 doi:10.20944/preprints201909.0019.v1

Trends in Virtual and Augmented Reality Research: A Review of Latest Eye Tracking Research Papers and Beyond

Jicheng Li*
Computer & Information Sciences
University of Delaware

Roghayeh Barmaki†
Computer & Information Sciences
University of Delaware

Abstract

Although proposition of "Virtual Reality" (VR) and "Augment Reality" (AR) can be traced back to the 60s, both areas are actually blooming in recent decades. Thanks to the latest deep learning techniques, an enormous advance in computer vision research community has taken place. Since VR and AR are highly related to computer vision tasks, these areas have enjoyed the benefits as well. Yet there is no sufficient survey on such impact and new research areas arising from it. This paper mainly focuses on the latest research progress in ACM Symposium on Eye Tracking Research & Applications (ETRA) 2019, as well as several recent representative paper works. It aims to figure out the influence of deep learning techniques on latest VR/AR research. Meanwhile, new issues have popped up with the development of VR and AR technology, such as privacy and computation efficiency. This paper draws attention to such newly produced topics as well. In addition, this paper also investigates on the effect of latest VR and AR techniques on people, such as level of teamwork in collaborative tasks, assistance and treatment to patients and the disabled, etc.

Keywords: Virtual Reality; Augmented Reality; Eye Tracking; Deep Learning; Gaze estimation; Collaborative Computing

1 Introduction

Virtual Reality (VR) is defined as a "computer-generated simulation of a three-dimensional image or environment that can be interacted with in a seemingly real or physical way by a person using special electronic equipment, such as a helmet with a screen inside or gloves fitted with sensors.", per Oxford dictionary. [4]. Unlike traditional user interfaces and human computer interaction methods, VR techniques create a simulated environment, place the user inside such an environment, and provide a nicely authentic experience to user.

There are two kinds of VR systems: immersive and non-immersive. Immersive VR systems typically require users to wear auxiliary apparatus, such as head-mounted display(HMD) and control sticks. An HMD will cover user's eyes and provide visual feedback from virtual environment, probably along with audio and vibration signals. And user could interact with virtual environments, i.e. hitting an object, by the control stick. There are already a bunch of fantastic immersive VR games on market, like Beat Saber. Non-immersive VR systems are powered by computers and allow users to explore the virtual environment with an unobstructed sight but bear a trade-off in experience. This paper mainly focuses on immersive VR systems.

Augmented reality (AR),by definition of Merriam-Webster dictionary, is "an enhanced version of reality created by the use of technology to overlay digital information on an image of something being viewed through a device (such as a smartphone camera)". Such overlaying method allows user to interact with the virtual images using real objects in a seamless way. The overlaid information can be either constructive or destructive, namely, either additive to or masking of the natural environment. AR is mainly characterised by [1]

- Combine real and virtual imagery;
- Interactive in real time;
- Register the virtual imagery with the real world.

. Good examples for AR application include Snapchat lenses, Pokemon Go, etc.

To explain the difference between VR and AR in a nutshell, AR adds digital elements to real world to create a live view which is visible by people with the help of digital products, i.e., smartphone and tablet, while VR usually blots out the real world to create a virtual, immersive world.

Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. [10] Deep convolutional neural networks(CNN), state of the art in computer vision field, have outperformed human performance in many vision tasks like classification. Deep CNNs also have a huge impact on AR and VR research. For example, deep learning method is a better solution for eye detection, which is a key step in eye tracking process.

2 Method

The main method used in this paper is to review papers published on ERTA 2019, including co-located events and workshops Communication by Gaze Interaction (COGAIN), Eye Tracking For Spatial

*e-mail: lijichen@udel.edu
†e-mail: rlb@udel.edu

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2019 doi:10.20944/preprints201909.0019.v1

Research (ET4S), Eye Tracking And Visualization (ETVIS), and Eye Tracking For The Web (ETWEB). In addition, several selected papers from other resources are also included. The search of papers has been made on three different search strings: "Virtual Reality", "Augmented Reality" and "Eye Tracking". Papers published within 3 years, containing one or more search strings were pre-selected. A finer screening was conducted to retain works related to deep learning, collaborative learning and multi-modality. This paper has explored datasets including Google Scholar, Web of Knowledge, Scopus, ACM Digital Library and IEEE Xplore Digital Library.

