How visual attention is modified by disparities and textures changes?

Dar'ya Khaustova, Jérome Fournier, Emmanuel Wyckens – France Télécom R&D France;
Olivier Le Meur – University of Rennes 1, France

ABSTRACT

The 3D image/video quality of experience is a multidimensional concept that depends on 2D image quality, depth quantity and visual comfort. The relationship between these parameters is not yet clearly defined. From this perspective, we aim to understand how texture complexity, depth quantity and visual comfort influence the way people observe 3D content in comparison with 2D. Six scenes with different structural parameters were generated using Blender software. For these six scenes, the following parameters were modified: texture complexity and the amount of depth changing the camera baseline and the convergence distance at the shooting side. Our study was conducted using an eye-tracker and a 3DTV display. During the eye-tracking experiment, each observer freely examined images with different depth levels and texture complexities. To avoid memory bias, we ensured that each observer had only seen scene content once. Collected fixation data were used to build saliency maps and to analyze differences between 2D and 3D conditions. Our results show that the introduction of disparity shortened saccade length; however fixation durations remained unaffected. An analysis of the saliency maps did not reveal any differences between 2D and 3D conditions for the viewing duration of 20 s. When the whole period was divided into smaller intervals, we found that for the first 4 s the introduced disparity was conducive to the section of saliency regions. However, this contribution is quite minimal if the correlation between saliency maps is analyzed. Nevertheless, we did not find that discomfort (comfort) had any influence on visual attention. We believe that existing metrics and methods are depth insensitive and do not reveal such differences. Based on the analysis of heat maps and paired t-tests of inter-observer visual congruency values we deduced that the selected areas of interest depend on texture complexities.

Keywords: visual attention, eye tracking, texture, disparity, discomfort, IOVC

1. INTRODUCTION

Visual attention is the process of selecting of important areas of interest out of all the abundant visual information that humans receive in every day life [1]. The selection of these regions is done with the help of eye movements, used to assign processing information to some parts of the visual field. There are different kinds of eye movements: saccades and fixations. Saccades are quick eye movements shifting from one fixation location to another. Fixations are slow eye movements that direct a small part of the visual field into the fovea in order to accurately inspect the location of the stimulus.

There are two main mechanisms of visual attention: bottom-up (a stimulus-dependent mechanism) and top-down (an observer dependent mechanism). Bottom-up is driven by low-level features [2], [3] with eye movements that are involuntary and unconscious. Top-down attention integrates high-level cognitive processes like prior knowledge, task and experience [4].

The production of 3D content is more complicated than 2D; improper shooting can cause eye strain and visual fatigue. Thus, the production of visually comfortable stereoscopic content is fundamental to ensure the deployment of 3D cinema as well as 3DTV at home. To deal with these problems, a number of studies on the influence of 3D on the perception have recently been launched. Some studies explore how stereopsis influences our perception and whether it causes a change in gaze behavior while watching stereoscopic content. Then, we divide all recent studies into two groups based on the temporal features of the stimuli: the first group is studying visual attention with still images and the second with stereoscopic video content.

M.Wexler & N.Ouarti [5] investigated how various aspects of 3D scenes affect visual behavior. In their experiment they used three types of inclined surfaces (grid, texture and dots). They demonstrated that saccades tend to follow surface depth gradients and found that vergence is dominated only by binocular disparity. Jansen *et al*. [6] studied the influence of disparity on fixations and saccades in free viewing of 2D and 3D images of natural scenes, pink noise and white noise. An analysis was performed using the data from the left eye. They found that disparity information had an influence on basic eye movements, causing an increase in the number of fixations, a decrease of fixation duration over time (only for pink and white noise); and a shortening of saccade length over time. The saliency of mean luminance, luminance contrast and texture contrast was compatible across 2D and 3D stimuli. Mean disparity had a time dependent effect for 3D stimuli. Disparity contrast was elevated at fixated regions in 3D noise images but not in 3D natural scenes. They reported that participants fixated closer locations earlier than more distant locations in the image. Previous works were supplemented by Wismeijer *et al*. [7]. They investigated whether saccades are aligned with individual depth cues, or with a combination of depth cues. As well as M.Wexler & N.Ouarti also did research in this area by conducting experiments that presented an incline in depth surfaces, in which combined monocular perspective cues and binocular disparity cues specified different plane orientations with different degrees of small and large conflict between two sets of cues. They found out that the distributions of spontaneous saccade directions followed the same pattern of depth cue combination as perceived surface orientation: a weighted linear combination of cues for small conflicts, and cue dominance for large conflicts. By examining the relationship between vergence and depth cues, they reached the same conclusion as Posner[1], that vergence is dominated only by binocular disparity.

Ramasamy *et al*.[8] analyzed the feasibility of using eye tracking for stereoscopic filmmaking in order to identify elements that distract the audience from the flow of the movie. They analyzed the gaze patterns of one scene to identify regions of interest in the frames. They found that in a scene showing a long deep hallway, in the stereoscopic version gaze points were concentrated at the far end of the scene and in the 2D version they were more spread. Other work that used video clips for the exploration of gaze patterns was presented by Häkkinen *et al*. [9]. They showed observers 2D and 3D versions of four short video sequences with durations from 5 to 22 seconds. For the observers, the task was to compare these two versions and report which version was better. They reached the opposite conclusion from the previous study by Ramasamy *et al*. and reported that in the 3D version eye movements were more widely spread. The next study by Huynh-Thu *et al*. [10] did not show any evidence that fixations were more widespread when viewing 3D. They did not find strong evidence of the opposite either. It was reported that "the spread of fixations depended highly on the content characteristics and narrative flow of the video, and not only on the depth effect provided by the 3D stereoscopic version". During the experiment they showed observers 21 video sequences with various durations from 8 to 143 seconds in 2D and 3D modes. They found that average fixation frequency and average fixation duration were smaller when viewing 3D stereoscopic content; while the average saccade velocity was higher when viewing 3D stereoscopic content. Another study by J.Gautier & O. Le Meur [11] investigated the influence of disparity on saliency, as well as center and depth biases. Their results claim that visual exploration is affected by the introduction of the binocular disparity,i.e. the participants tend to first look at closer areas (in terms of depth) and then direct their gaze to more widespread locations. They also proposed a time-dependent computational model to predict saliency on still pictures. The proposed approach combined low-level visual features,along with center and depth biases.

Still, there has been no relation established between visual attention, scene complexity and the depth component. Another open question is whether discomfort sometimes generated by 3D content has an influence on the way we observe images. To this end, we examined visual attention and investigated potential differences in attention behavior between viewing still images with different levels of texture complexity and disparities. At first, we analyzed basic eye movement properties: the length of saccades and the duration of fixations. The temporal evolution of these parameters was studied over a viewing duration of 20 s. To investigate if the observers' visual exploration pattern changed due to the introduced parameters, we computed saliency maps (SMs) for each image.

2. MATERIALS AND METHODS

The experimental study consists of three steps: (1) Estimate the camera baseline and convergence distance for the creation of comfortable and uncomfortable viewing conditions; (2) