Could the bubbleview metaphor be used to infer visual attention on 3D graphical content ?

Alexandre Bruckert, Mona Abid, Matthieu Perreira da Silva, Patrick Le Callet

COULD THE BUBBLEVIEW METAPHOR BE USED TO INFER VISUAL ATTENTION ON 3D GRAPHICAL CONTENT ?

Alexandre Bruckert, Mona Abid, Matthieu Perreira da Silva, Patrick Le Callet

Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000, Nantes, France

ABSTRACT

Understanding the deployment of human gaze on 3D graphical objects is of critical importance in order to propose rich and complex 3D environments without strong latency nor rendering constraints. However, the data needed to study this gaze deployment can be costly and difficult to obtain, especially in the context of the Covid-19 pandemic where in-lab experiments are strongly discouraged. In order to alleviate these issues, we propose to use the BubbleView metaphor as a way of crowdsourcing visual attention data on 3D graphical content. In this paper, we question the adequacy of this method to provide a reliable proxy for visual attention in the context of 3D graphical objects. Moreover, we show that data obtained in this manner can be used to train visual saliency models, with only a slight tradeoff in performances compared to the use of ground-truth eye-tracking data.

***Index Terms*—** Visual attention, eye-tracking, visual saliency, 3D graphical objects

1. INTRODUCTION

Three-dimensional graphics have taken a prominent place in numerous fields of applications, including video games, digital animation or scientific simulation. However, as these graphical content become more detailed and complex, compression and simplification operations are critically needed to ensure that 3D objects are displayed to the users without rendering nor latency issues. In order to ensure an optimal quality of experience (QoE) for the user, visual attention information is often used, as it gives an indication about the perceived experience. For instance, visual attention can be used to infer the memorability of an image [1], or to spatially guide the compression of an image or video [2].

In this context, visual saliency, which refers to the distribution of eye fixations on a given stimulus, is a particularly interesting feature to study. Indeed, predicting this distribution allows to preserve visually important areas during compression. Consequently, numerous models dedicated to predict this distribution have been proposed in the literature [3]. While most of the effort has been concentrated on 2D images or videos, only a few studies explored visual saliency on 3D meshes [4, 5, 6] or point clouds [7, 8].

However, gathering visual saliency data is expensive in terms of resources, as it requires costly eye-tracking equipment and human supervision during test sessions, and thus prevents the collection of large-scale datasets needed to train deep learning models. In order to address this problem, several alternatives for in-lab eye-tracking experiments have been proposed [9].

Webcam-based eye-tracking [10] was proposed as a direct way to collect gaze data through crowdsourcing, by following the pupil of an observer captured through their webcam. However, it appeared to be quite difficult to gather reliable data with this method, due to important variations of the quality of the equipment, lighting conditions, or the positions of the observers.

Auto-annotation methods rely on the viewer's self-assessment of where gazed-upon areas are located, usually by asking participants to paint or to click over the interesting regions [11, 12]. These methods however generate a lot of noise because of the intra-observer diversity in annotation style, and the top-down nature of the collected ground-truth.

Zooming interactions [13] consists in giving the observer an interactive zooming tool, and treating the zoom-upon areas as areas of interest, from which visual attention heatmaps can be drawn. However, this methods is more suited to multi-scale content.

Mouse tracking methods rely on the assumption that gaze and mouse tracks are highly correlated, especially on specific stimuli where a mouse cursor is expected by the participant, such as web pages [14, 15]. The **BubbleView** metaphor is another form of mouse-tracking proxy for visual attention: the image is entirely blurred, except for an area around the mouse cursor. This unblurred region can either be displayed by clicking [16], or continuously follow the mouse cursor [17]. Finally, the BubbleView metaphor shows very good correlation with eye-tracking data on various content, and allows for the collection of very large visual saliency datasets, such as SALICON [18]. In this work, we propose to evaluate the BubbleView metaphor for crowdsourcing visual attention data on 3D graphical objects. We introduce a novel dataset composed of mouse tracks on high resolution 3D objects collected under laboratory conditions. We then compare the mouse tracks to ground truth eye-tracking data. Finally, we retrain several visual saliency models using these

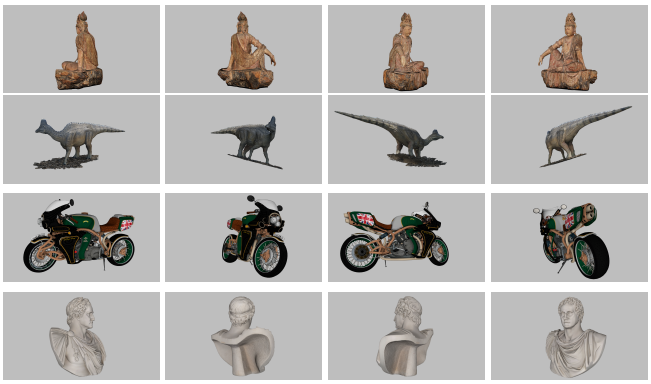

Fig. 1: Examples of the rendered stimuli

two datasets, in order to assess the interest of the BubbleView metaphor for predicting visual attention on 3D graphical content.