3 Research Trend And Future Directions

Upon reviewing work published in ETRA 2019 and all other selected papers, recent trends and several possible future directions are speculated for further research.

3.1 Visual Attention

J. L. Louedec et al. [11] proposed a deep neural network to predict visual attentions of a chess player combining bottom-up and top-down approaches. The proposed neural network was trained using eye tracking data of a player in game, and was capable of generating meaningful saliency map (see figure 1), a representation of a player's visual attention,on unseen game configurations.

Figure 1: Saliency map from a chess game. At left, input chess board image. At centre, eye tracking results (points represent eye fixations, lines are scan path between fixations). At right, saliency map computed from eye tracking. For each pixel, a probability between 0 (black) and 1 (white) is computed. Adopted from [11, Fig. 1].

In [16], the paper compared how spatial attention was oriented in virtual environment given cues in different modalities.The task for a participant was to create a sandwich with correct ingredients in VR given cues. Cues, either valid or invalid, were intended for facilitating or misleading the participant (see figure 2). Two types of attention orienting were involved, endogenous and exogenous orienting. Endogenous orienting implies top-down processing, while exogenous orienting implies bottom-up involuntary processing. It turned out that all valid cues made participants react faster. In addition, cues in different modalities had different effect. Directional arrow (visual endogenous) and 3D sound (auditory exogenous) oriented attention globally to the entire cued hemifield, in contrast, vocal instruction (auditory endogenous) and object highlighting (visual exogenous) allowed more local orientation.

[12] proposed a novel joint attention training approach using a Customizable Virtual Human (CVH) and a Virtual Reality (VR) game as assistance, namely, Imagination Drum, where a virtual teacher will teach user drum skills. The game allowed the user to

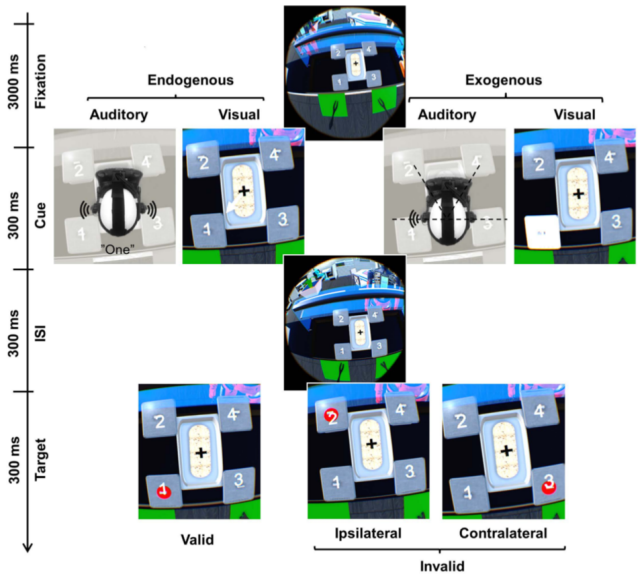

Figure 2: From top to bottom the time course of a trial: fixation of 3 seconds, cueing of 300 ms according to the block modality and type, 300 ms inter-stimulus interval, and 300 ms target presentation with a small red ball. Adopted from [16, Fig. 2].

customize the drum teacher (i.e., age, gender, hairstyle, skin, eye color, see figure 3), and provided feedback in multiple modalities (visual, sound, vibration, etc.) to arouse the enthusiasm of the user. To interact to the virtual teacher, the user could use a virtual drumstick to hit the drums follow the lead. The paper concluded that CVH made the participants gaze less at the irrelevant area of the games storyline, i.e. background.

