International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2022)
J.-M. Normand and H. Uchiyama (Editors)

Towards Improving Educational Virtual Reality by Classifying Distraction using Deep Learning

Adil Khokhar and Christoph W. Borst

University of Louisiana at Lafayette

Abstract

Distractions can cause students to miss out on critical information in educational Virtual Reality (VR) environments. Our work uses generalized features (angular velocities, positional velocities, pupil diameter, and eye openness) extracted from VR headset sensor data (head-tracking, hand-tracking, and eye-tracking) to train a deep CNN-LSTM classifier to detect distractors in our educational VR environment. We present preliminary results demonstrating a 94.93% accuracy for our classifier, an improvement in both the accuracy and generality of features used over two recent approaches. We believe that our work can be used to improve educational VR by providing a more accurate and generalizable approach for distractor detection.

CCS Concepts

• ***Computing methodologies*** → ***Machine learning;*** • ***Human-centered computing*** → ***Virtual reality;***

1. Introduction and Related Works

Consumer-level VR displays with varied sensing (head-tracking, eye-tracking, facial-tracking, and hand-tracking) such as the Meta Quest Pro are more common and affordable than ever, making it appealing for educational Virtual Reality (VR). Much work has evaluated what can be revealed about a person's physio-cognitive state through sensor data. Notable examples include measuring cognitive load with eye tracking [FNC19], disengagement with eye tracking [SDSB*16], emotional engagement with electrodermal activity [VCR*18], attention with eye tracking and facial thermal imaging [AKN*19], stress levels with cardiovascular activity [MHGP16], and attention with facial tracking and eye tracking [BSN*20]. Other works have trained classifiers on eye tracker data to detect distracted driving (SVM, 81.1% accuracy) [LJB13], lapses in focus (SVM, 80.6% accuracy) [YKM12], and student attention (XGBoost, 77% accuracy) [VDA*19]. Much research in this area focuses on using 2D desktop interfaces or webcams. External sensors for these can either require a user's face be visible, cannot be worn together with a headset, or impose restrictions on user movements to avoid noisy data [JWM*12], and can make it difficult to use them or existing datasets for VR studies; however, current consumer-level VR headsets have integrated sensors that VR researchers have used to detect emotion [XEAZ*21], identify users [AKB22], authenticate users [LNS*20], predict learning gains [MMDR20], predict user interaction [DJPZ*21], and predict cybersickness [IDQ21]. Of these, the study on cybersickness showed that deep learning can offer modest improvements to accuracy [IDQ21] compared to traditional machine learning methods.

Limited research has been done on classifying distraction in educational VR. Moore et al. [MMDR20] trained an SVM classifier with an accuracy of 86.7% to predict student performance levels in a VR training environment. Asish et al. [AHKB21] trained a CNN-LSTM classifier with an accuracy of 87.2% for detecting distraction in educational VR based on raw gaze and motion metrics. Limitations included overall detection per session rather than narrower per-event detection, learning with timestamps across these multi-event sessions with fixed distractor timing, absolute pupil positions, and upsampling by repeating data to balance class distribution. Training classifiers on such information can cause them to learn user-related or task-related features, such as fit position of a headset to a head shape or timing of interactions, which results in poor generalizability due to the model overfitting [KJ13].

Our work differs from prior approaches by studying event-based distractors that involve acknowledgement, extracting generalized features from gaze position, and improving on the prior deep learning classifier. We consider a deep learning approach to predict distractor presence based on VR headset sensor (head, hand, and eye) data collected during an educational VR experience. Students took a field trip to a virtual oil rig to learn about devices used in drilling operations. We collected and labelled data from 37 participants to train a deep learning model that detects distractors. Our work makes the following contributions:

1. Our approach uses generalized features extracted from VR headset sensor data (head, hand, and eye tracking) to train a deep CNN-LSTM classifier to detect distractors in an educational Virtual Reality (VR) environment.
2. Preliminary results that demonstrate a 94.93% accuracy for our classifier.

