1. Veslava OSIŃSKA[1], 2. Adam SZALACH[1], 3. Dominik M. PIOTROWSKI[2], 4. Jesus Casal MARTINEZ[3]

The Nicolaus Copernicus University in Toruń, Institute of Information and Communication Research (1),
The Nicolaus Copernicus University in Toruń, University Library (2),
Universitat Politècnica de València (3),
ORCID: 1.0000-0002-1306-7832; 2.0000-0001-8040-001X; 3.0000-0002-3372-4772; 4. 0000-0002-0690-0435

doi:10.15199/48.2025.03.66

Eye tracking and exploring AI potential in describing and generating images: a contribution to research on generative art

***Abstract**. Over the past few years, AI development has impacted fields like computer vision, image description, and generation. The article explored AI's capability to create descriptions and generate images, comparing these with human perception. Images were examined using eye tracking in a VR art gallery and on a desktop. The study involved expert and AI descriptions of BITSCOPE project images, followed by AI-generated images based on those descriptions, focusing on gaze plot metrics.*

Streszczenie.** W ciągu ostatnich kilku lat rozwój sztucznej inteligencji (SI) przyczynił się do postępów w takich dziedzinach jak widzenie komputerowe, opisywanie i generowanie obrazów. Analizy skupiły się na zdolności SI do tworzenia opisów i generowania obrazów, porównując je z ludzką percepcją. Obrazy były badane za pomocą śledzenia ruchu gałek ocznych w galerii sztuki VR oraz w środowisku stacjonarnym. Badanie obejmowało opisy obrazów projektu BITSCOPE dokonane przez eksperta i SI, a następnie generowane przez SI obrazy na podstawie tych opisów, koncentrując się na metrykach śledzenia wzroku. **(Eye tracking i badanie potencjału AI w opisywaniu i generowaniu obrazów: wkład w badania nad sztuką generatywną)

Keywords: artificial intelligence, eyetracking, virtual reality, computer vision.
Słowa kluczowe: sztuczna inteligencja, eyetracking, rzeczywistość wirtualna, widzenie komputerowe.

Introduction

Over the past few years, the development of artificial intelligence (AI) has brought significant advancements in various fields, including computer vision, image description, and generation. Understanding the potential applications of artificial intelligence in these areas is crucial, especially in the context of creative activities such as generative art. One of the most decisive steps forward in machine learning development has been the emergence of the family of algorithms known as generative adversarial networks (GANs), first proposed by Goodfellow et all [1]. Suppose previous models of neural networks (NN) like convolutional NN were used to recognize patterns in data, such as objects in images or words in text. In that case, GANs are used to create new data that follows a given pattern, such as realistic images or text. The novel approach is to generate new data by competition of two neural networks against each other. One network attempts to create new data, the second network strives to discern whether or not it is falsification. Through repeated training, both networks become better at their jobs.

With the appearance of the GANs we face a new way of creating the content, from artwork to music. It changes human factors in many fields such as healthcare, art, entertainment, and manufacturing. This kind of network enables the creation of synthetic content so realistic that it is often indistinguishable from human-generated creations. Some works are focused on how GANs are being incorporated into user pipelines for design practitioners [2] In recent years, more and more online tools based on GANs have been developed to enable users to create graphics easier and faster. 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 evaluation of student works, both essays and design.

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 be most similar to the original. The authors investigate textual descriptions used as initial data to generate images which were analysed in terms of user's recognition and perception. Last and key feature was studied by using eye tracking in two variants: desktop and VR environment.

Research questions

The following research questions were posed:

1. GAN use description of picture to be displayed. Textual descriptions as can be expected will be the key. The authors compare images generated based on AI and human expert descriptions. Comparison analysis was performed due to an eye-tracking experiment. Do artificial intelligence descriptions correspond to what humans perceive? This question evokes 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?
2. The most essential problem in the experiment is recognition of GAN images by users. By differentiating the users' groups it is possible to analyze how is visual perception functions in reaction of visual stimuli in different conditions, in particular in VR.
3. What kind of falsification (text or images) of GAN systems make greater impact on user?