2. DATA COLLECTION AND PROCESSING

2.1. Stimuli

We build our dataset as an extension of the 3DGC saliency dataset [19]. We selected 25 high-resolution 3D objects, with various semantics and shapes, and selected four viewpoints to render them, using the same method as for the 3DGC data (i.e. considering four views corresponding to four faces of a cube), thus yielding a total of 100 images. In order to avoid any bias due to the occupancy of the objects, we applied a gray background behind them. Figure 1 shows an example of such images.

2.2. Eye-tracking data

Participants and task: In order to build an eye-tracking ground-truth, we collected data from 34 volunteers. Participants were asked to freely explore each stimulus, without any specific task. A trial session was performed, in order to ensure that the participants understood the procedure.

Apparatus and protocol conditions: We used the *Eyelink 1000 Plus* eye-tracker, sampling at 1000Hz, in remote mode (i.e. using no chin rest). Stimuli were displayed in a random order, at full HD resolution (1920×1080) during three seconds, on a DELL P2417H monitor. The distance between the eyes of the observers and the screen was set at 96cm ($\pm$ 1cm). In this setting, one degree of visual angle amounts to roughly 64 pixels. The screen luminance was set at 200 cd/m², and the luminance for the room's walls was measured at 30 cd/m².

Data processing: In order to transform raw gaze points into fixations, we applied a threshold-based aggregation algorithm using motion, velocity and acceleration [20], and removed all fixations lasting less than 80ms, i.e. roughly the minimum amount of time to process foveal information. Visual saliency maps were obtained by aggregating the fixation points over all observers, and convolving them with a 2D Gaussian kernel, which standard deviation was set to one degree of visual angle, which is approximately the radius of the fovea.

A mask was then applied in order to only consider the surface of the rendered object, and not the background. To take into account gaze data at the border of the surface, the size of the mask was enlarged using a morphological dilation, similar to the process in [19].

2.3. BubbleView data

In this study, we chose to use the BubbleView method in a continuous way, not relying on clicks, in order to grasp both bottom-up and top-down processes. This way, the whole image is blurred, except a circular region around the center of the mouse.

BubbleView parameters: Several parameters are required to set up the BubbleView metaphor. First, the size of the bubble needs to be set at approximately the size of the fovea [16]. The blur sigma also needs to be set such that some details are visible to attract the attention towards different regions, but not too many details, to force exploration using the bubble. Finally, the display time of the stimulus is also important, in order to have a balance between top-down and bottom-up processes as close as the one in the eye-tracking study.

The size of the bubble was set to 1.5° of visual angle, i.e. 96 pixels. For the blurring parameter, we applied a homogeneous blur, which kernel size is designed to match the blur in the peripheral vision at the edges of the screen when fixating the center. To do so, we use the size of the screen to measure the maximal eccentricities values (i.e. the distance, in degrees of visual angle, from the center of the screen to the edges), and we derive from it the visual acuity, expressed in the Snellen decimal [21]. We can then derive the sigma for the blur necessary to approximate this acuity, 10 pixels in our case.

Finally, the viewing time was set at 4.5 seconds. Between each image, participants were asked to position their mouse cursor on a target at the center of the screen.

Participants: The dataset was split into two playlists, of 50 images each. The 60 participants were split into two group, each group viewing one of the two playlists. Participants were asked to explore freely each stimulus, without any specific task. This experiment was done in a laboratory setting, under supervision, in order to avoid spammers or outliers. The adjustment of this protocol for crowdsourcing settings, i.e. the design of efficient filters to ensure trustworthy data, in a context where the experiment is proposed to a large panel of online participants, will be explored in future works.

