressing collar, pressing collar), and pressfitting to be pressed.

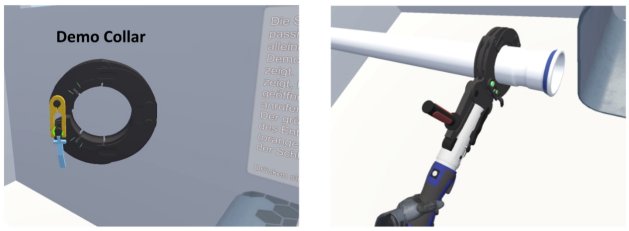

Figure 2: Objects in the VTE. Left: Enlarged version of the collar; Right: Pressing tool attached to pressing collar on pressfitting.

pressed.
2) Conducting the first stage pressing by hooking the pressing tool with the stage 1 adapter on the pressing collar and initiating the pressing process by either pushing the trackpad on the hand-held HTC VIVE controllers or the start button on the real tool.
3) Changing the pressing adapter on the pressing tool, i.e. mounting the stage 2 adapter. This step was animated since the sanitary trainees were already been trained in changing pressing adapters.
4) Conducting the second stage of pressing (similar to stage 1), which is also referred to as final pressing.

The training session ended with the participant pressing the buzzer after having read the final instructions on the instruction screen.

Apparatus: We used the HTC Vive Pro Eye with two SteamVR base stations 2.0, the HMD with the eye tracker, and the two controllers. A group of 31 participants was trained with controllers only, the remaining 17 participants were trained with Sensoryx VRfree gloves, and a real Geberit ACO 203plus pressing tool with two HTC Vive Trackers 2.0. Further, the SteamVR was set-up on a 6×3 meters space. The VTE was developed in Unity and the Tobii XR SDK was utilized to extract the eye tracking data. The VR computer had an Nvidia RTX 2080 GPU, an Intel i7-9700K CPU, and 32 GB RAM to ensure a frame rate of 90 Hertz on the HMD.

3.2 Analysis

The analysis has two goals: (i) To understand the relationship between behavioral data and performance and, (ii) to assess the practical utility of using behavioral data for predicting training success. To address the first goal, we extract meaningful characteristics from sensor data capturing bodily movements and investigate whether they exhibit a significant influence on training success. To this end, we employ classical statistical hypothesis testing, i.e. multiple linear regression. For the second goal, we apply machine learning models that are known for good predictive performance but are less suitable to prove the significance of an input variable to predict an output.

Statistical Analysis: To derive our features, i.e. input variables for the multiple linear regression, we transformed and aggregated the raw data (see Table 1). Then, we performed forward variable selection to improve interpretability by reducing the number of variables.
Raw Data: We consider three data sources as described in Sect. 3.1, i.e. head movement (rotation), eye gaze, and objects in focus (OF).
Transformation: Absolute values, such as angles of the head rotation, are strongly dependent on the experimental setup [20]. To be more general, we computed our features from the magnitude of differences of raw input values, i.e. we used $|d_i|$ with $d_i := value_i - value_{i+1}$. Absolute differences relate to change per time, i.e. velocity. We also used the magnitude of differences between differences, i.e. $|d_i^2| := |d_i - d_{i+1}|$, which corresponds to acceleration. For objects in focus, we computed the duration for each instance they were in focus. Features related to gaze duration reflect the importance of an object to the participant [2]. The transformed data consists of a sequence of values, where one value indicates the timespan in seconds the object was in focus.
Aggregation: The transformed data has a large number of values for each variable, e.g. thousands for each head movement variable. Thus, we aggregated further by computing the average (Avg) and standard deviation (Std) for each variable. This leads to interpretable, but interrelated measures. For example, features for different axes are correlated. To give another example, high average velocity might be seen as positive, since it indicates a person is moving quickly. However, it might also indicate that a person is moving hectically. If, in addition, the standard deviation is high, it indicates that a person is more likely to have periods of high velocity, e.g. fast transitioning between objects in focus, and periods of low velocity, e.g. focusing on a particular object while remaining still.
Variables and Model Selection: Applying the proposed aggregation measures on the transformed data yields a large number of variables, i.e. more than 30 that could be used in a linear regression model. However, given that we aim to predict only one value per participant, i.e. 48 values, utilizing 30 parameters might result in over-fitting. Interpretation is also difficult. Thus, we performed variable selections (Chapter 3 of [7]), i.e. we aim to find a subset of variables that yields a balance between model complexity and fit to the data. Following [7], we used Akaike's Information Criterion (AIC) for model selection. This criterion provides a trade-off between model simplicity and data fit. The model selection was done separately for each of the three data sources, resulting in the variables in Table 1.

Regression Analysis: Applying a linear regression analysis using the chosen variables (last row in Table 1) is done using Python's Statsmodels package. Variables are first standardized to make their regression coefficients more easily comparable. In total, we assessed four models, i.e. one model for the features of each behavioral data source (yielding three models), and one model using all features. We added two control variables; one taking into account whether trainees had previously been exposed to VR, and one taking into account whether a participant used a real pressing tool (with trackers mounted onto it) or only controllers. We investigated whether individual features bore a significant correlation with training success by investigating p-values. We assessed each model as a whole by discussing the adjusted R^2 and AIC. The former quantifies the explained variance of the training score taking into account model complexity, i.e. the number of parameters (one parameter per feature).

Prediction Task: We trained classifiers using the same features as for linear regression (see the last row in Table 1) to evaluate their utility in predicting a user's knowledge test performance. We employed Python's scikit-learn library. We divided our participants into two groups based on their training success using the median of the test scores. That is, participants who scored lower than the median score were put into the low performance group (LP) and those who scored equal to or higher than the median score were put into the high performance group (HP). The median value was taken

Table 1: Overview of raw data, its transformation, aggregation and resulting variables.

	Head Movement	Eye Gaze	Object in Focus
Raw Data	Angles H_x, H_y, H_z	Ray Direction D_x, D_y, D_z	Pressing Tool F_{pt}, Buzzer F_b, Pressing Collar on Pressfitting F_{pc}, Instruction Screen F_{is}, Demo Collar F_{dc}
Transformations	1st order differences to get velocity (v) 2nd order differences to get acceleration (a)		Duration (d)
Transformed Data	1st order diff.: H_x^v, H_y^v, H_z^v 2nd order diff.: H_x^a, H_y^a, H_z^a	D_x^v, D_y^v, D_z^v D_x^a, D_y^a, D_z^a	F_{pt}^d, F_{pc}^d, F_{dc}^d, F_b^d, F_{is}^d
Aggregation	Average (Avg), Standard deviation (Std)		
Variables for linear regression	$Avg(H_x^v), ..., Avg(H_z^a)$ $Std(H_x^v), ..., Std(H_z^a)$	$Avg(D_x^v), ..., Avg(D_z^a)$ $Std(D_x^v), ..., Std(D_z^a)$	$Avg(F_{pt}^d), ..., Avg(F_{dc}^d)$ $Std(F_{pt}^d), ..., Std(F_{dc}^d)$
Chosen variables based on AIC and forward selection	- $Std(H_z^v)$, $Std(H_z^a)$	$Avg(R_x^v)$, $Avg(R_y^a)$ $Std(R_x^v)$, $Std(R_z^a)$, $Std(R_x^a)$	$Avg(F_{pt}^d)$, $Avg(F_{pc}^d)$, $Avg(F_{dc}^d)$ -

Table 2: Descriptive results per group (mean and standard deviation).

Measure	Range	Controller	Real Tool	Combined
No. of Participants		31	17	48
Training Duration (minutes)		8.53± 2.08	9.87± 2.25	9.00± 2.18
Test Score	[0,8]	5.16± 2.03	5.76± 1.35	5.38± 1.81
SUS Presence	[0,8]	4.90± 0.89	4.66± 1.06	4.82± 0.94
Confidence	[1,7]	5.00± 1.29	4.94± 1.27	4.96± 1.28
Satisfaction	[1,7]	5.94± 0.86	6.29± 0.66	6.17± 0.81
ΔSSQ	[±235.62]	-2.90± 11.41	0.88± 10.87	-1.56± 11.26

to obtain groups with balanced sizes.

To evaluate the trained classifiers, we used the leave-one-out cross validation approach: We left one of n samples out, training a model with the remaining $n-1$ samples, and test the model using the left-out sample. This was done for each sample leading to n folds. The cross-validation accuracy is the average accuracy of all n folds.