Materials & Methods

Research Material

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 titled "Death of Jakub Jasiński" (orphaned work) from 1956, and the second was an oil painting on canvas by Leonardo de Mango titled "The Arrival of the Mahmal" (public domain) from 1921. The works are attached in the Annex. First, 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, specify the atmosphere, and provide information about the colors. To create descriptions by artificial intelligence, the free online tool 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 colors." The Image Creator from the Designer Image Creator application from Microsoft, which is based on DALL-E 3 [3], was used to generate images based on the created descriptions. DALL-E 3, developed by OpenAI, is an advanced AI model designed for generating highly detailed and realistic images from textual descriptions.

Experiment and equipment

The generated images were analyzed to identify the 8 mappings most similar to the created and generated descriptions, which were then examined. For the pilot experiment, 12 participants were selected, including 5 women (average age: 23.8) and 7 men (average age: 22), 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 eye-tracking methodology, and the second stage took place in a VR environment. These stages were separated by a minimum period of 5 days to allow partial forgetting of the descriptions, images, and responses from the previous part of the experiment.

The eye-tracking study was conducted in a specially prepared laboratory using the GazePoint GP3 HD Eye Tracker 150 Hz, the dedicated GazePoint Assistant software for control, and the Ogama Version 5.1 (30.03.2021) application. The application was also used for data collection and preliminary analysis. The experiment was carried out using a laboratory desktop computer and two 27-inch monitors (4K, 60Hz). The respondent and the examiner sat opposite each other, separated by the monitors, without maintaining eye contact. On the respondent's desk, in addition to the monitor and the measuring device, there was an additional pointing device used to change the image after observing it. The distance between the respondent and the monitor and eye tracker was approximately 1.5 meters. Before the experiment, the participants were informed about the topic and the measurement method. Each respondent was instructed about the study procedure and participated in a preliminary survey. To avoid potential distractions, only the experimenter and the respondent were present in the laboratory and the study was conducted according to the schematic presented in (Fig. 1). All tests were conducted under similar lighting conditions at the same time of day [4].

After familiarizing themselves with the tool and receiving a preliminary description of the study procedure, the respondent underwent a double ten-point calibration, which involved tracking a moving object on the screen with their eyes. Due to the requirements of the Ogama and GazePoint Assistant applications, the first calibration was preliminary and aimed to exclude vision dysfunctions and incorrect positioning relative to the measuring device.

The next step was to enter demographic data and conduct the proper calibration, which was evaluated on a point scale. The lower the point value, the more accurate the calibration. Based on documentation, previous experiments, and experience with the measuring device, it was established that values below 100 points are sufficient for accurate measurement. It should be noted that the device is sensitive to variable lighting conditions, vision dysfunctions, and other numerous variables.

