How Much Do We Pay Attention? A Comparative Study of User Gaze and Synthetic Vision during Navigation

Julia Melgaré
PUCRS, Escola Politécnica
Brazil

Guido Mainardi
PUCRS, Escola Politécnica
Brazil

Eduardo Alvarado
LIX, École Polytechnique, CNRS, IP Paris
France

Damien Rohmer
LIX, École Polytechnique, CNRS, IP Paris
France

Marie-Paule Cani
LIX, École Polytechnique, CNRS, IP Paris
France

Soraia Musse
PUCRS, Escola Politécnica
Brazil

(a) Recorded User Gaze Heatmap (b) Predicted Saliency Map (c) Comparison Between Maps

Figure 1: Comparison between user gaze and predicted saliency [4] during navigation in a stylized environment.

ABSTRACT

Data-driven saliency models have proved to accurately infer human visual attention on real images. Reusing such pre-trained models in video games holds significant potential to optimize game design and storytelling. However, the relevance of transferring saliency models trained on real static images to dynamic 3D animated scenes with stylized rendering and animated characters remains to be shown. In this work, we address this question by conducting a user study that quantitatively compares visual attention obtained through gaze recording with the prediction of a pre-trained static saliency model in a 3D animated game environment. Our results indicate that the tested model was able to accurately approximate human visual attention in such conditions.

CCS CONCEPTS

• **Computing methodologies** → **Perception**; **Animation**.

KEYWORDS

Visual attention, saliency model, user gaze, user interaction.

1 INTRODUCTION

Recent advances in gaze recording devices have facilitated the development of saliency models that quantify human attention based on specific geometric and colored features [8]. These models can take advantage of recent machine learning techniques, by being trained on large datasets of real-world annotated images. Once learned, such saliency models are computationally efficient and can robustly handle arbitrary images. Moreover, they have been proven to reproduce visual attention in some real-world scenarios, such as in robotics [6]. In the case of virtual scenarios, control-based gaze behavior [2] has been shown to be effective in the design of attention systems. However, it remains to be seen whether such efficient pre-trained saliency models can be a viable solution to predict user attention in animated 3D virtual environments featuring game-like rendering, instead of requiring the use of captured gaze-tracking techniques [7].

To explore this point, we conducted a study to compare the visual attention of humans in an animated stylized virtual environment with a saliency model trained on real gaze data, but using static, natural images. Participants' visual attention data was recorded using an eye-tracking software, while they perceived a 3D stylized environment with various attention-grabbing stimuli. Video recordings of the user-observed screen were also provided as input to a pre-existing saliency model, able to predict saliency on a frame-by-frame basis. Finally, we compared the output maps of this saliency model to the user gaze heatmaps produced by eye-tracking software during the user experiment. An example of similarity measured between the two outputs is depicted in Figure 1. We opted for Kroner et al. [4]'s CNN-based visual saliency model which has already been used as a gaze model in photo-realistic virtual environments [3]. In contrast, we tested the limits of this pre-trained model for navigation in stylized scenes, which challenges its robustness.

MIG '23, November 15–17, 2023, Rennes, France

2 METHODOLOGY

In our user experiment, participants sat in a well-lit room behind a clear background. They faced an RGB camera[1] used by the eye-tracking software GazeRecorder[2] and a monitor screen displaying the virtual world. After calibrating the software to capture landmarks on the subject's image, we record their gaze data as they navigated the virtual environment through a fixed trajectory which lasted approximately 38 seconds. Once the trajectory was complete, we would stop recording and collect the output provided by GazeRecorder, consisting of two videos: one captures the user's screen view, while the other presents the same screen recording overlaid with a heatmap representation of the gaze data. The stylized 3D virtual environment presented a track through a forest, populated by various elements designed to capture the user's visual attention during navigation, such as static salient objects (fruit trees, buildings and flowers) and moving animated objects (animals). We conducted this experiment with a total of 50 participants, where 90% were between 18 and 35 years old, 76% had completed at least high school, and 70% were familiar with computer graphics. The gender distribution was balanced at 25 males and 25 females.