Figure 3: The User Interface for customizing the CVH Teacher. Adopted from [12, Fig. 3].

In [18], participants were required to take nursing training using Rapid Response Training System (RRTS), a virtual environment(see figure 4) where participants could monitor and communicate with a virtual patient. The trainee had to respond properly based on the patient's vital signs and verbal feedback. Trainees were randomly assigned to one of three conditions,wherein the virtual patient either: (1) Not animated; (2)Played idle animations; (3) Played idle animations and provided appropriate eye contact, lip-synced speech and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2019 doi:10.20944/preprints201909.0019.v1

facial gestures. It proved that conversational and non-conversational animations successfully elicited visual attention of the trainee.

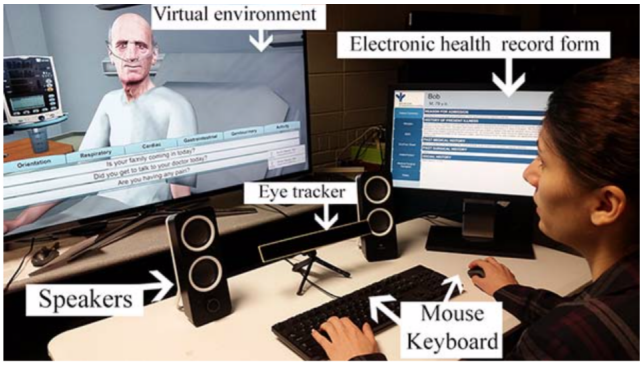

Figure 4: Screenshot shows a participant interacting with the virtual patient in the RRTS, and recording his vitals in the EHR screen. Adopted from [18, Fig. 1].

Rawan et al. [14] proposed a novel data-driven optimization approach for automatically analyzing visual attention and placing visual elements in 3D virtual environments. They created a virtual museum and asked participants to explore the museum freely. The gaze data of the participants were recorded by eye tracker. Then they used the gaze data to train a regression mode, to predict gaze duration for each position of the museum. And the layout of the museum was optimized by placing artworks on locations where participants will spend more time on, as demonstrated in figure 5.

Figure 5: Left: an input 3D scene with its corresponding layout. Right: the optimal placement of visual elements that will attain the target gaze duration. The eyes depict the camera location and angle in taking the screenshots. Adopted from [14, Fig. 1].

In [8], a novel, data-driven eye-head coordination model, SGaze, was promoted. SGaze was capable of real-time gaze prediction for immersive HMD-based applications without any external hardware or eye tracker (see figure 6). It was developed based on the fact that there was a linear correlation between gaze positions and head rotation angular, and there exists a latency between eye movements and head movements. Yet SGaze was designed for passive tasks, i.e. free exploration of visual scenes, where long saccades seldom exist.

3.2 Collaborative Learning

Collaborative learning is a widely-used education pattern featured by small group interaction and team-based evaluation metrics. Typically two or more participants are assigned in the same group and work for a common purpose, which encourages them to learn via teamwork. Compared with individual learning or lecture-based learning,

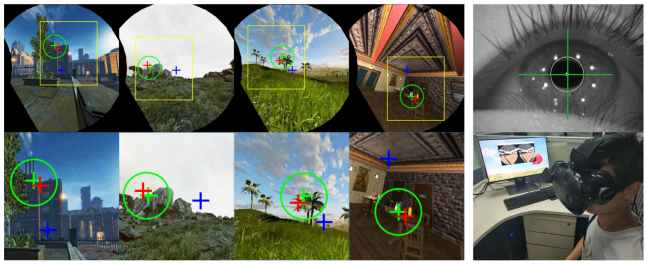

Figure 6: Green cross: ground truth; red cross: gaze prediction using SGaze; blue cross: mean baseline; Green Circle: foreal region with a 15° field of view. The upper row shows the images captured from an HMDs screen, with zoomed-in view in the lower row. The rightmost column demonstrates experiment setup. Adopted from [8, Fig. 1].

collaborative learning as an active learning approach can increase students learning motivation and improve knowledge retention [5].