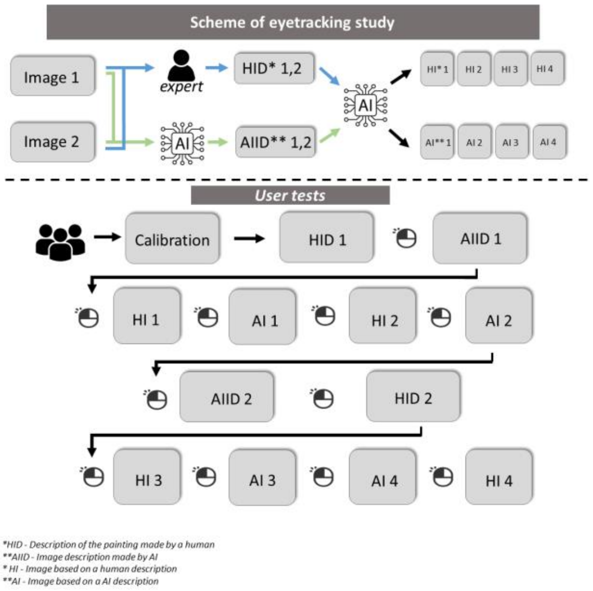

Scheme of eyetracking study
Image 1
Image 2
expert
HID* 1,2
AI
AIID** 1,2
AI
HI* 1
HI 2
HI 3
HI 4
AI** 1
AI 2
AI 3
AI 4
User tests
Calibration
HID 1
AIID 1
HI 1
AI 1
HI 2
AI 2
AIID 2
HID 2
HI 3
AI 3
AI 4
HI 4
*HID - Description of the painting made by a human
**AIID - Image description made by AI
* HI - Image based on a human description
**AI - Image based on a AI description

Fig. 1. Scheme of eyetracking study (The experiment performed in VR space looked the same. The exception was the presentation of all four descriptions before starting the study)

The study began with a slide reminding participants of the rules during the experiment. Progression to each subsequent slide was initiated by the respondent clicking the left mouse button. The next two slides contained descriptions of the painting "Death of Jakub Jasiński," one created by a human and the other by AI, followed by 4 slides with generated descriptions. In the second stage, there were two more descriptions and 4 more images.

While viewing the images, 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?". After the experiment finished, users were not informed of their results, the original images were not presented, nor were they told whether their answers were correct. No irregularities were observed during the entire study.

The second stage of the study, as mentioned earlier, was conducted 5 days later at the same location and with the same respondents. This stage was carried out using a prototype virtual gallery based on the Unity engine, created for the BITSCOPE (bitscope.umk.pl) project. The experiment used the HTC VIVE PRO EYE headset. Only 4 respondents reported previous experience with VR for entertainment purposes, and only one had used an HTC device. None of the respondents had prior experience with any form of virtual gallery or museum in VR.

Due to the capabilities of the application, the users were provided with printed descriptions of the images generated by AI and humans to familiarize themselves with. The content of the descriptions was not labeled or titled in any way.

The 8 images were arranged, with 2 on each of the 4 walls, in an environment created by the application that included various decorations such as virtual plants, sculptures, and objects to enhance immersion. The respondents approached the images in any order and, as with the previous stage, had to answer the question: "Do you think the displayed image was generated based on a description by a human or artificial intelligence?"

Application

BITSCOPE is an advanced application designed to create dynamic, immersive environments that simulate a VR experience in an art gallery or museum. The application is highly customizable, allowing each user and session to be unique. The main features of BITSCOPE include:

- Studies configuration: An experimenter can configure a study based on various parameters. These include selecting images from a library, determining the number of artworks, choosing textures for floors and walls, setting distractor elements, and adjusting the colour and intensity of lighting. Additional specific parameters can also be set according to the study's requirements.
- User Experience: The user, equipped with the HTC VIVE Pro Eye headset, engages in the VR experience by moving through different rooms. The application allows interaction with the environment and uses the teleportation metaphor for long-distance movements. The user can freely navigate between rooms, exploring and interacting with the displayed artworks.
- Data Tracking and Storage: The application collects and stores raw eye-tracking data, recording the user's interaction with the artworks. Furthermore, BITSCOPE synchronizes with the OpenVIBE EEG application, sending temporal markers that are recorded in the user's brain activity log, which is monitored externally.