DOI: 10.2312/egve.20221279

Figure 1: *Left: A student looks at the correct object the teacher is pointing at (a gaze trail is shown here but is not visible to the student). Right: The teacher points and a student's gaze drifts towards the wrong object due to a distraction prompt on their in-game mobile device.*

2. Environment

Virtual oil rig platforms have been used for training workers. [SDA*16]. In our VR oil rig environment, students learn about equipment on a virtual oil rig. Our pre-recorded teacher is a 3D RGBD-based avatar composed from prerecorded color and depth videos captured by a Kinect V2 at 30 FPS, similar to the implementation from Borst et al. [EBWC16, KYB19]. This teacher agent points out and explains equipment used on oil rigs while the student holds a virtual mobile device for general use and interactions.

3. Methods

3.1. Experiment Design

To support our experiment on distractors, the virtual mobile device provides a way to experimentally simulate distractions in a controlled manner, without easily being ignored, and with a standard interaction tool that fits the theme. At critical moments during our educational presentation while the teacher points out and explains a device, the virtual mobile device presents a distractor by showing a text message with accompanying vibration and sound effect, as shown in Figure 1. Distractor occurrence was randomized by only occurring in a randomized subset of teacher clips so as to reduce student anticipation of distractors. We had two distractor levels:

- **No Distractor:** The teacher avatar points to a device and asks the student to look. No distractor is presented.
- **Distractor:** One to two seconds before the teacher avatar points at a device and asks the student to look, text on the mobile device prompts the student to interact with an object. The object is highlighted and the student needs to point and click on it.

3.2. Subjects and Apparatus

Subjects were 31 male and 6 female students from a Computer Science department, for 37 total subjects, with 28 undergraduate and 9 graduate students. Ages were from 18 to 40 years (median of 21). 9 subjects indicated prior experience of a VR field trip. 10 subjects indicated substantial prior VR experience in the form of owning a VR headset. In addition, 1 subject indicated knowledge of devices used on the oil rig. The apparatus for the experiment included a Vive Pro Eye headset, a Vive wand, a logitech R400 clicker, a large Samsung TV, and a desktop with an Intel Core i9 10900K CPU processor, GeForce 2080 graphics card, and 64GB of memory.

3.3. Data Collection Procedure

Subjects were given an overview before donning a VR headset. The proctor assisted the subject with fitting the headset as needed and confirmed that subjects saw clearly and were comfortable before they were handed the Vive Wand Controller and lowered the headset's headphones. For eye tracker calibration, the subject adjusted the IPD using a calibration indicator and looked at 5 calibration dots. The system calculated the offsets between each recorded gaze point for each calibration dot and made appropriate adjustments. If the offset was too large then the calibration process was repeated. The proctor confirmed the resulting accuracy using a particle trail that visualized tracked eye gaze ray while the subject looked at spheres arranged in a grid. Subsequently, the particle trail was disabled and the student received a tutorial by the pre-recorded teacher avatar that introduced the oil rig and in-game controls.

After the tutorial, the subject experienced the main educational presentation wherein a prerecorded teacher pointed out devices used on the virtual oil rig and explained how they work. The presentation included 6 areas. In each area, the teacher agent pointed out and explained three devices. Distractions were timed to occur prior to the pointing and explanation of a device. Only one of those times involved a distractor occurring. The order of which device the distractor occurred at was randomized for each area to reduce student anticipation of a distraction every time the teacher pointed. When a distractor occurred, the student was expected to acknowledge it