Špakov et al. [19] investigated how sharing visual attention in a collaborative game will affect overall game performance. The theme of the game was to explore a darkened house to find keys located in different rooms (see figure 7). Players had to collaborate with their head gaze(the direction that a person is facing) or eye gaze (where a person is looking at) information shared to each other. Two versions of the game, desktop version (low-level immersion) and VR-HMD version (high-level immersion) were developed and tested. And experiment result shows that sharing eye-gaze information in the high immersion condition produced better performance.

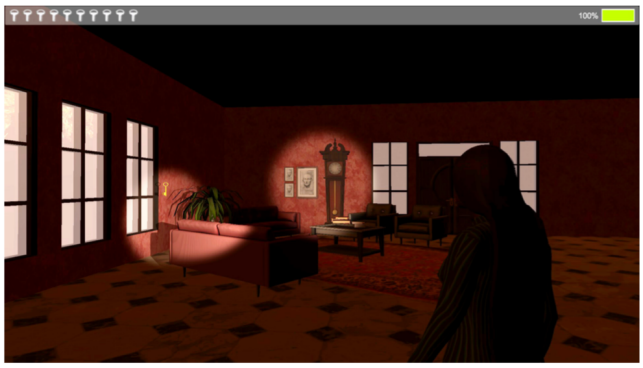

Figure 7: A screenshot from the game. Highlighted areas represent visual attention of both players.Adopted from [19, Fig. 2].

[13] aims to investigate the effects that an HMD-based AR system can have on eye contact behaviour between professional participants in a collaborative task. The participants, professionals from police, fire department and air force, worked through three different scenarios, alternating between HMDs (see figure 8) and regular paper maps, with the purpose of managing the crisis response to a simulated major forest fire collaboratively. For both HMDs and paper map teams, eye contact were pretty low (on average 2% for paper map and 0.2% for AR). Yet confidence and trust in the artefacts was rated significantly higher with HMDs than without. Contrary to popular assumptions, the decrease in eye contact with HMDs does not seem to have a direct effect on the collaboration in a professional,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2019 doi:10.20944/preprints201909.0019.v1

task-oriented context.

Figure 8: Participants using HMD-based AR system as auxiliary technical tools. Adopted from [13, Fig. 2].

In [6], a local worker aimed to assembly Lego while sharing video and virtual gaze information with a remote helper. The remote helper can provide feedback using a virtual pointer on the live video view. There was a significant improvement in worker's performance with remote help than without.

3.3 Emotion Detection

S. Chen et al. [2] introduced a novel real-time system that was able to capture and reconstruct 3D faces wearing HMDs and robustly recover eye gaze. Since user's face was partially occluded by HMD, they inserted two infrared cameras in VR glasses to capture eye images. In addition, facial images were captured by an extra infrared camera mounted outside of HMD, and head posture was collected by sensor of a mobile phone fitted to the VR glasses. An avatar driven by the captured 3D face was able to reflect user's emotion properly. The workflow is demonstrated in figure 9.

[7] presented an algorithm to automatically infer expressions by analyzing only a partially occluded face while the user is engaged in VR, and generated dynamic avatars in real-time which functioned as an expressive surrogate for the user. A novel approach to increase accuracy of deep convolutional neural networks, “personalization”, was introduced. The paper advocated that images of the users eyes captured from an IR gaze-tracking camera within a VR headset are sufficient to infer a selected subset of facial expressions , without the use of any fixed external camera. Figure 10 shows how the model works.