We directly compare the user gaze heatmaps obtained from GazeRecorder with the saliency maps generated by the model when we provide the participants' screen recordings as input. Considering the gaze heatmap of a participant generated by GazeRecorder H_u (Figure 1 (a)), and the Saliency map H_s generated using the saliency model (Figure 1 (b)), we convert both maps into binary images, by assigning an intensity value of 1 to any pixel in the map with an intensity greater than zero. We call B_u and B_s the resulting binary maps. We consider the saliency model to correctly predict the participant's attention if $B_u \cap B_s \neq \emptyset$, i.e., if there is any overlapping pixel between the two binary images. This operation is calculated on a frame-by-frame basis for each participant's gaze heatmap video, as illustrated in Figure 1(c). Given that, the accuracy of the artificial visual attention generated by the saliency model in comparison to that of the users is defined as the number of frames where an overlap between B_u and B_s was found, divided by the total number of frames of the user's video.

3 RESULTS

To ensure a fair comparison with the saliency model, we only considered user recordings that presented corresponding gaze data in at least 70% of the frames regarding the total duration of the experiment, which reduced the sample size to 9 users. This is caused by gaze detection issues due to changes in the environment or slight movements from the user. Figure 2 shows the calculated accuracy of the saliency model's predicted visual attention for this sample of participants, with the average being 93.9%.

Indeed, our results indicate that Kroner et al.'s saliency prediction model [4] effectively replicated human visual attention in a stylized virtual environment with reasonable accuracy, when compared to gazes captured from our small sample of participants. This is particularly noteworthy when we consider results obtained in previous studies [1, 5], where other deep learning models were less successful in matching humans' visual attention.

[1]The camera used had a max resolution of 720p and 30 fps.
[2]https://gazerecorder.com/

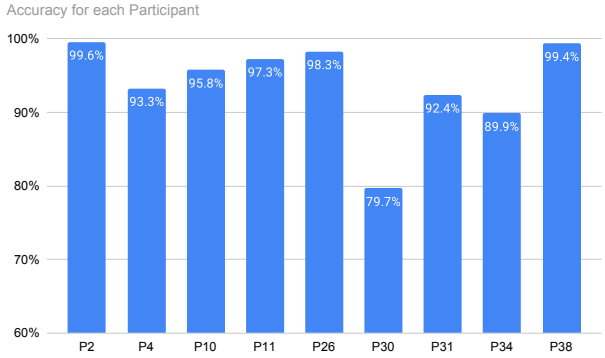

Figure 2: Accuracy of the saliency model [4]'s predicted visual attention compared to each participant's gaze behavior.

4 CONCLUSION

This work evaluated how a visual saliency prediction model trained with real image gaze data [4] performs when predicting salient regions in a stylized, game-like virtual environment. We achieved that by comparing the model's visual attention behavior with that of real humans exposed to the same stimuli in the same virtual world. As for limitations, using a camera-based eye-tracking software for capturing gaze data from the participants introduced some inaccuracy which led to us not being able to use all of the data collected in this study. We chose to thoroughly evaluate only one particular saliency model as it had already been tested in realistic virtual environments. However, for future work it would be interesting to evaluate other saliency models found in the literature, and to use different comparison metrics. Overall, our study provides a valuable insight for the game and virtual reality communities, by showing that some existing ML-based visual saliency models are already applicable to gaze prediction in non-photorealistic virtual navigation scenarios.

ACKNOWLEDGMENTS

The authors would like to thank CAPES, CNPQ, and PUCRS for financially supporting this project.

REFERENCES

[1] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017).
[2] Haegwang Eom, Daseong Han, Joseph S. Shin, and Junyong Noh. 2019. Model Predictive Control with a Visuomotor System for Physics-Based Character Animation. *ACM Trans. Graph.*, Article 3 (2019).
[3] Ific Goudé, Alexandre Bruckert, Anne-Hélène Olivier, Julien Pettré, Rémi Cozot, Kadi Bouatouch, Marc Christie, and Ludovic Hoyet. 2023. Real-time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters. *IEEE Transactions on Visualization and Computer Graphics* (2023).
[4] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. 2020. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks* 129 (2020).
[5] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia* 23 (2020).
[6] Ekaterina Potapova, Michael Zillich, and Markus Vincze. 2017. Survey of recent advances in 3D visual attention for robotics. *The International Journal of Robotics Research* 36, 11 (2017).
[7] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018).
[8] Qi Zhao and Christof Koch. 2013. Learning saliency-based visual attention: A review. *Signal Processing* 93, 6 (2013).