Data processing: To infer visual saliency maps from the collected BubbleView data, we followed the work in [17]. For

each pixel of a stimulus, we compute the total aggregated time during which the mouse cursor was at this specific location. The resulting map was then convolved with a 2D Gaussian kernel, emulating the size of the fovea, i.e. 1° of visual angle. Similarly to the eye-tracking saliency maps, a mask was applied and dilated to only account for the surface of the object. Figure 2 shows an example of this process, for both eye-tracking and BubbleView data.

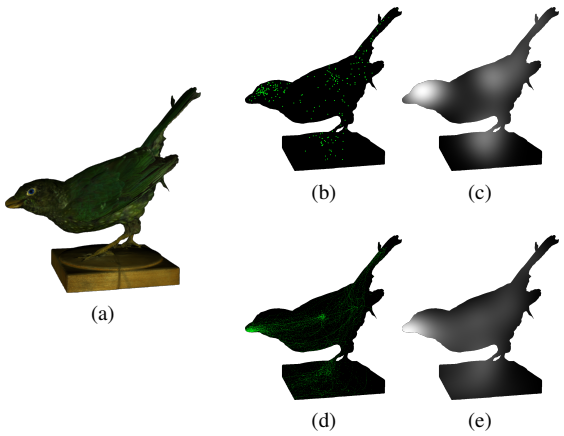

Fig. 2: Example of eye-tracking and BubbleView tracks transformed into saliency maps. (a) Original stimulus, (b) Eye fixation points, (c) Ground-truth visual saliency map, (d) BubbleView tracks, (e) BubbleView saliency map.

3. EXPERIMENTS

In this section, we propose two different evaluations of the BubbleView metaphor for 3D graphical objects. First, we evaluate how similar the saliency maps generated with the BubbleView paradigm are to the eye-tracking ground-truth; then, we evaluate and compare how visual saliency models trained with BubbleView or eye-tracking perform.

3.1. Similarity of the BubbleView metaphor with eye-tracking

In order to assess how close BubbleView saliency maps are to the eye-tracking ground-truth, we relied on four metrics commonly used in the context of visual saliency [22]: Pearson's correlation coefficient (CC), Normalized Scanpath Saliency (NSS), Kullback-Leilbler Divergence (KLD) and Similarity (SIM).

As recommended in [9], we also compared these scores as a percentage of a baseline score. To compute this baseline for the NSS, we used the inter-observer congruency, as described in [23]: for each observer, we create a fixation map based only on their fixations, and compared it to the saliency map created by aggregating all of the other observers. Scores were then averaged over observers and images. For CC, KLD and SIM metrics, the IOC baseline is computed by randomly spliting the observers into two groups, and comparing the saliency maps of both groups. The percentage of the IOC is then given by normalizing the scores with the baselines [9].

Table 1: Average similarity scores ($\pm$ standard deviation) between BubbleView saliency maps and ground-truth eye-tracking data. First line is the score between the BubbleView map and the eye-tracking ground-truth, and second line is the same score as a percentage of the IOC.

CC ↑	NSS ↑	KLD ↓	SIM ↑
0.90 ± 0.05	2.90 ± 0.57	0.19 ± 0.09	0.79 ± 0.05
79.6%	86.2%	73.9%	81.7%

Table 1 shows these similarity scores. Overall, we observe very high similarity compared to scores usually obtained on natural images [9, 17]. This can be explained by two factors: the simplicity of the considered stimuli, which do not present any background to explore, thus focusing the mouse of the participant on the content, and the relatively small size of the objects, allowing for a fast first exploration, and time to come back to interesting or salient areas afterwards.

Inter-observer congruency (IOC) is a measure of the similarity of gaze patterns between several observers watching at the same stimulus. In the context of 3D graphical objects and QoE, inter-observer visual congruency is a particularly useful metric. Indeed, it can be used to characterize the visual attention complexity of a rendered view of a 3D object [6], and such views can then be compared together, for instance to infer a preferred point of view.

Therefore, we are interested in knowing whether or not the BubbleView metaphor preserves this congruency. As explained in the previous section, we can compute IOC scores using a leave-one-out approach, comparing each observer with the aggregation of the others. However, while this approach can be replicated on BubbleView tracks, the comparison with eye-tracking scores might be uncertain due to the different nature of data. Since our work here is focused on visual saliency, we use a heuristic by computing Shannon's entropy of the saliency map. This way, when entropy is high, salient areas are dispersed over the whole surface of the object, and thus IOC scores are low, and vice versa.

We computed the entropy of both BubbleView and eye-tracking saliency maps for each view of the objects, and compared them together. We obtain a Pearson's correlation coefficient between entropies of 0.886 ($p - value \ll 10^{-5}$), showing that the BubbleView metaphor preserves relatively well information about the dispersion of gaze tracks on 3D graphical objects.

3.2. Visual saliency models

In order to evaluate the usefulness of BubbleView data for training visual saliency models on 3D graphical content, we selected and retrained three models that are representative of the state-of-the-art: MSINet [24], SAM-Resnet [25] and EML-Net [26].

Those three models were evaluated in three different configurations: using their base weights (i.e. pretrained on the SALICON dataset [18]), fine-tuned on the eye-tracking data, and fined-tuned using the BubbleView data. In each of the fine-tuning conditions, 85 images were randomly chosen to fine-tune the model, and the 15 left were used for validation. Each training was repeated 10 times, with a new random split each time, and the performances averaged, to ensure consistency in the results. The images were resized to fit the original input size of each model, using padding to conserve the aspect ratio. For the SAM-Resnet model, considering the high number of parameters of this model, and the low number of stimuli, we froze the weights of the Resnet-50 feature extractor.

To test these models, we used the 84 images of the 3DGC eye-tracking dataset [19]. As described in Section 2.2, we apply a dilated mask to only keep the saliency values on the surface and on the border of the object.

The predicted saliency maps are then scored against the eye-tracking ground-truth using the same four metrics as in Section 3.1. The results for each metric, model and training setting are summarized in Table 2.

Overall, EML-Net gives the best results, while SAM-Resnet exhibits the lowest performances. All of the models, when used in their basic settings, show relatively poor performances (under 1.5 in NSS, and under 0.7 in CC). This is easily explained by the nature of the SALICON dataset, which is composed of natural images, and not 3D graphical content.

As expected, fine-tuning on the eye-tracking dataset shows a significant improvement in performances compared to the base model, even with the low amount of training data that we used. This procedure allows the models to adjust their weights to the specific features of the considered content, but also to the properties of the ground-truth saliency maps, which can slightly differ due to the diversity of experimental conditions in which eye-tracking data are collected.

Interestingly, we observe that for the three models, fine-tuning on the BubbleView dataset also significantly improves the performances, although not as much as the eye-tracking data. Nonetheless, it seems that BubbleView data can be successfully used to improve the prediction of visual saliency on 3D graphical objects.

Similarly to Section 3.1, we also evaluated the performances of the models in each setting to predict the visual attention complexity. We compared the entropy of the predicted saliency map to the ground truth using Pearson's correlation coefficient. Table 3 compiles these scores. The EML-Net model shows the highest correlation, especially when fine-tuned using eye-tracking data ($CC = 0.88$). We also observe a systematic improvement of the correlation when any kind of fine-tuning is used, including with the BubbleView data.

Table 2: Scores of the three visual saliency models on the test dataset. Best performances are bolded. For each model, three settings are compared: base, BubbleView fine-tuned (BV) and eye-tracking fine-tuned (ET).

		CC ↑	NSS ↑	KLD ↓	SIM ↑
SAM-Resnet	Base	0.582	1.023	0.527	0.621
	BV	0.679	1.215	0.463	0.632
	ET	0.711	1.320	0.401	0.649
MSINet	Base	0.641	1.289	0.368	0.633
	BV	0.690	1.587	0.303	0.650
	ET	0.724	1.613	0.287	0.675
EML-Net	Base	0.684	1.432	0.239	0.645
	BV	0.741	1.721	0.207	0.678
	ET	**0.758**	**1.932**	**0.196**	**0.691**

Table 3: Pearson's correlation coefficient between the entropies of ground-truth and predicted saliency maps, for base model (B), BubbleView (BV) and eye-tracking (ET) fine-tuned.

SAM-Resnet			MSI-Net			EML-Net		
B	BV	ET	B	BV	ET	B	BV	ET
0.52	0.63	0.68	0.75	0.81	0.85	0.82	0.86	**0.88**

4. CONCLUSION

In this work, we explored the BubbleView metaphor's ability to be a reliable proxy for gathering visual attention data on 3D graphical objects. Based on attention data gathered with both eye-tracking and BubbleView experiments, we showed strong similarities between both collecting methods in the considered features, namely the visual saliency and the visual attention complexity. It should be noted that 3D graphical objects, due to their nature, are particularly good candidates for the BubbleView metaphor: indeed, their simplicity and the absence of background allows the observers to explore the stimuli very quickly and easily, which is not always the case with natural images, for instance. We also showed that BubbleView data can be reliably used to fine-tune visual saliency models in this context, and can improve the prediction of visual attention complexity.