3.4 Accessibility and Inclusion

Joint attention training holds the potential to help children recover from Austim Spectrum Disorder(ASD). A child with ASD typically have restricted interests and repetitive behaviors, and suffers from communication and interaction with other people. [12] proposed a novel joint attention training approach using a Customizable Virtual Human (CVH) and a Virtual Reality (VR) game as assistance,

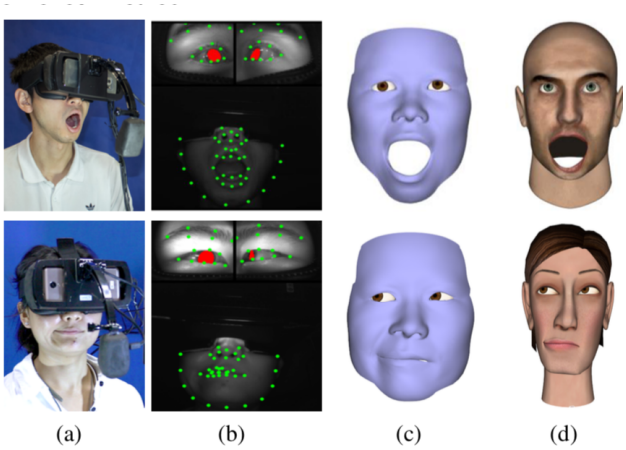

Figure 9: 3D facial expression reconstruction and eye gaze tracking. (a) The picture captured by an extra RGB camera to show the setup. (b) The three captured IR images. (c) The reconstructed 3D face and eye gaze. (d) An avatar driven by the captured 3D face. Adopted from [2, Fig. 3].

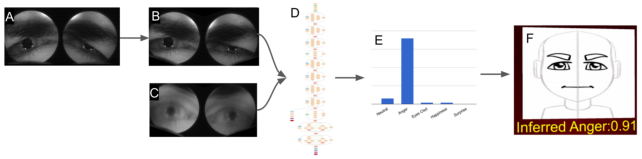

Figure 10: A: Raw eye images from the HMD. B: Rectified eye images. C: The average neutral image for this user session, used for personalization. D: The difference between the rectified headset image and the mean neutral image is the input to a deep neural network. In the non personalization case, the mean neutral image is not subtracted from the rectified image. E: Output takes the form of a distribution over expressions. F: This distribution is used to generate an expressive avatar. Adopted from [7, Fig. 3].

Imagination Drum, where a virtual teacher will teach user drum skills. As mentioned above, Imagination Drum did help on capturing children’s attention, enhancing children's communication ability, as well as improving level of engagement in social activities such as attending a music class at school.

Z.Chen et al. [3] proposed a strabismus recognition approach (see figure 11) using eye-tracking data and deep convolutional neural networks. By converting raw gaze data to gaze deviation(GaDe) maps and GaDe images, the gaze pattern was retained and can also be feed as input for convolutional neural networks. A GaDe image (see figure 12) was generated by three GaDe maps, which serve as R, G, and B channels of the image. Each GaDe map was generated based on the fixation accuracy (Euclidean distance between gaze position and target position) of left gaze, right gaze and center gaze (mean of left and right gaze), respectively. On each GaDe map, the position of a gaze point is determined by eye tracking record, and the value of a gaze point is calculated by converting its fixation accuracy to a meaningful RGB value. Evidently, VGG-S [15] get the best performance for this task.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2019 doi:10.20944/preprints201909.0019.v1

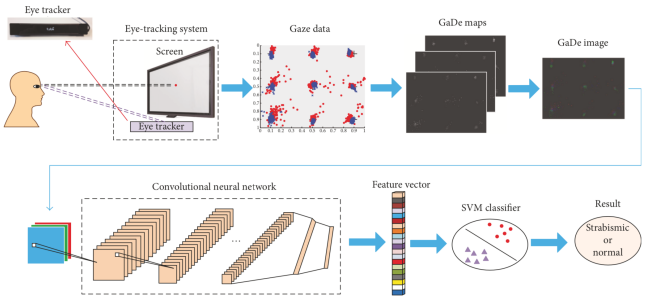

Figure 11: The proposed strabismus recognition framework. Adopted from [3, Fig. 1].