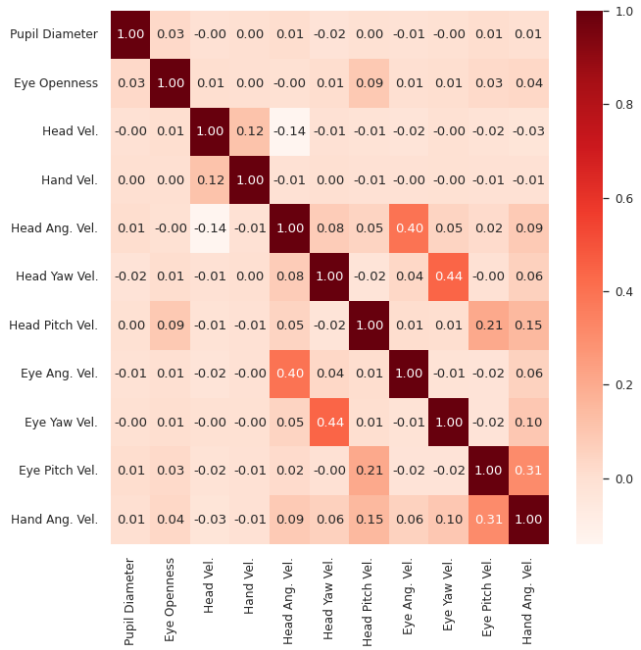

Figure 2: *Correlation matrix with a heatmap indicating which features are related to each other.*

by completing the associated point-and-click task. If the student ignored the distractor, then, at the end of the area, the teacher paused and reminded the student to address outstanding prompts, remaining paused until the student acknowledged the distractor. Refer to Khokhar et. al [KB22] for more details.

We collected raw gaze data as reported from the Vive Pro Eye device at 120hz. This included pupil diameter, eye openness, eye wideness, 2d gaze position, a 3d gaze origin, a 3d gaze direction, and a bitmask. The bitmask consisted of 5-bits that indicated whether the 2d gaze origin, 3d gaze origin, 3d gaze direction, pupil diameter, and/or eye openness fields had an invalid reading for that frame. We computed the angle between the subject's eye gaze direction and the direction from the center of the subject's eye to the teacher, to any pointed-at objects, to the phone, and to any distraction objects. We also computed these angles using the subject's head gaze direction. Angles were calculated between two direction vectors as $\theta = 2 \cdot atan2(||u-v||, ||u+v||)$ [DJPZ*21]. Finally, we recorded the world-space positions of the phone, head, and hand in the virtual environment.

3.4. Ground Truth Construction

We considered two classes representing our conditions for classification: distractor and no-distractor. We took a 5 second window of data from the interval when the distractor was present when the teacher pointed. Some people in the distractor interval may have tried to ignore the distractor and not perform the required distractor task, and we still label them as being in the distractor condition. We also took a 5 second window of data from the interval where there was no distractor during teacher pointing and labelled them as having no distractor. Our dataset consisted of 230680 data points with an equal distribution per class.

Layer	Type	Output Shape	# Param	Dropout	Activation
1	Conv1D	(11, 32)	128	-	ReLU
2	Conv1D	(11, 64)	6208	-	ReLU
3	Conv1D	(11, 128)	41088	-	ReLU
4	MaxPool	(6, 128)	0	-	-
5	Dropout	(6, 128)	0	0.2	-
6	Conv1D	(6, 256)	229632	-	ReLU
7	Conv1D	(6, 512)	393728	-	ReLU
8	MaxPool	(3, 512)	0	-	-
9	LSTM	(128)	328192	-	ReLU
10	Dropout	(3, 512)	0	0.5	-
11	Flatten	1536	0	-	-
12	Dense	512	786944	-	ReLU
13	Dense	1024	525312	-	ReLU
14	Dense	8	8200	-	Softmax

Table 1: *Architecture for Deep CNN-LSTM. Layers 1, 2, and 7 had a kernel size of 3. Layer 3 had a kernel size of 5. Layer 6 had a kernel size of 7. Max pool layers are size 3 with stride of 2. The LSTM layer had 128 units.*

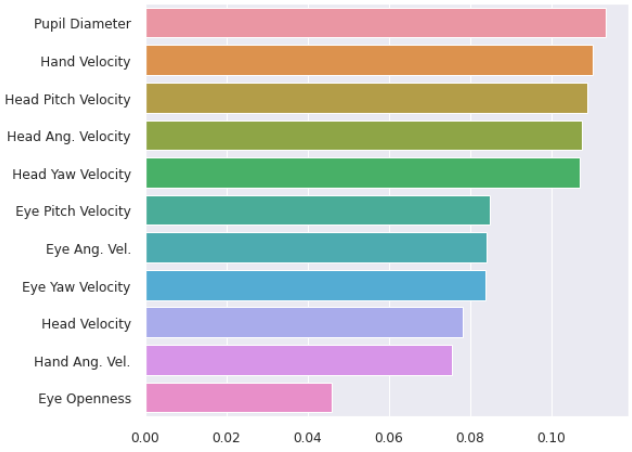

Figure 3: *Feature importance scores (x-axis) for selected features (y-axis). The score describes the decrease in mean accuracy with a random forest classifier when values for that feature are randomly shuffled.*

3.5. Data Pre-Processing

To narrow down features that were suitable for our dataset, we used the chi-squared test to compare fields between the two conditions. We then calculated the correlation matrix (Figure 2) and importance scores (Figure 3) for remaining features.

We avoided using gaze angles or differences directly because they may be less generalizable than features like movement velocities, for example, heights of positional data can be dependent on participant height, or include task-dependent information such

Name	Accuracy	Loss
CNN (Zhao et al. [ZLC*17])	77.73%	0.4562
CNN-LSTM (Asish et al. [AHKB21])	84.49%	0.4771
Deep CNN-LSTM (Ours)	**90.39%**	**0.2754**

Table 2: *Accuracy and loss of each model in classifying distractor presence.*

Name	Class	Precision	Recall	F1-Score
CNN	No Distr.	0.75	0.83	0.79
	Distractor	0.81	0.72	0.76
CNN-LSTM	No Distr.	0.83	0.86	0.85
	Distractor	0.86	0.83	0.84
Deep CNN-LSTM	**No Distr.**	**0.91**	**0.90**	**0.90**
	Distractor	**0.90**	**0.91**	**0.90**

Table 3: *Precision, recall, and F1-scores of each model in classifying distractor presence.*

as looking at a helicopter [KJ13, IDQ21]. Instead, we computed velocities for each angle as inter-frame differences per time step. We separated the 3d eye gaze and head gaze directions into yaw and pitch using the arctangent of a y component divided by the x component and the arctangent of a z component divided by the x component, respectively. Invalid values in fields indicated by the validation bitmask were replaced with linearly interpolated values based on the last valid and next valid values. A moving average filter reduced noisy data by taking the average of 6 neighboring data points for each field. Left and right eye openness values were averaged to produce a combined eye openness value. The left and right pupil diameter values were combined similarly. A base pupil diameter value was then computed by averaging the combined pupil diameter over each trial, the window of data that composed the lecture, pointing, and associated distractor for that area. We subtracted the pupil diameter value for that frame by the base pupil diameter to obtain a standardized distance metric for the pupil diameter. Finally, we normalized the data through standardization as follows:

$$D_n = \frac{D_n - D_{avg}}{std.dev.}$$

3.6. Deep Learning Classifiers

We considered three deep learning approaches for model classification: 1D CNN, CNN-LSTM, and our Deep CNN-LSTM. We used the Keras library to implement the models and the TensorFlow backend to train them. For all models we used a batch size of 512. We split our dataset into training (80%) and test (20%) sets, randomly choosing from 444 event windows (37 subjects X 6 areas X 2 conditions), with the training set used to train the model and the test set used to test classifier accuracy.

CNN: CNNs can classify raw time series data by extracting features from a sequence of observations [ZLC*17]. We used a convolution kernel size of 3 and 128 filter maps for the CNN layer.

CNN-LSTM: The CNN-LSTM model is a hybrid of a convolutional neural network (CNN) and a long short-term memory (LSTM) network. The CNN extracts spatial features from the input data and the LSTM extracts temporal features and performed well in prior work [AKB21].

Deep CNN-LSTM: This is our own architecture. See Table 1 for kernel size and filters for each CNN layer. The design of our CNN network is inspired from the VGG16 network [SZ15]; however, differing from it, we remove the 5th convolutional layer and change from 2d convolutions to 1d convolutions to classify time series data. We also add the LSTM network from the CNN-LSTM model to extract temporal features.

4. Results and Discussion

We evaluated the performance of our models using the F1 score, precision, and recall for each class. The F1 score is a measure of a test's accuracy. It is the harmonic mean of precision and recall. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Results are shown in Table 2 and Table 3.

This suggests that our deep CNN-LSTM model is able to capture the spatial and temporal features of gaze data better than LSTMs. Limitations with our approach included the fact that we only used a small subset of the data for training and testing, so we could not explore unsupervised or self-supervised methods. This limited us to classical unsupervised approaches, such as k-means, which perform poorly as distraction classifiers [AKB21]. Data were labelled based on the condition provided but this does not guarantee that all people were distracted or not distracted in those conditions. Further consideration of this during labelling could increase classifier accuracy. Our gaze angles did not include a notion of depth. Finally, we did not validate our model's generalizability on data from an offline environment.

5. Conclusion and Future Work

Our work presented an approach that improved the accuracy of and generality of features for training our deep CNN-LSTM classifier for detecting distractors in educational VR when compared to two recent approaches. Future work will consider validating the model's generalizability in an offline environment and will explore additional sensing such as EEG, ECG, etc. We will collect additional data so we can explore applying recent advances in deep learning approaches for anomaly detection, such as transformer-based methods with attention mechanisms. These aggregate and tokenize features into learnable embeddings that take advantage of an attention mechanism to improve learning on spatiotemporal features [VSP*17]. Variational auto-encoders are also used to detect anomalies and can learn a latent space representation of features to improve spatiotemporal learning [SBZ*22]. Adding modularity, by using only signals that a user opts to share, can address privacy concerns (e.g., machine unlearning [CZW*21]). This would work with gatekeeping modules that prevent detection of user properties by limiting features to those needed for an activity [DJHBJ21]. Finally, we would like to extend our work to include other types of distraction such as cognitive distraction and emotional distraction. We believe this will improve educational VR by providing a more accurate and robust distraction classifier.

References

[AHKB21] ASISH S. M., HOSSAIN E., KULSHRESHTH A. K., BORST C. W.: Deep Learning on Eye Gaze Data to Classify Student Distraction Level in an Educational VR Environment. *ICAT-EGVE 2021 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments* (2021), 10 pages. Artwork Size: 10 pages ISBN: 9783038681427 Publisher: The Eurographics Association Version Number: 037-046. URL: https://diglib.eg.org/handle/10.2312/egve20211326, doi:10.2312/EGVE.20211326. 1, 4

[AKB21] ASISH S. M., KULSHRESHTH A. K., BORST C. W.: Supervised vs Unsupervised Learning on Gaze Data to Classify Student Distraction Level in an Educational VR Environment. In *Symposium on Spatial User Interaction* (New York, NY, USA, Nov. 2021), SUI '21, Association for Computing Machinery, pp. 1–2. URL: https://doi.org/10.1145/3485279.3488283, doi:10.1145/3485279.3488283. 4

[AKB22] ASISH S. M., KULSHRESHTH A. K., BORST C. W.: User Identification Utilizing Minimal Eye-Gaze Features in Virtual Reality Applications. *Virtual Worlds 1*, 1 (Dec. 2022), 42–61. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. URL: https://www.mdpi.com/2813-2084/1/1/4, doi:10.3390/virtualworlds1010004. 1

[AKN*19] ABDELRAHMAN Y., KHAN A. A., NEWN J., VELLOSO E., SAFWAT S. A., BAILEY J., BULLING A., VETERE F., SCHMIDT A.: Classifying Attention Types with Thermal Imaging and Eye Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 3 (Sept. 2019), 1–27. URL: https://dl.acm.org/doi/10.1145/3351227, doi:10.1145/3351227. 1

[BSN*20] BABAEI E., SRIVASTAVA N., NEWN J., ZHOU Q., DINGLER T., VELLOSO E.: Faces of Focus: A Study on the Facial Cues of Attentional States. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, Apr. 2020), ACM, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3313831.3376566, doi:10.1145/3313831.3376566. 1

[CZW*21] CHEN M., ZHANG Z., WANG T., BACKES M., HUMBERT M., ZHANG Y.: When Machine Unlearning Jeopardizes Privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, Nov. 2021), CCS '21, Association for Computing Machinery, pp. 896–911. URL: https://doi.org/10.1145/3460120.3484756, doi:10.1145/3460120.3484756. 4