Our focus is on predicting training success of the entire training session but using only parts of the entire training session, i.e. we use the first four minutes for each user. Since the average duration of a user is about nine minutes with a standard deviation of two minutes, this corresponds to using about half of all data for a user. The motivation is to assess the capability to predict final success to alter training, e.g. in case performance is likely to be non-satisfactory.

4 Results

4.1 Descriptive Results

We elaborate on the combined results of both conditions in terms of mean and standard deviation (Std). All results are shown in Table 2. The mean SUS presence score was 4.83 (Std=0.94) of 7, which can be considered reasonable. The mean knowledge test score was 5.38 (Std=1.81) out of 8, indicating that most participants acquired substantial knowledge. The overall effectiveness of the VTE is underpinned by the mean satisfaction with the training content's quality of 6.17 (Std=0.81) of 7 and a mean confidence that their answers would be correct of 4.96 (Std=1.28) of 7.

4.2 Statistical Analysis using Linear Regression

We first discuss the feature selection process as described in Table 3.2, i.e. the step to reduce the number of features from above 30 to 10 (see last two rows in Table 1).

Variable Selection: For head movements, features for x, y, and z have high correlations. Thus, only a single feature, i.e. z (corresponding to head pitch) is chosen. The mean velocity is not chosen, since moving fast on average is not a key factor to determine success.

Table 3: Regression models on knowledge test scores.

Features	Head Mov.	Eye Gaze	OF	Head + Eye + OF
Std(H_z^a)	-2.39*	-	-	-1.86
Std(H_z^v)	2.23*	-	-	2.35*
Avg(R_x^v)	-	-0.76	-	-0.30
Avg(R_y^a)	-	1.31*	-	0.74
Std(R_x^v)	-	3.87**	-	2.85*
Std(R_z^v)	-	-1.81**	-	-1.32*
Std(R_x^a)	-	-2.55*	-	-1.92
Avg(F_{pt}^d)	-	-	-0.60*	-0.61*
Avg(F_{pc}^d)	-	-	0.65*	0.49
Avg(F_{dc}^d)	-	-	0.57*	0.52*
Intercept	4.32***	4.20***	4.41***	4.29***
Real Tool	0.91	1.38*	1.27*	0.82
Used VR	1.17*	1.10*	0.83	1.27*
Metrics				
R^2	0.187	0.332	0.247	0.522
Adj. R^2	0.111	0.216	0.157	0.358
AIC	193.3	189.8	191.6	183.8

Fast movements can be ambiguous indicating both hectic, error-rich behavior as well as fast, focused behavior. For eye gaze, features correlate between 0.6 and 0.92. Thus, only five out of twelve features are chosen. The interpretation of standard deviation of a single variable is qualitatively the same as for head movements.

The average focus duration on objects plays a role, indicating that for some objects it could be good or bad to investigate them for only a short amount of time on average. The standard deviation is irrelevant, indicating that it does not matter how the total gaze time at the object is distributed. For example, it is irrelevant whether one stares at an object for a long time early in the study followed by short periods of focus, or always looks at it for the same duration.

Linear Regression: Results of the linear regression analysis are shown in Table 3. The models consisting of just a single behavioral data source perform similarly. Eye gaze performs best in terms of AIC. High standard deviation of velocity $Std(H_z^v)$ is an indicator for success. It is large if people alternate between two extreme states, e.g. standing still to focus on one object, and moving fast. Low $Std(H_z^a)$ is also an indicator of success. High $Std(H_z^v)$ and low $Std(H_z^a)$ can characterize a person transitioning between times of moving fast and moving fairly slowly with constant accelerations. That is, there is no complete standing still but also only a few rapid transitions from high velocity in one direction followed by the opposite direction. Hectic and frequent movements, i.e. rotating the head quickly up

and down in an abrupt manner, are also indicative of large (and short) accelerations. Since all users exhibit times of low accelerations, large accelerations tend to cause large $Std(H_z^a)$, which is indicative of lower training success. The interpretation of $Std(R_x^v)$ and $Std(R_x^a)$ is as for head movements. When it comes to looking up and down, it is preferable if people maintain their height or maintain an exploration at constant low velocity, as indicated by the negative coefficient of $Std(R_z^v)$. Sequences of fast moving eyes (up and down) followed by (short periods) of little movement can indicate uncertainty in terms of handling the tool, e.g. double-checking.

