Eye tracking as a tool for analysing human -AI image interactions

Veslava Osińska
Nicolaus Copernicus University in Toruń
Institute of Information and Communication Research
Toruń, Poland
wieo@umk.pl,
0000−0002−1306−7832

Adam Szalach
Nicolaus Copernicus University in Toruń
Institute of Information and Communication Research
Toruń, Poland
aszalach@umk.pl,
0000-0001-8040-001X

Dominik M. Piotrowski
Nicolaus Copernicus University in Toruń
Nicolaus Copernicus University Library
Toruń, Poland
dpi@umk.pl,
0000-0002-3372-4772

***Abstract*— In recent years, artificial intelligence (AI) has significantly advanced fields like computer vision, image description, and generation, proving particularly relevant in creative areas such as generative art. This research aimed to explore AI's capabilities in creating and describing images compared to human perception. It included a comparative analysis of visual perception using eyetracking techniques in two settings: a VR art gallery created for the BITSCOPE project and a stationary ET study of individual images. The images, sourced from the BITSCOPE project's CHIST-ERA IV collection, were initially described by an expert following specific instructions, which were then used by AI to generate corresponding images. The eyetracking study focused on key areas and gaze plot sequences, using a gaze plot similarity metric based on topology and path length, enabled by the size of the research group.**

Keywords*— *artificial intelligence, eyetracking, virtual reality, computer vision

I. Introduction

In recent years, more and more online tools based on GANs (the Generative Adversarial Network) have been developed to enable users to create graphics easier and faster [1]. The democratisation of such tools and techniques highlights the problem of recognizing the machine's contribution to the final project. This has become a hot problem during the evaluation of student work, both essays and design.

The current research aimed to recognize the capabilities of artificial intelligence in the reconstruction of image content based on textual descriptions. It has become common for AI to assist humans in producing content. Thus the authors compare two descriptions: AI-origin and human-origin. The main problem the authors strive to solve is revealing which descriptions make the images generated most similar to the original. The authors investigate textual descriptions used as initial data to generate images. The last and key feature was studied using eye tracking in the desktop and VR environments.

GAN uses the description of a picture to be displayed [2]. Textual descriptions, as can be expected, are key. The authors compare images generated based on AI and human expert descriptions. A comparative analysis was performed due to an eye-tracking experiment. Do artificial intelligence descriptions correspond to what humans perceive? This question raises the next one: does artificial intelligence or a human better describe an image? Based on which descriptions will the images generated be most similar to the original?

II. Materials & Methods

Two images were selected for the study, which came from a collection gathered for the BITSCOPE project under the CHIST-ERA IV program. The first was a drawing by Jerzy Hoppen entitled "Death of Jakub Jasiński" (orphaned work) from 1956, and the second was an oil painting on canvas by Leonardo de Mango entitled "The Arrival of the Mahmal" (public domain) from 1921. The images were described by a senior curator from the Nicolaus Copernicus University Library in Toruń (Poland), who is an art historian by training, and who was initially instructed to write what he saw in the paintings. To create descriptions by artificial intelligence, the free online tool, the Pally Image Description Generator, was selected, where additional optional information was provided: "Describe what can be seen in the picture, specify the atmosphere, and provide information about the colours." The Image Creator from the Designer Image Creator application from Microsoft, which is based on DALL-E 3 , was used to generate images based on the created descriptions. DALL-E 3, was developed by OpenAI

The generated images were analysed to identify the 8 mappings (4 based on human descriptions and 4 AI origin descriptions) and the ones most similar to the created and generated descriptions, which were then examined. For the pilot experiment, 12 participants were selected, all with at least a secondary or bachelor's education, studying in the fields of computer science or journalism. All respondents reported no interest in art, infrequent visits to art galleries, and low interest in AI-generated graphics, except for 3 students who were interested in computer graphics due to their specialization.

The study was divided into two stages. The first stage involved the eye-tracking desktop experiment (GazePoint GP3 HD Eye Tracker 150 Hz), and the second stage took place in a VR environment (the BITSCOPE application with an HTC VIVE Pro Eye headset). These stages were separated by a minimum period of 5 days to allow to partially forget the previous part of the experiment.

Each respondent was asked to answer one question: "Do you think the displayed image was generated based on a description by a human or artificial intelligence?". The users were not informed of their results, the original images were not presented, nor were they told whether their answers were correct. The second stage of the study was conducted 5 days later, at the same location, and with the same respondents. Only 4 respondents reported previous experience with VR for entertainment purposes, and only one had used an HTC device.