[DJHBJ21] DAVID-JOHN B., HOSFELT D., BUTLER K. R. B., JAIN E.: A privacy-preserving approach to streaming eye-tracking data. *undefined* (2021). URL: https://www.semanticscholar.org/paper/A-privacy-preserving-approach-to-streaming-data-David-John-Hosfelt/194a82df379e1182ddd88b94a30d147d920f2fab, doi:10.1109/TVCG.2021.3067787. 4

[DJPZ*21] DAVID-JOHN B., PEACOCK C., ZHANG T., MURDISON T. S., BENKO H., JONKER T. R.: Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications* (New York, NY, USA, May 2021), ETRA '21 Short Papers, Association for Computing Machinery, pp. 1–7. URL: https://doi.org/10.1145/3448018.3458008, doi:10.1145/3448018.3458008. 1, 3

[EBWC16] EKONG S., BORST C. W., WOODWORTH J., CHAMBERS T. L.: Teacher-Student VR Telepresence with Networked Depth Camera Mesh and Heterogeneous Displays. In *Advances in Visual Computing* (Cham, 2016), Bebis G., Boyle R., Parvin B., Koracin D., Porikli F., Skaff S., Entezari A., Min J., Iwai D., Sadagic A., Scheidegger C., Isenberg T., (Eds.), Lecture Notes in Computer Science, Springer International Publishing, pp. 246–258. doi:10.1007/978-3-319-50832-0_24. 2

[FNC19] FOWLER A., NESBITT K., CANOSSA A.: Identifying Cognitive Load in a Computer Game: An exploratory study of young children. In *2019 IEEE Conference on Games (CoG)* (London, United Kingdom, Aug. 2019), IEEE, pp. 1–6. URL: https://ieeexplore.ieee.org/document/8848064/, doi:10.1109/CIG.2019.8848064. 1

[IDQ21] ISLAM R., DESAI K., QUARLES J.: VR Sickness Prediction from Integrated HMD's Sensors using Multimodal Deep Fusion Network, Aug. 2021. arXiv:2108.06437 [cs]. URL: http://arxiv.org/abs/2108.06437, doi:10.48550/arXiv.2108.06437. 1, 4

[JWM*12] JANSEN M., WHITE T. P., MULLINGER K. J., LIDDLE E. B., GOWLAND P. A., FRANCIS S. T., BOWTELL R., LIDDLE P. F.: Motion-related artefacts in EEG predict neuronally plausible patterns of activation in fMRI data. *Neuroimage 59*, 1-3 (Jan. 2012), 261–270. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221044/, doi:10.1016/j.neuroimage.2011.06.094. 1

[KB22] KHOKHAR A., BORST C. W.: Modifying Pedagogical Agent Spatial Guidance Sequences to Respond to Eye-Tracked Student Gaze in VR. In *Symposium on Spatial User Interaction* (2022), SUI '22, Association for Computing Machinery, p. 12. URL: https://doi.org/10.1145/3565970.3567697, doi:10.1145/3565970.3567697. 3

[KJ13] KUHN M., JOHNSON K.: *Applied Predictive Modeling*. Springer New York, New York, NY, 2013. URL: http://link.springer.com/10.1007/978-1-4614-6849-3, doi:10.1007/978-1-4614-6849-3. 1, 4

[KYB19] KHOKHAR A., YOSHIMURA A., BORST C. W.: Pedagogical Agent Responsive to Eye Tracking in Educational VR. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Mar. 2019), pp. 1018–1019. ISSN: 2642-5254. doi:10.1109/VR.2019.8797896. 2

[LJB13] LI N., JAIN J. J., BUSSO C.: Modeling of Driver Behavior in Real World Scenarios Using Multiple Noninvasive Sensors. *IEEE Transactions on Multimedia 15*, 5 (Aug. 2013), 1213–1225. URL: http://ieeexplore.ieee.org/document/6416069/, doi:10.1109/TMM.2013.2241416. 1

