# INFLUENCE OF VIEWPOINT ON VISUAL SALIENCY MODELS FOR VOLUMETRIC CONTENT

*Mona Abid, Matthieu Perreira Da Silva, Patrick Le Callet*

University of Nantes, LS2N, CNRS, UMR 6004, Nantes, France.

## ABSTRACT

In order to predict where humans look in a 3D immersive environment, saliency can be computed using either 3D saliency models or view-based approaches (2D projection). In fact, building a 3D complete model is still a challenging task that is not investigated enough in the research field while 2D imaging approaches have been extensively studied and have shown solid performances.

As 6 degrees of freedom are allowed in volumetric videos, users are able to navigate through the content in different manners. In this case, 2D saliency models might be less robust if applied naïvely, since advanced parameters such as viewing distance are not considered in such models.

The aim of this paper is to investigate the influence of viewpoint on 2D saliency models when applied on volumetric data and this to get a better understanding of how viewpoint information could be integrated into view-based approaches.

To do so, a subjective psycho-visual experiment was conducted and a fine analysis was led using the variance analysis statistical method.

***Index Terms*—** Visual attention models, volumetric data, view-based approach, perception, computer vision.

## 1. INTRODUCTION

With the emergence of volumetric videos (viewing 3D environment with 6 degrees of freedom), understanding human visual attention mechanisms and interaction in immersive scenes are of great importance in perception. Since 6 degrees of freedom are allowed in volumetric videos, users are able to navigate through the content in different manners. Therefore, viewpoint (depending on both visualization distance and viewing angle) is more likely to change according to the user's interaction in virtual scenes. When looking at a scene with no specific task, humans do not focus on each region of the image with the same intensity [1]. In fact, attentive mechanisms guide their gazes on salient and relevant parts. In order to predict where humans look in a 3D environment, saliency can be computed. As defined in [2], visual saliency is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention. In other words, our attention is attracted to visually salient stimuli.

Being an active research area in the computer vision community, visual saliency modeling aims to predict human fixations as a way to detect the regions attracting the human gaze [3]. This modeling not only gives an insight into the complex human visual system but also shows much potential in the wide range of applications using computational saliency such as image object segmentation [4], object recognition [5], video compression [6], tracking [7], etc. Since the few existing 3D models consider geometry information only without texture or shading [8], applying them in an immersive environment is very restricted (because of the lack of texture for example). On the other hand, several promising 2D models that showed high performances could be applied by considering 2D projection views of 3D data, rendered by a specific rule. Investigating the impact of viewing distance when considering volumetric data could help us understand the influence of such parameter on the human gaze and therefore give us an insight of how to integrate this parameter in view-based models to adapt the latter for immersive imaging.

## 2. RELATED WORK

Although visual saliency concept firstly arose in psychological and neurobiological context, it generated a noticeable interest in neuroscience and computer vision communities. One of the earliest models was proposed by Itti et al [9]. It is an implementation of general computational frameworks and psychological theories of bottom-up attention based on center-surround mechanisms. Saliency models could be divided into 2 main categories: conventional models [1] and deep models [10]. In fact, conventional saliency prediction methods define features that capture low-level cues such as color, contrast, intensity, edge, orientation, and texture or semantic concepts such as faces, people, text, etc. Whereas, deep models are based on automatic hierarchical features extraction and end-to-end task learning. Thanks to annotated datasets that are publicly available and to the development of different deep learning architectures, the imitation of the selective human visual system have been progressing for the past years. Compared with conventional saliency models,

This work was funded by the French National Research Agency as part of ANR-PISCo project (ANR-17-CE33-0005).

deep saliency models achieved much higher performances when dealing with human eye fixation prediction [10].
Since the user's interaction is possible in volumetric videos, the viewing distance from which a 3D object is seen is more likely to change. In this case, although advanced 2D saliency approaches showed impressive results in 2D imaging, they might be impaired if applied naïvely. In fact, the viewing information is not included as a parameter either in the annotated datasets (used to train models) such as MIT300 [11], CAT 2000 [12] that were established for the MIT saliency Benchmark ($saliency.mit.edu$) or in the computational model itself.

## 3. EXPERIMENTAL METHODOLOGY

In order to evaluate how much computational saliency prediction models and human saliency are aligned, a subjective psycho-visual study was conducted. The main purpose is to investigate how current state-of-the-art saliency models perform when varying the visual angle. For the sake of simplification, a static environment is considered. The reason behind this simplification is the lack of dynamic saliency models and the large variation of gaze data. In fact, observers tend to agree less on dynamic scenes when collecting ground truth data [8]. One can interact with the static environment by moving forward, backward. This generates a zoom-in, zoom-out leading to 3 main variations: A change in the level of details, the presence of occlusions and a change in the viewing distance. Given different 3D objects, an eye-tracking experiment was conducted. The details of stimuli generation, experiment setup and data collection are presented in the following parts.