III. RESULTS

Time characteristics of eye gaze

The human origin description and generated images were coded as "H". Two variations, H1 and H2, were related to the first sample picture, and accordingly H3 and H4 for the second. AI originated pictures were signed as AI1, AI2 (for the first sample) and AI3, AI4 (for the second sample).

TABLE I. RECOGNITION RATE OF IMAGES AND PERCEPTION TIME

Image code	recognition %		Slide's average perception time, s	
	desktop	*VR*	*desktop*	*VR*
H1	16.7	67.7	32.9	27.1
H2	58.3	83.3	33.0	39.3
AI1	33.3	75.0	33.0	50.0
AI2	41.7	58.3	23.6	27.3
H3	58.3	83.3	19.6	23.6
H4	75.0	63.7	18.9	27.4
AI3	58.3	58.3	22.0	37.9
AI4	75.0	58.3	20.6	33.7

As the table 1 presents, it is surprising that the largest average perception time points to the largest recognition rate in the VR case. The table by 96 rows (8x12) was analyzed by statistical tests. Biserial correlation test by Monte Carlo method revealed statistical significance for association between two variables: time and recognition of AI-origin images (N=48, p=0.011, α=0.05 two-tailed). The longer an observer perceives such kinds of graphics in VR, there is more chance of a true answer. But this correlation is weak r=0.374. Alternatively, display time of human-origin images can not be associated with their recognition (N=48, p=0.935, α=0.05). No statistically significant correlation between time and recognition rate were noted in the case of desktop experiment (N=48, p=0.071 (AI), p=0.641 (H), α=0.05).

Fixations based measures such as fixations count, frequency per s, fixations duration and saccade-based measures [3] (velocity and pixel length) have been averaged for each participant and used for statistical inference. Student tests show there are no statistical differences between the perception of human and AI originated images (N=48, p=0.347, α=0.05) within the desktop environment. The same results have been given for VR statistics.

However, the comparison of time parameters between two environments revealed noticeable differences. If desktop image average exposition time equals 25.5 s then VR perception per image lasts longer by 30% (33.3 s). The difference is statistically significant: N=96, p=0.031, α=0.05. The variances are also different that are statistically significant: p<0.0001, α=0.05). Eye gaze parameters such as fixations count or fixation/saccade ratio with comparison to desktop experiment differ by even two orders. This can be explained by two distinct ways of measuring and devices used for eye gaze (stationary eye tracker versus google eye tracking module). In the case of desktop experiment central vision was directly registered, while VR eye tracker also collected peripheral signals.

Similarity of paths

Eye gaze path can be recorded into a string by assigning for example the letter of the field that contains the current fixation. It is possible to identify similarity between two paths defined as the percentage of letters that the first string matches the second string. If we use Levenshtein distance, d (x,y), which computes the minimal cost of transforming string x to string y [4]. Then similarity function s (x,y) can be defined as follows:

$$s(x,y) = 1 - \ (d(x,y))/\max \ \ [\![(length(x), length\,(y))]\!]$$

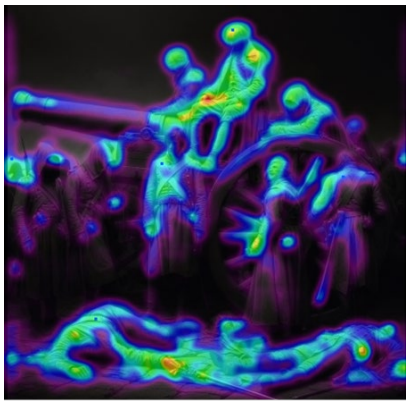
The sequence similarity is the percent of character sequences that are concordant in both strings [1]. For eye gaze movements this is the percentage of locations both scan paths have passed by, independently of time and sequence [17]. For example location similarity of eye gaze paths while users read extended human-origin description was very high 76.7%, and for AI description 63%. It means all participants captured largely the same key words during reading. All calculations were made by setting picture's grid into 5x5 for desktop conditions and 13x13 for VR environment because of different distance from observer to image plane. Table 2 presents comparison of calculated similarities values in both terms location and sequence.