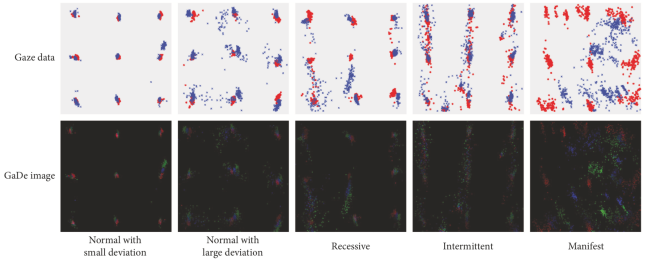

Figure 12: Examples of gaze data and corresponding GaDe images.Red $*$: left gaze data. Blue $\times$: right gaze data. Colors in the second row represent the R, G, and B channels of GaDe images. The The first two columns represent normal data with small deviation and large deviation. The third, fourth, and fifth columns represent data of, respectively, recessive strabismus, intermittent strabismus, and manifest strabismus. Adopted from [3, Fig. 4].

3.5 Privacy

Steil et al. [17] introduced privacy concern derived by first-person camera of eye tracking facilities, etc. To avoid unexpected private issues, privacy sensitive scenes were detected by scene camera. If a scene image was classified as sensitive scene by pre-trained deep neural network, then camera shutter will block the camera. Figure 13 demonstrates how the model, PrivacEye, was triggered.

3.6 Efficiency

By capturing the IR reflection from the eye using only a sparse grid of IR detectors, photosensor oculography (PSOG) is a promising solution for reducing the computational requirements of eye tracking sensors in wireless virtual and augmented reality platforms. Yet PSOG devices suffer from a performance degradation in the presence of sensor shifts. Katrychuk et al. [9] proposed a novel machine learning-based solution (see figure 14) for addressing the issue and thus increase computation efficiency.

4 Conclusion

This paper has reviewed the development in VR research presented over the last few years at ETRA and other established resources, with a particular focus on eye tracking, deep learning and gaze estimation and collaborative learning. Currently, there are huge advance in bringing VR and AR research from laboratories to industry, especially in game markets. Benefit from hardware upgrades and

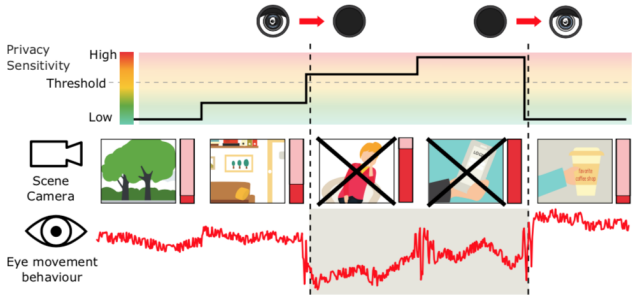

Figure 13: PrivacEye uses a mechanical camera shutter (top) to preserve users 'and bystanders 'privacy with head-mounted eye trackers. Privacy-sensitive situations are detected by combining deep scene image and eye movement features (middle) while changes in eye movement behaviour alone trigger the reopening of the camera shutter (bottom).Adopted from [17, Fig. 1].

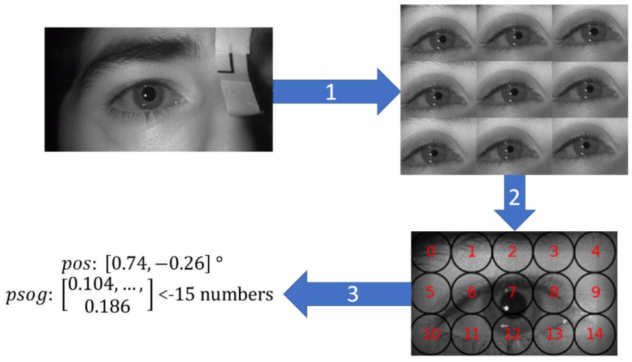

Figure 14: Preprocessing pipeline: (1) Account for head movements, then shift and crop. (2) For each cropped image,simulate PSOG sensor output(for each detection window one standard deviation of the gaussian kernel is depicted). (3) Save raw sensor output with corresponding eye gaze position. Adopted from [9, Fig. 2].

blooming in artificial intelligence, there is still huge potential for VR and AR technology in a wide range of areas.