### 3.1. Data collection - Stimuli generation

Our dataset aims to provide comprehensive and diverse coverage of objects (visual angles) for eye-tracking analysis. Stimuli are generated from several 3D objects by varying the visual angle parameter which is directly related to the viewing distance. This leads to a change in the level of detail and presence of occlusion. Objects sizes are quite different and their content is quite diverse: richness, brightness, and complexity of the contents. Moreover, their inherent representation is different. This means that some objects are modeled by textured meshes others are modeled by colored vertices (mean number of vertices per object is equal to 500k). The visualization and the manipulation of the different objects used were done using Unity software which is a game engine that is used to create both 2D and 3D games as well as to produce computer simulations. To capture the rendering of the 3D objects, one plan projection was considered as well as 3 scales for each object. The stimulus scale variation is illustrated in figure 2. A semantic definition to differentiate the 3 scales of a given object was introduced. The biggest scale of a given object is defined as the maximal vertical occupancy of the rendering on the monitor screen which resolution is $1920 \times 1080$ and also by making sure that the whole content of the rendered image remains visible.

### 3.2. Experiment Design

In order to design the psycho-visual experiment, dedicated software for eye-tracking experiments, (Experiment builder; SR research) was used. Based on the MIT Saliency Benchmark protocol [13, 14], every stimulus is displayed for 3 seconds. To make sure that one observer sees each stimulus only once, 2 sessions were created when designing the experimental protocol. One observer can attend one session only to minimize the "memory effect" that could bias the eye-tracking experiment. Each session contains 3 series of 12 stimuli that are randomized. Before each series, a calibration of the eye-tracker is applied.

### 3.3. Material setup and subjective test

#### *3.3.1. Eye-tracker characteristics and configuration*

In order to conduct the subjective experiment and collect subjects eye gaze data, the eye-tracker *EyeLink 1000 Plus* (Desktop System) by SR research was used. It has a sampling rate of $1000Hz$, allows tracking both eyes simultaneously and reports an average accuracy between $0.25°$ and $0.5°$ under recommended conditions. The only distance requested during the eye-tracker configuration is the distance between the screen and the lens which was set to $430$ *mm* in this experiment. The distance between the observer and the screen is around $1000$ *mm* leading to a pixel density value of $66$ *pixel/degree*.

#### *3.3.2. Participant setup and gaze data acquisition*

In remote tracking mode, a target sticker is put on the forehead of the participants so that head movements can be compensated during tracking. Stimuli were displayed on a monitor with a refresh rate of $60Hz$. The distance between the observer and the eye-tracker was defined in such a way as to guarantee an accurate recording while also ensuring comfortable viewing for the observer. This distance range is $[550mm - 600mm]$.
30 observers (university students, aged between 20 and 24) volunteered to participate to this experiment. All participants had normal/ corrected-to-normal visual acuity and normal color vision. The conducted experiment was based on a free viewing task. A calibration step was performed for each observer before each session. It consists of presenting 13 calibration points in a spherical shape followed by a validation step.

## 4. ANALYSIS PROTOCOL AND RESULTS

### 4.1. Data Analysis

Data was recorded via the eye-tracker every $1$ *ms*. It includes saccades, fixations, and blinks. In order to detect saccades,

three thresholds were used: motion (°), velocity ($°/sec$), and acceleration ($°/sec^2$). The saccadic motion threshold is used to delay the onset of a saccade until the eye has moved significantly. A velocity threshold of $22°/sec$ allows detection of saccades as small as $0.3°$. Acceleration data has a threshold of $8000°/sec^2$ as recommended for cognitive research. In order the get the saliency map corresponding to every stimulus, a Gaussian distribution having a standard deviation corresponding to 1 degree of visual angle was applied on each fixation point. For each considered saliency model, 72 saliency maps were computed (24 objects $\times$ 3 views). Moreover, 72 human fixation maps and 72 ground truth saliency maps were generated from the acquired gaze data. With such data, saliency models performances were investigated considering different viewing distances. In order to show the overall differences between semantic scales, medians and standard deviation were computed according to the vertical visual angle using box plots as illustrated in figure 1.

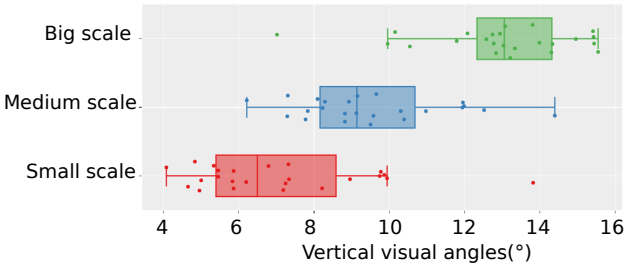



**Fig. 1**. Semantic scales box plot along vertical visual angles.

### 4.2. Results

#### *4.2.1. Evaluation metrics and post-processing*

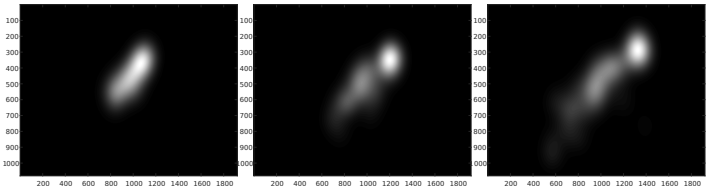
Many techniques exist to measure the agreement between model prediction and human eye fixations. But it is hard to achieve a fair comparison for saliency models by one single metric since every metric provides a slightly different information [13]. Among widely accepted and standard metrics for saliency evaluation, a quantitative experiment was carried out by considering a variety of metrics such as Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), Area under the ROC curve (AUC) in its different variants of AUC-Judd and AUC-Borji, and Kullback-Leibler Divergence (KLD). These different metrics can be split into 2 different categories. The main difference between them concerns ground-truth representation. Metrics such as CC, SIM, KLD could be classified as distribution based since they use the saliency map as ground-truth, others such as NSS, AUC-Judd, and AUC-Borji could be classified as location-based categories since they use the fixation map.

Before using different metrics to evaluate the models, an important step was integrated in order to make sure that there is no bias introduced by the size of the image content. In fact, as some metrics consider the number of salient and non-salient points (typically metrics using ROC) and in order to have a fair comparison between different object scales, cropping was applied on the inputs before applying different metrics. The used bounding box changes from one content to another based on the window size ensuring the whole image content wrapping as illustrated in figure 2 where the bounding box is plotted in red. Moreover, in order to decrease objects location impact when looking at different stimuli, the rendered objects are always screen-centered as shown in figure 2.

Small scale Medium scale Big scale

**Fig. 2**. Bounding boxes applied on the content of one 3D object

Once ground truth fixation maps and saliency maps generated, different computational models were applied on all stimuli to get their respective saliency maps. Afterwards, different metrics were evaluated for each stimulus on its cropped window. Figure 3 illustrates the ground truth saliency map of the stimuli presented in figure 2.



Small scale Medium scale Big scale

**Fig. 3**. Ground truth saliency maps of one 3D object in 3 scales

#### *4.2.2. Experimental results and interpretation*

In order to study the impact of the rendered object size, some top performance saliency models [15, 16] were selected among many existing ones. For the following metrics: NSS, AUC Judd, AUC Borji, CC and SIM, the higher the score value is the closer the computational model compared to the ground truth is. Contrary to KLD metric that computes the divergence between the saliency model and the ground truth. Therefore the lower the value of KLD is the better performance the model has. An overall analysis was firstly conducted in order to get an idea of how scores are spread out for a given model and a given metric along with image content size. Table 1 synthesizes the mean and the standard deviation score values according to 3 groups of scales (Small, Medium, Big) when applied on 24 stimuli for a given model and a given metric. In table 1, only 3 most used metrics [17] are presented for the sake of simplification. Such preliminary

analysis shows that the CC of the Salicon model is higher than the CC of both versions of SAM model for all scales. However, the standard deviation is quite high which means that score values are quite dispersed. Moreover, a closer look into different score values in table 1 shows that CC metric has very close mean values for the 3 different scales. Such behavior could be explained by the fact that the used correlation coefficient is the Pearson Correlation Coefficient (PCC) which is a linear correlation between two variables. One can assume that Spearman correlation which measures the non-parametric statistical dependency could be more suitable. Overall, these observations motivate the necessity of a finer analysis.

| Saliency model | Metrics | Mean ± Standard deviation | | |
|---|---|---|---|---|
| | | Small | Medium | Big |
| Salicon | CC ↑ | 0.64 ± 0.23 | 0.65 ± 0.24 | 0.66 ± 0.20 |
| | NSS ↑ | 0.89 ± 0.44 | 1.04 ± 0.50 | 1.25± 0.53 |
| | KLD ↓ | 0.28 ± 0.16 | 0.36 ± 0.28 | 0.54 ± 0.42 |
| SAM-Vgg | CC ↑ | 0.34 ± 0.29 | 0.32 ± 0.20 | 0.36 ± 0.17 |
| | NSS ↑ | 0.48 ± 0.25 | 0.56 ± 0.29 | 0.73± 0.27 |
| | KLD ↓ | 0.69 ± 0.37 | 0.89 ± 0.35 | 1.18 ± 0.56 |
| SAM-Resnet | CC ↑ | 0.41 ± 0.29 | 0.40 ± 0.24 | 0.43 ± 0.20 |
| | NSS ↑ | 0.58 ± 0.37 | 0.70 ± 0.36 | 0.83± 0.37 |
| | KLD ↓ | 0.54 ± 0.35 | 0.75 ± 0.36 | 1.06 ± 0.61 |