TABLE II. PERCENTUAL SIMILARITY OF USERS' EYE GAZE PATHS

Image code	Local similarity		Sequence similarity	
	desktop	*VR*	*desktop*	*VR*
H1	74.6	53.9	21.5	20.1
H2	49.0	44.1	13.3	19.5
AI1	57.4	47.1	22.6	19.9
AI2	48.6	39.4	18.6	19.6
H3	50.5	41.0	19.5	18.0
H4	49.9	42.3	16.5	19.0
AI3	49.1	41.0	23.1	19.6
AI4	54.1	40.1	16.2	18.5

The participants' attention focuses at the same areas of images but in with different order – that is loci similarity is usually higher than sequence similarity. As we can see two pictures (H1 and AI1) became the winners in eye gaze similarities comparison (Table 2). H1 coded image reaches the value as for long textual description. The results are worth confronting with the scan path generated by a computational model of vision (Fig. 1, 2) based on human attention. This model reproduces the attentional scan paths by detecting local spatial discontinuities in intensity, color, and orientation, and finally combines them into a unique ''master'' or ''saliency'' map [5]

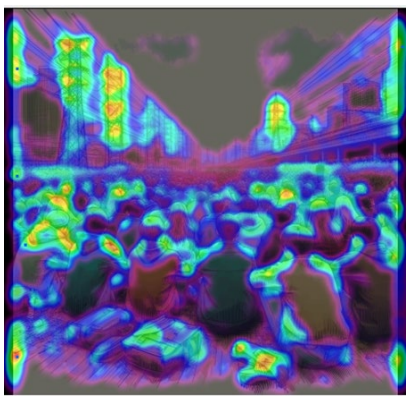
A



B

Fig. 1. Fixation map for Human origin (H1) picture (A) and its saliency orientation map (B)

A



B

Fig. 2. Fixation map for Human origin (AI2) picture (A) and its saliency orientation map (B).

IV. Conclusion

In the current study an expert created very extensive descriptions concerning visible characters but also the common emotional atmosphere of original images. Human-created texts consist of 106 words (for the first picture) and 64 words (for second picture) and were read several times longer in comparison with short AI-generated descriptions that counted 31 words both. The measure based on path location revealed that the participants read and insight long human-created descriptions in similar manier - similarity reached even 76.7%.. As for now AI image description generators are still far from human-produced description. They lack narration, fluency and association with historical, cultural facts characteristic for expert's or even average person's knowledge. They can, however, be used in areas such as search engine optimization or content accessibility for visually impaired individuals.

Displaying images in the first phase (desktop) and in the second phase (VR) eye tracking measures have been compared and gave noticeable differences. If desktop image average exposition time equals 25.5 s then VR perception per image lasts longer by 30% (33.3 s).

The measure that became common for two phases is eye gaze path similarity, evaluated by using Levenstein distance. Most users concentrate their visual attention on the same places in the case of two images: H1 and AI1 (Tab. 2). The same results relate to both environments. Theoretical model produced heat maps of these images showing significant covering of both patterns. Thus the concordance of theoretical and empirical results points an experiment should be developed in this direction.

The conclusion drawn from this research underscores the noticeable need for further improvement of AI tools in the field of computer vision, particularly in the context of image description. Future research aimed to recognize the capabilities of human perception of GAN images with the incorporation of emotions registering. Improvement in this area can contribute to a better understanding and utilization of artificial intelligence in the domain of generative art, enabling the creation of more refined and meaningful works.

Acknowledgments

The research is a part of project Bitscope is supported by the National Science Centre, Poland, under CHIST-ERA IV programme, which has received funding from the EU Horizon 2020 Research and Innovation Programme, under Grant Agreement no 857925.

References

[1] I. J Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. (2014). "Generative adversarial nets," in Proceedings of the 27th international conference on neural information processing systems - Vol. 2NIPS'14. (Cambridge, MA: MIT Press), 2672–2680.

[2] F. Chen, et al. (2021) 'Show, Rethink, And Tell: Image Caption Generation With Hierarchical Topic Cues', in. 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE Computer Society, pp. 1–6, [online:] https://doi.org/10.1109/ICME51207.2021.9428353, access June 19, 2024.

[3] K. Holmqvist, et al., Eyetracking: a comprehensive guide to methods and measures. Oxford University Press 2015.

[4] V. Levenshtein, (1966). Binary Codes Capable of Correcting Deletions and Insertions and Reversals. Soviet Physics Doklady, (8), 707-710.

[5] L. Itti, C. Koch (2001). Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging 10(1), 161–169.