[LNS*20] LUO S., NGUYEN A., SONG C., LIN F., XU W., YAN Z.: OcuLock: Exploring Human Visual System for Authentication in Virtual Reality Head-mounted Display. In *Proceedings 2020 Network and Distributed System Security Symposium* (San Diego, CA, 2020), Internet Society. URL: https://www.ndss-symposium.org/wp-content/uploads/2020/02/24079.pdf, doi:10.14722/ndss.2020.24079. 1

[MHGP16] MCDUFF D. J., HERNANDEZ J., GONTAREK S., PICARD R. W.: COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera. *Conference on Human Factors in Computing Systems - Proceedings* (2016), 4000–4004. ISBN: 9781450333627. doi:10.1145/2858036.2858247. 1

[MMDR20] MOORE A. G., MCMAHAN R. P., DONG H., RUOZZI N.: Extracting Velocity-Based User-Tracking Features to Predict Learning Gains in a Virtual Reality Training Application. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Porto de Galinhas, Brazil, Nov. 2020), IEEE, pp. 694–703. URL: https://ieeexplore.ieee.org/document/9284660/, doi:10.1109/ISMAR50242.2020.00099. 1

[SBZ*22] SHU X., BAO T., ZHOU Y., XU R., LI Y., ZHANG K.: Unsupervised dam anomaly detection with spatial–temporal variational autoencoder. *Structural Health Monitoring* (Apr. 2022), 14759217211073301. Publisher: SAGE Publications. URL: https://doi.org/10.1177/14759217211073301, doi:10.1177/14759217211073301. 4

[SDA*16] SANTOS I., DAM P., ARANTES P., RAPOSO A., SOARES L.: Simulation training in oil platforms. In *2016 XVIII symposium on virtual and augmented reality (SVR)* (2016), pp. 47–53. tex.organization: IEEE. 2

[SDSB*16] SABATOS-DEVITO M., SCHIPUL S. E., BULLUCK J. C.,

BELGER A., BARANEK G. T.: Eye tracking reveals impaired attentional disengagement associated with sensory response patterns in children with autism. *Journal of autism and developmental disorders 46*, 4 (2016), 1319–1333. Publisher: Springer. 1

[SZ15] SIMONYAN K., ZISSERMAN A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, Apr. 2015. arXiv:1409.1556 [cs]. URL: http://arxiv.org/abs/1409.1556, doi:10.48550/arXiv.1409.1556. 4

[VCR*18] VILLANUEVA I., CAMPBELL B., RAIKES A., JONES S., PUTNEY L.: A Multimodal Exploration of Engineering Students' Emotions and Electrodermal Activity in Design Activities: A Multimodal Exploration of Engineering Students' Emotions. *Journal of Engineering Education 107* (Sept. 2018). doi:10.1002/jee.20225. 1

[VDA*19] VELIYATH N., DE P., ALLEN A. A., HODGES C. B., MITRA A.: Modeling Students' Attention in the Classroom using Eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference on ZZZ - ACM SE '19* (Kennesaw, GA, USA, 2019), ACM Press, pp. 2–9. URL: http://dl.acm.org/citation.cfm?doid=3299815.3314424, doi:10.1145/3299815.3314424. 1

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.: Attention Is All You Need, Dec. 2017. arXiv:1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762, doi:10.48550/arXiv.1706.03762. 4

[XEAZ*21] XUE T., EL ALI A., ZHANG T., DING G., CESAR P.: CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. *IEEE Transactions on Multimedia* (2021), 1–1. Conference Name: IEEE Transactions on Multimedia. doi:10.1109/TMM.2021.3124080. 1

[YKM12] YONETANI R., KAWASHIMA H., MATSUYAMA T.: Multimode saliency dynamics model for analyzing gaze and attention. In *Proceedings of the symposium on eye tracking research and applications* (2012), pp. 115–122. 1

[ZLC*17] ZHAO B., LU H., CHEN S., LIU J., WU D.: Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics 28* (Feb. 2017), 162–169. doi:10.21629/JSEE.2017.01.18. 4