For objects in focus, studying properly the demo collar as well as the pressing collar correlates positively with training performance. The demo collar is not needed to solve the task, but it is an indicator whether exploring the non-mounted collar is beneficial to understand the problem. Spending prolonged time per interaction with the pressing tool is an indicator of poor performance. This is aligned with the actual task: The key challenge is mounting the tool on the collar, where the complexity is more in understanding the pressing collar on the pressfitting rather than the pressing tool. Therefore, the collar should be in focus more frequently and for a longer time. Short interaction times with the pressing collar, i.e. short times the object is in focus, typically indicate that the user did not actually conduct any steps needed to solve the task but conducted a short visual inspection during trouble shooting.

Finally, the model which uses all three behavioral data sources is shown to be the best of all four models as hinted by showing the largest adjusted R^2 value and the lowest AIC value. Thus, the more behavioral data sources are used, the better the model. Also, at least one predictor from each data source remains significant in the model with all data sources showing that data sources are complementary.

Table 4: Classifier performance for predicting knowledge test scores using features computed using the first four minutes of a user.

Features	Linear Regr.	Logistic Regr.	SVM
Head Movements	.562	.667	.562
Eye Gaze	.604	.604	.500
Object in Focus (OF)	.479	.500	.458
Head + Eye + OF	.625	.688	.646

Table 5: Correlation of features computed using the first four minutes and the resulting data.

Features	Correlation	Features	Correlation
Std(R_z^v)	0.795	Std(H_z^a)	0.627
Std(R_x^a)	0.764	Std(H_z^v)	0.662
Avg(R_x^v)	0.773	Avg(F_{pt}^d)	-0.104
Avg(R_y^a)	0.822	Avg(F_{pc}^d)	0.411
Std(R_x^v)	0.736	Avg(F_{dc}^d)	-0.299

The type of interface, i.e. whether people used hand-held controllers emulating the pressing tool or the real tool was significant for some models. The coefficient was also small, indicating a limited dependence on performance whether the real tool or emulation is used. Any variation in the training setup leading to more variance in sensor values and, in turn, of extracted features is non-desirable from an analytics perspective, since variance makes prediction more difficult. The fact that results are significant despite variations in the type of interface is an indication of the robustness of the relationship between our proposed features and training performance.

An exemplary comparison of behavioral data of high and low performing users is depicted in Fig. 3. It shows the duration for each time one of the three considered objects of the VTE was in focus. It can be seen that the average duration differs between them for each object in alignment with our regression analysis. A high performance participant shows shorter focus times for the pressing tool, but longer for the two objects than a low performance participant.

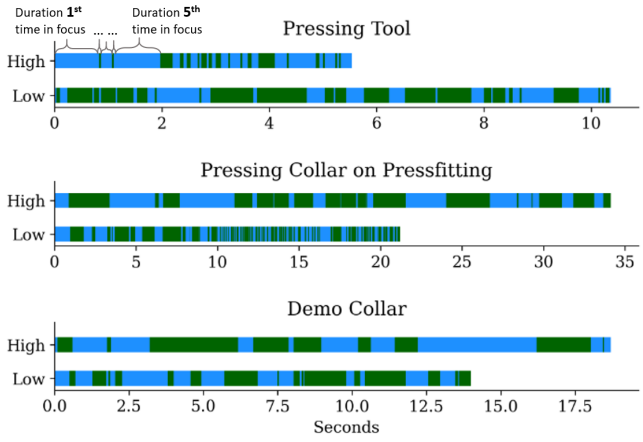

Figure 3: Cumulative focus duration for objects for high and low performing participants. A bar segment is an occurrence of focusing on an object. The length of the i-th bar segment shows the duration of the i-th time an object was in focus.

4.3 Prediction

In a study with 48 participants, each participant accounts for about 2.1%. Thus, we consider models that differ by two or fewer users, i.e. about 4.2%, as not clearly distinct. Prediction results for three different models are shown in Table 4. Overall logistic regression performs best, outperforming the baseline of 50%. Aligned with the linear regression analysis, more information leads to better results, i.e. the model with all three data sources performs best. It seems that objects in focus are not helpful in predicting training success. This is expected, since these objects are mostly unused in the first four minutes of the experiment, where workers focus on reading instructions. Thus, the interaction with those objects is initially nonexistent, which is different in the latter part of the experiment. More generally, the question arises if the performance metrics are stable over time, i.e. if the correlation between the values for the first four minutes and the rest of the experiment is high (see Table 5). For objects in focus, features computed using only the first four minutes of all data are poorly correlated with those for the remaining data. In fact, if only all data (except the first four minutes, where interaction does not take place anyway) is taken into account, predictive accuracy for all three methods is above 60%. The results in Table 5 reveal that correlations for other measures are high, explaining why these variables are suitable to predict the overall success given the first four minutes.

