Virtual Reality, Eye Tracking and Machine Learning: Analysis of Learning Outcomes in Off-the-Shelve VR-Software

Johannes Tümler [1], Juan Enrique Erazo Sanchez[1] and Christian Hänig [1]

Abstract: The combination of Virtual Reality (VR) and eye tracking allows to analyze how students use the presented VR content for learning. Here, we propose a novel approach to analyze eye tracking data in VR, even if no access to the VR software source code is given. This proof-of-concept leverages image classification methods to identify objects that captured the students' attention in VR. The method allows analysis of individual learning strategies and correlate those to individual learning outcomes.

Keywords: Virtual Reality, Education, Eye Tracking, Machine Learning, Image Classification

1 Introduction and Goals

Virtual Reality allows users to immerse in a completely computer-generated interactive experience in real-time. This makes it an ideal technology to let students learn and explore in a safe, controllable, reproducible environment. Because of its attractiveness for learning and teaching, there is a lot of research to apply VR in education. Especially important are studies of the effectiveness of VR implementations and as well as questions on how to embed such systems into regular teaching processes.

Eye tracking is a technology to *find spatial focus points of user's vision over time*. The combination of virtual reality and eye tracking consequently offers benefits to understand where users focus their view in a virtual environment. Clay et al. give insight on basic principles, limitations and application examples for eye tracking in VR [CKK19]. Rappa et al. analyzed papers that incorporated eye tracking in VR learning [Ra22]. Today, several VR headsets exist, that have included eye tracking capabilities. The best approach to carry out eye tracking studies in VR would be to embed it into the VR software source code for maximum flexibility. Unfortunately, most of the commercial VR learning software does neither offer eye tracking integration nor provide software source code. Therefore, an alternative approach would be helpful for eye tracking analysis in *off-the-shelve software*.

[1] Anhalt University of Applied Sciences, Bernburger Straße 55, 06366 Köthen, Germany,
Ingenieurinformatik, johannes.tuemler@hs-anhalt.de, https://orcid.org/0000-0002-4788-2667
Master of Data Science student programme, juanenrique.erazosanchez@student.hs-anhalt.de
Artificial Intelligence, christian.haenig@hs-anhalt.de, https://orcid.org/0000-0001-7775-4386

doi: 10.18420/delfi2023-32

Here, we present a novel approach to analyze eye tracking in off-the-shelve VR software. The goal is to use machine learning to identify VR scene objects that the users have looked at. This allows teachers or researchers to analyze their student's gaze behavior in a post-intervention step. The research was conducted in form of a student project at Anhalt University of Applied Sciences in Köthen, Germany.

2 Data Acquisition

The data used here was recorded in a previous study [IT22]: Having biomedical curricula at our university, we used a biomedical software: ShareCare YOU VR[2] is a VR simulation to teach and learn the human body, anatomical structures, organs, and their functions (Figure 1). In consent with N=20 participants we took video recordings from a Varjo VR-2 headset during a thirty-minute virtual learning experience.

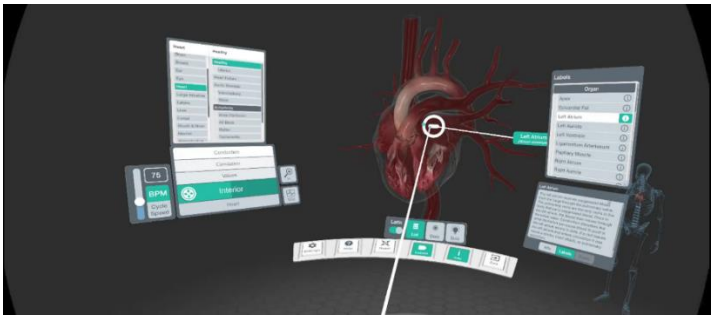

Figure 1: Screenshot of user's view in ShareCare YOU VR software. Billboards show a menu and additional text on the left and right side. A dock gives access to features at the bottom. The currently activated 3D object is visible in the center (here: human heart).

The study incorporated both a pre- and a post-intervention quiz to assess the participants' knowledge of human heart anatomy before and after using the VR learning software. Looking at the gained score percentage, exactly those with high/low starting knowledge scored a low/high absolute gain in score, corresponding to outcomes of Zinchenko et al. [Zi20]. Unfortunately, due to loss of data, not all the 20 participants' data sets were still available for our new analysis. Therefore, in the following approach we used only ten data sets (IDs 4, 5, 6, 7, 8, 9, 10, 11 ,17, 18) to provide a proof-of-concept.