References

[1] R. T. Azuma. A survey of augmented reality. *Presence: Teleoper. Virtual Environ.*, 6(4):355–385, Aug. 1997. doi: 10.1162/pres.1997.6.4.355

[2] S. Chen, L. Gao, Y. Lai, P. L. Rosin, and S. Xia. Real-time 3d face reconstruction and gaze tracking for virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 525–526, March 2018. doi: 10.1109/VR.2018.8446494

[3] L. W.-L. Chen Zenghai, Fu Hong and C. Zheru. Strabismus recognition using eye-tracking data and convolutional neural networks. *Journal of Healthcare Engineering*, 2018. doi: 10.1155/2018/7692198

[4] L. Freina and M. Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The International Scientific Conference eLearning and Software for Education*, vol. 1, p. 133. " Carol I" National Defence University, 2015.

[5] Z. Guo and R. Barmaki. Collaboration analysis using object detection. In *Proceedings of the 12th International Conference on Educational*

Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019, 2019.

[6] K. Gupta, G. A. Lee, and M. Billinghurst. Do you see what i see? the effect of gaze tracking on task space remote collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2413–2422, Nov 2016. doi: 10.1109/TVCG.2016.2593778

[7] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. A. Essa. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. *CoRR*, abs/1707.07204, 2017.

[8] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, May 2019. doi: 10.1109/TVCG.2019.2899187

[9] D. Katrychuk, H. K. Griffith, and O. V. Komogortsev. Power-efficient and shift-robust eye-tracking sensor for portable vr headsets. *11th ACM Symposium on Eye Tracking Research & Applications*, p. 19, 2019.

[10] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[11] J. L. Louedec, T. Guntz, J. L. Crowley, and D. Vaufreydaz. Deep learning investigation for chess player attention prediction using eye-tracking and game data. *11th ACM Symposium on Eye Tracking Research & Applications*, p. 1, 2019.

[12] C. Mei, B. T. Zahed, L. Mason, and J. Ouarles. Towards joint attention training for children with asd - a vr game approach and eye gaze exploration. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 289–296, March 2018. doi: 10.1109/VR.2018.8446242

[13] E. Prytz, S. Nilsson, and A. Jnsson. The importance of eye-contact for collaboration in ar systems. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pp. 119–126, Oct 2010. doi: 10.1109/ISMAR.2010.5643559

[14] H. H. Y. S. M. P. Rawan Alghofaili, Michael S Solah and L.-F. Yu. Optimizing visual element placement via visual attention analysis. 2019.

[15] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, p. arXiv:1409.1556, Sep 2014.

[16] R. Soret, P. Charras, C. Hurter, and V. Peysakhovich. Attentional orienting in virtual reality using endogenous and exogenous cues in auditory and visual modalities. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, pp. 86:1–86:8. ACM, New York, NY, USA, 2019. doi: 10.1145/3317959.3321490

[17] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling. Privaceye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. *11th ACM Symposium on Eye Tracking Research & Applications*, 2019.

[18] M. Volonte, A. Robb, A. T. Duchowski, and S. V. Babu. Empirical evaluation of virtual human conversational and affective animations on visual attention in inter-personal simulations. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 25–32, March 2018. doi: 10.1109/VR.2018.8446364

[19] O. Špakov, H. Istance, K.-J. Räihä, T. Viitanen, and H. Siirtola. Eye gaze and head gaze in collaborative games. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, pp. 85:1–85:9. ACM, New York, NY, USA, 2019. doi: 10.1145/3317959.3321489