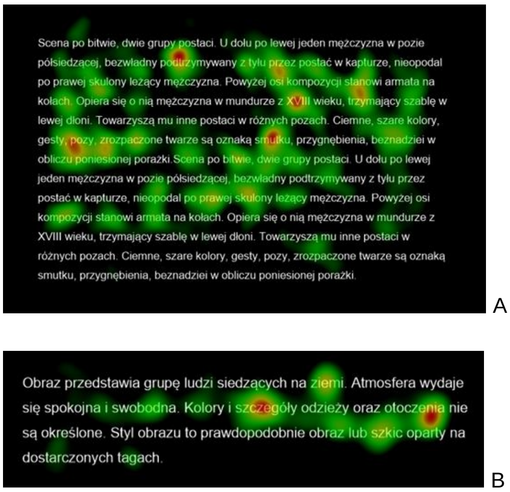
BITSCOPE was developed using Unity v2022.3.13, a robust and flexible platform for creating virtual environments and VR applications. The technical aspects of the development include:

- Modeling and Animation: Advanced 3D modeling tools were used to create the visual elements of the environment, from wall and floor textures to representations of artworks.
- Programming and Scripting: Custom scripts that control BITSCOPE's functionality were developed in C# using Visual Studio 2022. These scripts manage everything from the environment configuration by the experimenter to the collection of eye-tracking data and user interaction with the environment.
- User Interaction: The application is designed to provide an intuitive and seamless experience. The teleportation metaphor facilitates long-distance movements within the environment, and the free movement between rooms ensures unrestricted exploration.
- Data Synchronization: Integration with OpenVIBE for EEG recording is a key component of BITSCOPE. The application sends precise temporal markers that synchronize with brain activity data, allowing detailed analysis of the user's interaction in the context of their neural activity.
- Optimization and Testing: Extensive testing was conducted to ensure the stability and performance of the application. Optimization efforts focused on delivering a smooth VR experience with minimal latency, which is crucial for user immersion and the accuracy of eye and brain tracking.

Results

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).

AI description consisted on 31 terms has a reading average time 16 s, where the sentences are constructed according to well-known rules among practitioners. The sequence of minimal presentation should be following:



Scena po bitwie, dwie grupy postaci. U dołu po lewej jeden mężczyzna w pozie półsiedzącej, bezwładny podtrzymywany z tyłu przez postać w kapturze, nieopodal po prawej skulony leżący mężczyzna. Powyżej osi kompozycji stanowi armata na kołach. Opiera się o nią mężczyzna w mundurze z XVIII wieku, trzymający szablę w lewej dłoni. Towarzyszą mu inne postaci w różnych pozach. Ciemne, szare kolory, gesty, pozy, zrozpaczone twarze są oznaką smutku, przygnębienia, beznadziei w obliczu poniesionej porażki.Scena po bitwie, dwie grupy postaci. U dołu po lewej jeden mężczyzna w pozie półsiedzącej, bezwładny podtrzymywany z tyłu przez postać w kapturze, nieopodal po prawej skulony leżący mężczyzna. Powyżej osi kompozycji stanowi armata na kołach. Opiera się o nią mężczyzna w mundurze z XVIII wieku, trzymający szablę w lewej dłoni. Towarzyszą mu inne postaci w różnych pozach. Ciemne, szare kolory, gesty, pozy, zrozpaczone twarze są oznaką smutku, przygnębienia, beznadziei w obliczu poniesionej porażki.

Obraz przedstawia grupę ludzi siedzących na ziemi. Atmosfera wydaje się spokojna i swobodna. Kolory i szczegóły odzieży oraz otoczenia nie są określone. Styl obrazu to prawdopodobnie obraz lub szkic oparty na dostarczonych tagach.

Fig. 2. Eye tracking heat maps of human (A) and AI description (B).

scene, emotions, colors and style. The content is not fluent like in the expert’s text. An expert created a short history taking into account historical epochs. Eye tracking patterns of both descriptions are presented at Fig. 2. Human origin text is more detailed and counts much more, 106 terms and the reading time 64.4.

Recognition rate and time

As Table 1 shows, temporal characteristics do not show essential differences in dynamics of image perception.

We can see that in the case of human text, visual attention focuses on emotions, device („armata”), details of scene, stewards (colors, poses) and also digits like dates. Users have difficulties extracting essential words from other, very brief descriptions of AI origin. This causes more repetitions of eye gazing during reading.