Considering the high cost of eye-tracking experiments, it appears that BubbleView can be a good cost-effective alternative for 3D graphical content. However, it remains to evaluate how this metaphor would perform outside of the lab, with different devices and an increased probability of spammers.

5. REFERENCES

[1] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *IEEE ICIP*, 2013, pp. 196–200.

[2] S. Zhu, Q. Chang, and Q. Li, "Video saliency aware intelligent hd video compression with the improvement of visual quality and the reduction of coding complexity," *Neural Computing and Applications*, vol. 34, no. 1, pp. 7955–7974, 2022.

[3] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE PAMI*, vol. 43, no. 02, pp. 679–700, 2021.

[4] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 659–666, 2005.

[5] R. Song, Y. Liu, R. R. Martin, and P. L. Rosin, "Mesh saliency via spectral processing," *ACM Trans. Graph.*, vol. 33, no. 1, 2014.

[6] M. Abid, M. Perreira Da Silva, and P. Le Callet, "Towards visual saliency computation on 3d graphical contents for interactive visualization," in *IEEE ICIP*, 2020, pp. 3448–3452.

[7] O. Akman and P. Jonker, "Computing saliency map from spatial information in point cloud data," in *Advanced Concepts for Intelligent Vision Systems*. 2010, pp. 290–299, Springer Berlin Heidelberg.

[8] X. Ding, W. Lin, Z. Chen, and X. Zhang, "Point cloud saliency detection by local and global feature fusion," *IEEE TIP*, vol. 28, no. 11, pp. 5379–5393, 2019.

[9] A. Newman, B. McNamara, C. Fosco, Y. B. Zhang, P. Sukhum, M. Tancik, N. W. Kim, and Z. Bylinskii, "Turkeyes: A web-based toolbox for crowdsourcing attention data," in *ACM CHI*, 2020, p. 1–13.

[10] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *ACM ETRA*, 2006, p. 87–94.

[11] P. O'Donovan, A. Agarwala, and A. Hertzmann, "Learning layouts for single-page graphic designs," *IEEE TVCG*, vol. 20, no. 8, pp. 1200–1213, 2014.

[12] S. Cheng, Z. Sun, X. Ma, J. L. Forlizzi, S. E. Hudson, and A. Dey, "Social eye tracking: Gaze recall with online crowds," in *ACM CSCW*, 2015, p. 454–463.

[13] A. Carlier, A. Shafiei, J. Badie, S. Bensiali, and W. T. Ooi, "Cozi: Crowdsourced and content-based zoomable video player," in *ACM MM*, 2011, p. 829–830.

[14] K. Rodden, X. Fu, A. Aula, and I. Spiro, "Eye-mouse coordination patterns on web search results pages," in *ACM CHI*, 2008, p. 2997–3002.

[15] J. Huang, R. White, and G. Buscher, "User see, user point: Gaze and cursor alignment in web search," in *ACM CHI*, 2012, p. 1341–1350.

[16] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister, "Bubbleview: An interface for crowdsourcing image importance maps and tracking visual attention," *ACM Trans. Comput.-Hum. Interact.*, vol. 24, no. 5, 2017.

[17] W. Ellahi, T. Vigier, and P. Le Callet, "Evaluation of the bubble view metaphor for the crowdsourcing study of visual attention deployment in tone-mapped images," in *IEEE EUVIP*, 2021, pp. 1–6.

[18] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *IEEE CVPR*, 2015, pp. 1072–1080.

[19] M. Abid, M. Perreira Da Silva, and P. Le Callet, "Perceptual characterization of 3d graphical contents based on attention complexity measures," in *ACM QoEVMA workshop*, 2020.

[20] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *ETRA*, 2000, p. 71–78.

[21] G Westheimer, "Visual acuity. chapter 17," *Adler's Physiology of the Eye, Clinical Application. St. Louis: The CV Mosby Company*, 1987.

[22] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE PAMI*, vol. 41, no. 3, pp. 740–757, 2019.

[23] A. Bruckert, Y. H. Lam, M. Christie, and O. Le Meur, "Deep learning for inter-observer congruency prediction," in *IEEE ICIP*, 2019, pp. 3766–3770.

[24] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder-decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.

[25] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," *IEEE TIP*, vol. 27, no. 10, pp. 5142–5154, 2018.

[26] S. Jia and N. D. B. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," *Image and Vision Computing*, vol. 95, pp. 103887, 2020.