[2] https://store.steampowered.com/app/724590/Sharecare_YOU_VR/

3 Method and Implementation

Data annotation provides the ground truth labels necessary for training and evaluation of machine learning models. In our study, we used the annotation tool makesense.ai to annotate the images in our dataset. From the existing video recordings, at first, we extracted each video frame as image and determined the user's 2D focus point using the existing eye tracking method. Second, we reduced the size of each image to the area around the user's focus point. The annotator has chosen a receptive field of 400x400 pixels, which provides enough visual information for the classification of the target classes. We identified six relevant classes based on the content of the videos: The virtual depiction of the human heart, three display types (left/right/main), the dock allowing users to control the simulation and the "booting screen" from before application start.

Each training image was assigned exactly one class, which was determined by the human annotator as the most relevant class. In total, we manually annotated 1978 images, resulting in a reasonably balanced dataset (see Table 1). While there is some variation in the number of samples per class, there are no classes that are significantly over- or under-represented (as in 1:100, 1:1000 or 1:10000 [CJK04]).

Class	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Total
Boot Screen	63	108	68	29	23	29	320
Left Display	24	70	84	22	36	14	250
Right Display	119	161	103	72	179	58	692
Dock	5	23	4	3	27	16	78
Main Display	19	63	3	19	17	46	167
Heart	87	98	56	90	68	72	471
							1978

Table 1: Frequency of target classes on a per-video (user) level and in total.

To classify the images in our dataset, we experimented with ResNet models [He16] of different sizes. We based our choice on previous work in the field of computer vision, which has shown that ResNet architectures are parameter-efficient for comparable image classification tasks [Do20]. In all conducted experiments, we initialized our models with weights been pre-trained on the ImageNet1k dataset [Ru15] to benefit from robust features trained on a large dataset.

To train our ResNet models, we split our annotated dataset of 1978 images into train, validation, and test sets. Since our data comes from videos, we used a leave-one-out approach on video level for cross-validation [HTF09]. For each experiment, the training set was built from four videos, the validation and test set from one video each. We have chosen this approach over the well-established stratified n-fold cross-validation [ZM00] because this approach ensured that no frames from the same video were present in both the training and test sets. Temporally consecutive frames from the same video are very similar to each other, which would lead to an overestimation of our model's generalization

ability if frames from the same video would be distributed among train and test sets. In total, we conducted 30 independent runs for each of the ResNet variants (18, 34, 50, 101, and 152 layers) and evaluated their performance on the held-out test set. After training and validating our ResNet models on the annotated dataset, we evaluated their performance using Precision, Recall, and F1-Score metrics [CH92] (Table 2). Based on our experiments, the ResNet34 model achieved the highest mean F1-Score of 0.9663, closely followed by the other models.

Model	Parameters	Mean F-Score	Min F-Score	Max F-Score	σ
ResNet18	11M	0.9638	0.8979	0.9830	0.0227
ResNet34	21M	**0.9663**	**0.9277**	**0.9906**	0.0207
ResNet50	23M	0.9635	0.8936	0.9843	0.0291
ResNet101	42M	0.9559	0.8809	0.9872	0.0295
ResNet152	58M	0.9620	0.8979	0.9872	**0.0200**

Table 2: Resulting F-Scores of all ResNet models (evaluated on test sets).

To gain a deeper understanding of the performance of our models, we analyzed misclassifications. Most were due to the presence of multiple objects in a single frame. In these cases, the model correctly predicted one of the visible objects, while the human annotator defined another object as the dominant and thus, correct one.

4 ML-based Analysis of the Previous Study Data

From previous experiments we able to assume that users focus points would not erratically jump back and forth between the target classes multiple times per second. Therefore, we extracted one frame per second from the VR eye tracking data to obtain a representative sample of participants' visual attention during their VR experience. Subsequently, we applied our trained ResNet34 model to classify the extracted frames, enabling us to identify the participants' focus objects among the object classes. We computed Pearson correlation coefficients between knowledge gain, the relative and absolute time spent on various classes (Table 3). For that, we excluded participants #4 and #17, because these had either a large or a very low previous knowledge.

Object Class	left_display	right_display	dock	main_display	heart
Average Viewing Time	14.4%	37.1%	6.6%	9.9%	32.0%
Correlation Coefficient	-0.72	-0.46	-0.02	0.39	0.72
Avg. Sec. per Participant	43	110	20	28	91
Correlation Coefficient	-0.30	-0.67	-0.40	-0.11	0.19

Table 3: Correlations between relative viewing time and gained knowledge (first two rows), correlations between absolute viewing time and gained knowledge (last two rows)