5 Discussion and Limitations

In this paper, we propose interpretable metrics and techniques (linear regression) as well as predictive models. While interpretability is commonly at odds with performance in machine learning, we believe that interpretable metrics and statistical analysis are favorable to contribute to scientific knowledge aligned with [11, 23].

Our predictive results outperform the baseline of 50% by 10-20%. This is comparable to other methods [11,20,23], though comparisons are difficult since tasks and VTEs differ. Those applications that require higher accuracy might still benefit from the current approach by exploiting the recall-precision trade-off. That is, decisions, where the model is highly uncertain (based on decision confidence scores produced by many ML models) are deferred to humans. In practice, when designing a dynamic VTE adjusting difficulty based on user skills, this might mean that the VTE is adapted for users that are

very strongly under- or overperforming. Better accuracy is achieved by increasing data, i.e. having more participants, or techniques such as multi-task learning leveraging data from multiple experiments as well as further increasing data sources. More data might make other approaches like deep learning working on raw data more suitable. More participants and higher sampling frequency likely also allows to distinguish more diverse behaviors, e.g. see [1] for eye gaze.

It is preferable to balance the number of users in both conditions, which would allow for a statistical analysis of differences between both conditions. On the positive side, our models could predict performance despite the additional variance stemming from these two conditions. Another limitation comes from the authentic setting the study was conducted in. Although the training task was not part of the participants' curriculum, our results might also still be influenced by the expertise reversal effect [14]. Furthermore, the population used in this study is male and young. An additional study covering more age groups and both genders would help to generalize our statements. Objects in focus provide means of analysis, but are highly case-specific. They also require some effort to implement. In contrast, head movement and eye gaze information are easier to obtain and metrics are easy to compute. However, their interpretation also shows case dependencies, though we conjecture that our findings generalize to similar settings, i.e. where careful and focused interaction with various objects is required. Furthermore, since the metrics such as velocity and acceleration are well interpretable, analyzing them for different contexts should not be a major concern though it would be preferable to identify general behavior that is always associated with a desired property such as training success.

6 Conclusion and Future Work

We show that users' head movements, eye gaze and objects in focus are related to training success in VTEs and that the combination of these behavioral data sources is useful. Furthermore, we predict users' training success already after 4 minutes of training, outperforming the baseline by 10-20%. This performance could already be sufficient for creating a system, which decides in real-time, if a user receives an extended or shortened training session. Thus, future work will focus on the implementation and evaluation of such a system. Additionally, we will focus on strengthening the prediction by considering additional sources of behavioral data (e.g. pupillometry), larger user studies including more balanced samples and improving our model, e.g. by partitioning data into segments [20].

Acknowledgments

This work was partly funded by the Innosuisse 43670.1 INNO-ICT project. Further, the authors express their gratitude to Müller Wüst AG, Geberit AG, and Berufsschule Lenzburg for their support.

References

[1] R. Andersson, M. Nyström, and K. Holmqvist. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research 3(3)*, 2010.

[2] P. Bera, P. Soffer, and J. Parsons. Using eye tracking to expose cognitive processes in understanding conceptual models. *MIS Quarterly*, 43(4):1105–1126, 2019.

[3] F. Buttussi and L. Chittaro. Effects of Different Types of Virtual Reality Display on Presence and Learning in a Safety Training Scenario. *IEEE Trans. Vis. Comput. Graph.*, 24(2):1063–1076, 2018.

[4] X. Corbillon, F. De Simone, and G. Simon. 360-degree video head movement dataset. In *Proc. Multimed. Syst. Conf.*, 2017.

[5] M. El Haj and A. A. Moustafa. Pupil dilation as an indicator of future thinking. *Neurological Sci.*, 42(2):647–653, 2021.

[6] J. Gisler, C. Hirt, A. Kunz, and V. Holzwarth. Designing virtual training environments: Does immersion increase task performance? In *2020 Int. Conf. Cyberworlds (CW)*, pages 125–128.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. 2009.