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

Table. 1. 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

more chance of a true answer. But this correlation is weak *r*=0.374. Alternatively, display time of human-origin images cannot 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).

Time characteristics of eye gaze

Fixations based measures such as fixations count, frequency per s, fixations duration and saccade based measures (velocity and pixel length) have been averaged for each participant and used for statistical inference. Student tests show there are no statistical differences between 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*=0.0. 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. The field of view and also depth to image plane requires larger visual activity from the user to perceive stimuli. All this requires more careful planning of the experiment under identical measurement conditions.

Similarity of paths

A scan path T of length n - 1 can be expressed as a sequence of *n* fixations with a two-dimensional position p_j [5]:

$$(1) \quad T \coloneqq p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$$
$$(p_j \coloneqq (x_j, y_j) \in \mathbb{N} \times \mathbb{N}, 1 \le j \le n)$$

Eye gaze path can be recoded 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*[6]. Then similarity function *s*(*x*,*y*) can be defined as follows:

$$(2) \quad s(x,y) = 1 - \frac{d(x,y)}{\max(length(x), length\,(y))}$$

The sequence similarity is the percent of character sequences that are concordant in both strings [7]. For eye gaze movements this is the percentage of locations both scan paths have passed by, independently of time and sequence [8]. 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. It is worth to mention that the order of displayed pictures is different as presented in a table, in particular, VR experiment allows the user to view images according own preferences.

Table 2. 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

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. 3 B) based on human attention.

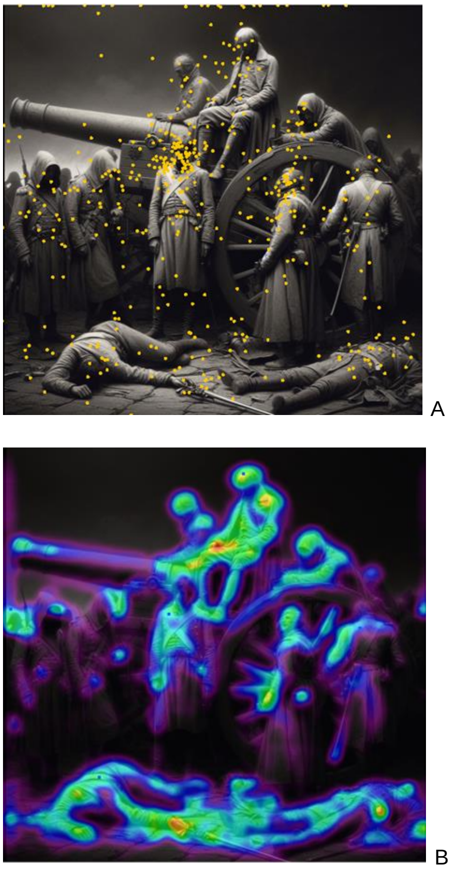


Fig. 3. Fixation map for Human origin (H1) picture (A) and its saliency orientation map (B).