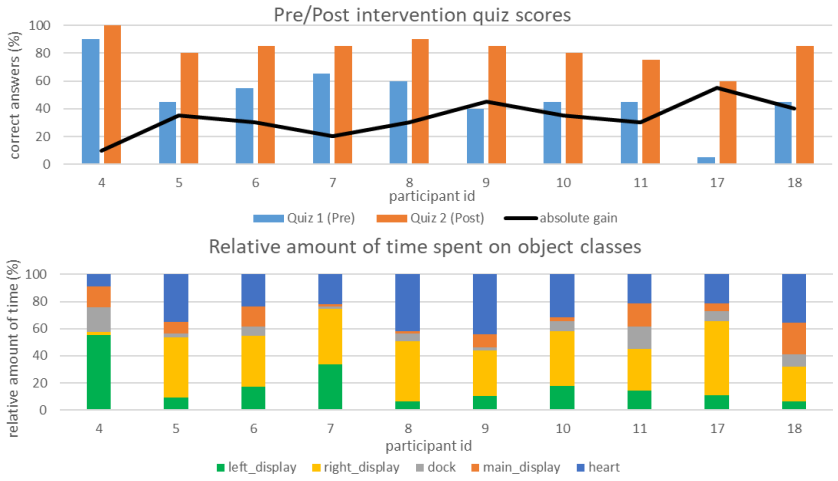

Figure 2: Gained knowledge (top), relative amount of time spent with a certain object class per participant, retrieved using the novel machine learning-based method (bottom).

Figure 2 illustrates the distribution of attention across the five main object classes. The learning experience of the participants and the factors influencing their learning outcomes were analyzed, considering various aspects such as individual pre-existing knowledge, the use of different displays, and overall learning duration. For most participants, the left display exhibited a moderate to strong negative correlation with learning outcomes. This suggests that the left display may not have effectively conveyed information. Interactions with the 3D VR heart model positively correlated with learning outcomes, ranging from mediocre to strong associations. This implies that focusing on the key object of the virtual environment was beneficial for understanding of the subject. The absolute duration of learning had, at most, a weak influence on learning outcomes. This result emphasizes that the quality of the learning experience and engagement with relevant objects may be more crucial than the time spent in the virtual environment.

5 Summary and Discussion

This study has two main contributions: (1) A machine learning model was developed and successfully applied to identify specific gazed-at regions in an off-the-shelve VR software. (2) The result of the machine learning analysis was used to examine possible correlations between learner's viewing behaviour and learning outcome.

We aimed at a proof-of-concept for the method and therefore used a relatively low number of data sets and object classes. We found correlations between the knowledge gained and time spent on specific VR objects. Future research should aim to replicate these findings with a larger and more diverse participant pool to better understand factors influencing

learning outcomes. The data used to create the image classification model was taken from real study videos. Therefore, the quality of the machine learning training data was not optimal. For future analyses, we suggest to first generate training data from a controlled dataset, for example by recording in-app videos with only known objects in sight and with controlled headset movements. Following our ideas, future research based on eye tracking data in off-the-shelve VR software may become easier.

Bibliography

[Ch92] Chinchor, N.: MUC-4 Evaluation Metrics. In: Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992.

[CJK04] Nitesh, V. C.; Nathalie, J.; Aleksander, K.: Editorial: special issue on learning from imbalanced data sets. Sigkdd Explorations, 6(1), 1-6, 2004.

[CKK19] Clay, V.; König, P.; König, S.: Eye tracking in virtual reality. Journal of Eye Movement Research, vol. 12, no. 1, 2019. doi: 10.16910/JEMR.12.1.3

[Do20] Dosovitskiy, A. et al.: An image is worth 16x16 words: Transformers for image recognition at scale. 2020. doi: 10.48550/arXiv.2010.11929

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[HTF09] Hastie, T.; Tibshirani, R.; Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition. Springer, New York, 2009. doi: 10.1007/978-0-387-84858-7

[IT22] Igbudu, M.; Tümler, J.: Investigating Eye-Tracking in 3rd Party Off-the-Shelve Software. Proceedings of DELFI Workshops 2022, pp. 47-55. Gesellschaft für Informatik e.V., 2022. doi: 10.18420/delfi2022-ws-14

[Ra22] Rappa, N.A. et al.: The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review. Interactive Learning Environments, vol. 30, no. 7, pp. 1338–1350, 2022. doi: 10.1080/10494820.2019.1702560

[Ru15] Russakovsky, O. et al.: Imagenet large scale visual recognition challenge. International journal of computer vision, 115, 211-252, 2015. doi: 10.1007/s11263-015-0816-y

[Zi20] Zinchenko, Y.P.; Khoroshikh, P.P.; Sergievich, A.A.; Smirnov, A.S.; Tumyalis, A.V.; Kovalev, A.I.; Gutnikov, S.A.; Golokhvast, K.S.: Virtual reality is more efficient in learning human heart anatomy especially for subjects with low baseline knowledge. New Ideas Psychol., vol. 59, 2020. doi: 10.1016/j.newideapsych.2020.100786

[ZM00] Zeng, X.; Martinez, T. R.: Distribution-balanced stratified cross-validation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1), 1-12, 2000. doi: 10.1080/095281300146272