[8] C. Hirt, M. Eckard, and A. Kunz. Stress generation and non-intrusive measurement in virtual environments using eye tracking. *J. Ambient Intell. Human. Comput.*, pages 1–13, 2020.

[9] C. Hirt, V. Holzwarth, J. Gisler, J. Schneider, and A. Kunz. Virtual learning environment for an industrial assembly task. In *2019 9th Int. Conf. Consum. Electron. (ICCE-Berlin)*, pages 337–342.

[10] V. Holzwarth, S. Steiner, J. Schneider, J. vom Brocke, and A. Kunz. Bim-enabled issue and progress tracking services using mixed reality. In *Smart Services Summit*, pages 49–58. 2021.

[11] V. Holzwarth, et al. Towards estimating affective states in virtual reality based on behavioral data. *Virtual Reality*, pages 1–14, 2021.

[12] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Trans. Vis. Comput. Graph.*, 25(5):2002–2010, 2019.

[13] R. Islam, et al. Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals. In *2020 IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*.

[14] S. Kalyuga. Expertise reversal effect and its implications for learner-tailored instruction. *Educ. Psychol. Rev.*, 19(4):509–539, 2007.

[15] H. Kampling. Feeling presence in immersive virtual reality for individual learning. In *Proc. Int. Conf. Inf. Syst. (ICIS)*.

[16] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *Int. J. Aviation Psychol.*, 3(3):203–220, 1993.

[17] A. B. Khedher and C. Frasson. Predicting user learning performance from eye movements during interaction with a serious game. In *Proc. EdMedia + Innovate Learn. 2016*, pages 1504–1511.

[18] G. I. Lee and M. R. Lee. Can a virtual reality surgical simulation training provide a self-driven and mentor-free skills learning? *Surgical Endoscopy*, 32(1):62–72, 2018.

[19] J. Liebers, et al. Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization. In *Proc. 2021 CHI Conf. Human Factors Comput. Syst.*, pages 1–11.

[20] A. G. Moore, R. P. McMahan, H. Dong, and N. Ruozzi. Extracting Velocity-Based User-Tracking Features to Predict Learning Gains in a Virtual Reality Training Application. In *2020 IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)*, pages 694–703.

[21] M. Mu, M. Dohan, A. Goodyear, G. Hill, C. Johns, and A. Mauthe. User attention and behaviour in virtual reality art encounter, 2020.

[22] T. Nescher and A. Kunz. Using head tracking data for robust short term path prediction of human locomotion. *Transactions on Computational Science XVIII*, pages 172–191, 2013.

[23] J. Orlosky, B. Huynh, and T. Hollerer. Using eye tracked virtual reality to classify understanding of vocabulary in recall tasks. In *Proc. Int. Conf. Artificial Intelligence and Virtual Reality (AIVR)*, pages 66–73.

[24] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proc. 2019 CHI Conf. Human Factors Comput. Syst.*, page 1–12.

[25] N. A. Rappa, S. Ledger, T. Teo, K. W. Wong, B. Power, and B. Hilliard. The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review. *Interactive Learn. Environ.*, pages 1–13, 2019.

[26] J. Schneider and J. P. Handali. Personalized explanation for machine learning: a conceptualization. In *Eur. Conf. Inf. Syst. (ECIS)*, 2019.

[27] M. Slater, M. Usoh, and A. Steed. Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 1994.

[28] S. Viriyasiripong, et al. Accelerometer measurement of head movement during laparoscopic surgery as a tool to evaluate skill development of surgeons. *J. Surgical Educ.*, 73(4):589–594, 2016.

[29] X. Wang, L. Lin, M. Han, and J. M. Spector. Impacts of cues on learning: Using eye-tracking technologies to examine the functions and designs of added cues in short instructional videos. *Comput. Human Behav.*, 107:106279, 2020.

[30] A. S. Won, B. Perone, M. Friend, and J. N. Bailenson. Identifying anxiety through tracked head movements in a virtual classroom. *Cyberpsychology, Behav, and Social Netw.*, 19(6):380–387, 2016.

[31] C. Wu, Z. Tan, Z. Wang, and S. Yang. A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. In *Proc. 8th ACM Multimed. Syst. Conf.*, pages 193–198, 2017.