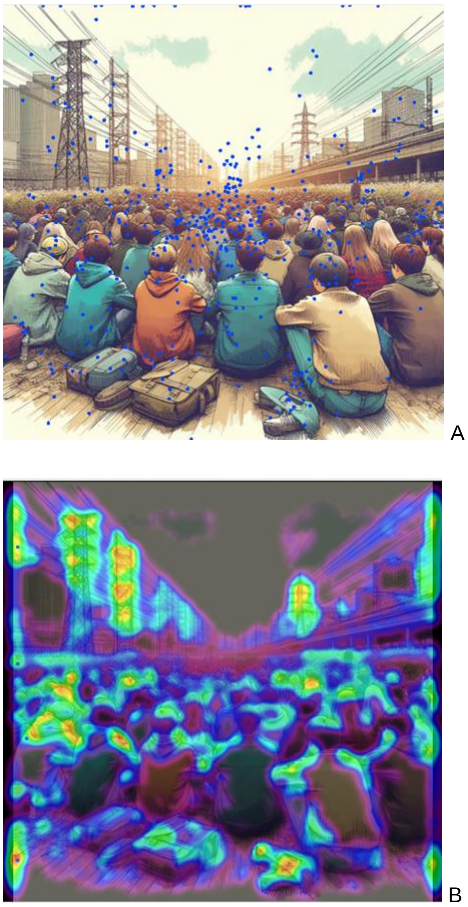


Fig. 4. Fixation map for Human origin (AI2) picture (A) and its saliency orientation map (B).

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 [9].

Discussion and conclusion

Using contemporary AI tools, it is possible to generate not only images but also descriptions of images. The authors used free tool Pally [10], making possible the addition of extra instructions. Nonetheless, other applications of this type are also available on the internet. These include Vision Studio [11] (add captions to images), which allows for easy creation of image summaries, or Astica [12], an application used for generating textual descriptions of images. The GPT-4o model, available through OpenAI's ChatGPT interface, also appears to be intriguing. This model could be utilized in future research.

AI image description generators analyze graphical units and generate descriptive text captions or summaries. As for now these tools are still far from human-produced description. They lack narration, fluency and association with historical, cultural facts characteristic of 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.

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 the second picture) and were read several times longer in comparison with short AI-generated descriptions that counted 31 words both. The velocity of reading varies from 1.52 to 1.65 words/s for human text and from 1.33 to 1.50 words/s for AI-origin items. The measure based on path location revealed that the participants read and insight long human-created descriptions in a similar manner - similarity reached even 76.7%.

Displaying images in the first phase (desktop) was carried out in predefined order. However in the second phase (VR) participants selected their own sequence of exhibited items. This randomness allows us to avoid bias related to remembering graphics and finally verify recognition rate and visual perception assessments. From the other side, the VR phase where the different device (eye tracker) and application for gathering eye gaze (own Bitscope product) were used, in time analyses should be considered separately from the first experiment. Bitscope procedure returns eye gaze data for full scenes instead of separate pictures like in the case of desktop measurements. Probably these different conditions cause fixations count for the images in the VR environment to be bigger by two orders than for the same items displayed on the screen. Moreover, an interesting finding was given for the VR environment. AI-origin graphics was longer perceived by users and the time of observation was correlated with the true answer, which was proved by statistical test.

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 (Fig. 3 and Fig. 4) showing significant covering of both patterns. Thus the concordance of theoretical and empirical results points an experiment should be developed in this direction. We can presume, the clue in this type of experiment is selection of graphics potentially inducing a universe pattern of visual attention. Clear and legible description appears to be added value in interaction between user and image.

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. In subsequent research, it is worth testing the effectiveness of description tools in the reference to perception of users watching generated images.

Future research is 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.

***Authors**: Veslava Osińska, Nicolaus Copernicus University in Toruń, Institute of Information and Communication Research, Bojarskiego 1, 87-100 Toruń, E-mail: wieo@umk.pl; Adam Szalach, Nicolaus Copernicus University in Toruń, Institute of Information and Communication Research, Bojarskiego 1, 87-100 Toruń, E-mail: aszalach@umk.pl; Dominik M Piotrowski, Nicolaus Copernicus University in Toruń, University Library, Gagarina 13, 87-100 Toruń, E-mail: dpi@umk.pl; Jesus Casal MARTINEZ, Universitat Politècnica de València, Camí de Vera, s/n, Algirós, 46022 València, E-mail: jcasmar2@htech.upv.es.*

REFERENCES

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

[2] Hughes R.T., Zhu L., Bednarz T., Generative Adversarial Networks–Enabled Human–Artificial Intelligence Collaborative Applications for Creative and Design Industries: A Systematic Review of Current Approaches and Trends. *Front. Artif. Intell*. 4:604234 (2021), doi: 10.3389/frai.2021.604234

[3] How to use AI image prompts to generate art using DALL-E, access June 19 (2024): https://create.microsoft.com/en-us/learn/articles/how-to-image-prompts-dall-e-ai

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

[5] Raschke, M., Blascheck, T., Burch, M., Visual Analysis of Eye Tracking Data. In: Huang, W. (eds) *Handbook of Human Centric Visualization.* Springer New York (2014), access June 19 (2024): https://doi.org/10.1007/978-1-4614-7485-2_15

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

[7] Brandt, S. A., & Stark, L. W., Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene. *Journal of Cognitive Neuroscience*, 9 (1), (1997), 27-38. MIT Press, access June 19, 2024: doi: 10.1162/jocn.1997.9.1.27

[8] Voßkühler, A., Nordmeier, V., Kuchinke, L., & Jacobs, A.M., OGAMA - OpenGazeAndMouseAnalyzer: Open source software designed to analyze eye and mouse movements in slideshow study designs, *Behavior Research Methods*, 40(4) (2008), 1150-1162.

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

[10] Image Description Generator - 100% Free, No Login „Pallyy", access June 19 (2024): https://pallyy.com/tools/image-description-generator

[11] Vision Studio, access June 19 (2024): https://portal.vision.cognitive.azure.com/demo/image-captioning

[12] Describe Images With AI: Scenes, Faces, Objects or Text | astica ai, access June 19 (2024): https://astica.ai/vision/describe-images

Annex

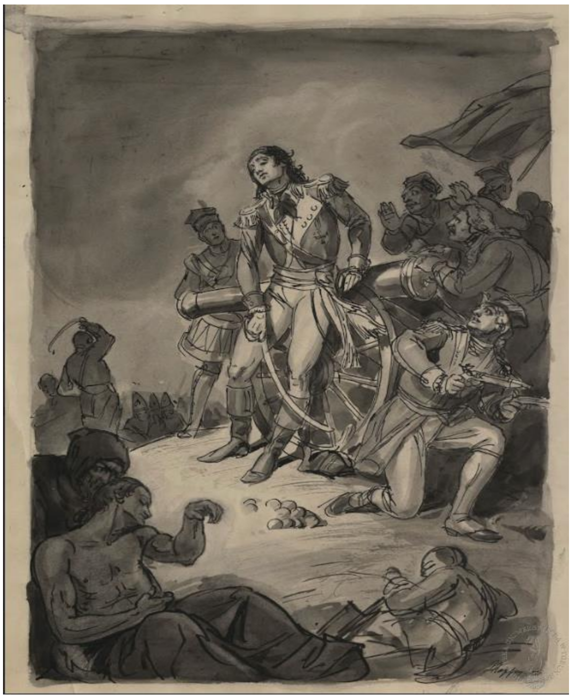


Image 1: Jerzy Hoppen, "Death of Jakub Jasiński", 1956.

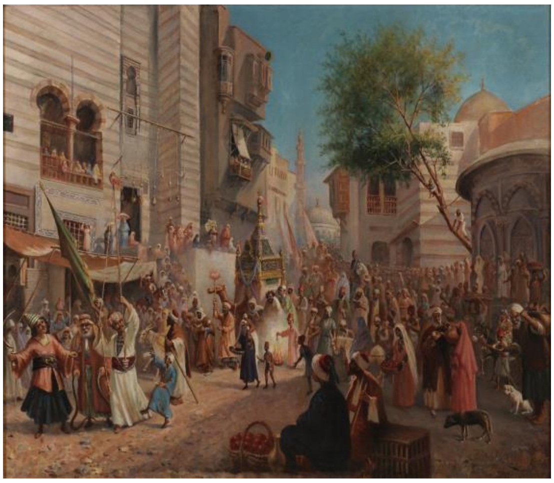


Image 2: Leonardo de Mango, "The Arrival of the Mahmal", 1921.