**Table 1**. Metrics evaluation for different saliency models

As already mentioned, the preliminary analysis is not sufficient to conduct a consistent comparison therefore an analysis of variance (ANOVA) was led. It is a collection of statistical models used to evaluate the dependency of a quantitative variables with qualitative variables. The aim of the applied ANOVA method is to evaluate if the studied dependency between a given model and the image size content is statistically significant or not. ANOVA can determine whether the means of the three scaling groups are different. According to the computed p-values for a given model and different metrics, it is possible to explain information diversity when considering different viewing distances and this by investigating the statistical significance. Table 2 synthesizes the outputs of the ANOVA including p-values and f ratios. The former indicates if there is a statistical significance between group means and the latter determines whether the variability between group means is larger than the variability of the observations within the groups. The main difference between p-value and f ratio is that they are inversely proportional and that p-value is a probability, while the f ratio is a statistical test.

| Saliency model | Metrics | p-value | Statistical significance | f ratio |
|---|---|---|---|---|
| Salicon | CC | 0.9680 | ✗ | – |
| | NSS | 0.0461 | ✓ | 3.22 |
| | KLD | 0.0134 | ✓ | 4.59 |
| SAM-Vgg | CC | 0.8715 | ✗ | – |
| | NSS | 0.0062 | ✓ | 5.47 |
| | KLD | 0.0011 | ✓ | 7.54 |
| SAM-Resnet | CC | 0.9381 | ✗ | – |
| | NSS | 0.0670 | ✗ | – |
| | KLD | 0.0008 | ✓ | 7.91 |

**Table 2**. Overall ANOVA output values for different metrics

As shown in table 2, some metrics have no significant statistical difference for all models such as CC metric whereas some others have a significant statistical difference for all models such as KLD metric. Concerning the NSS metric, statistical significance depends on the saliency model. Overall analysis showed that for the smaller scale of image rendering, saliency models have slightly better scores when compared to the different ground-truth images. This gives information about the sensitivity of the saliency models to the stimuli scales. Further analysis should be led in order to rank models according to different scales. It is important to mention that the preliminary analysis and the ANOVA are complementary.

## 5. CONCLUSION AND FUTURE WORK

Lately, human interaction in 3D immersive scenes has been an active research topic. In fact, many efforts have been made in order to understand and predict where humans look in such volumetric scenes. Depending on the viewing distance, the viewpoint is more likely to change according to the users interaction in the virtual scene. In this paper, a subjective study of the perceptual effect of the viewpoint was conducted.
A first investigation of the impact of visual angle on saliency models showed that a fine grain analysis should be conducted in order to ensure that the interpretation of different results is consistent. In fact, the overall analyses showed that for the smaller scale of image rendering saliency models have slightly better scores when compared to the different ground-truth images. This behavior should be reassessed by conducting complementary analyses.
Furthermore, pair comparisons could be applied as well as linear mixed models to take into consideration the variety between stimuli. For future work, in addition to different analyses, another aspect of visual attention should be investigated in immersive imaging. Apart from the viewing distance, the impact of the viewing angle should also be studied.

## 6. REFERENCES


[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[2] L. Itti, "Visual salience," *Scholarpedia*, vol. 2, no. 9, p. 3327, 2007.

[3] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[4] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5038–5047.

[5] Y. Ma, B. Zheng, Y. Guo, Y. Lei, and J. Zhang, "Boosting multi-view convolutional neural networks for 3d object recognition via view saliency," in *Chinese Conference on Image and Graphics Technologies*. Springer, 2017, pp. 199–209.

[6] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.

[7] C. Aytekin, F. Cricri, and E. Aksu, "Saliency-enhanced robust visual tracking," 02 2018.

[8] G. Lavoué, F. Cordier, H. Seo, and M.-C. Larabi, "Visual attention for rendered 3d shapes," *Comput. Graph. Forum*, vol. 37, pp. 191–203, 2018.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[10] A. Borji, "Saliency prediction in the deep learning era: An empirical investigation," *arXiv preprint arXiv:1810.03716*, 2018.

[11] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[12] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

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

[14] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[15] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080.

[16] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[17] O. Le Meur, D. Barba, P. Le Callet, and D. Thoreau, "A human visual model-based approach of the visual attention and performance evaluation," in *Human Vision and Electronic Imaging X*, vol. 5666. International Society for Optics and Photonics, 2005, pp. 258–268.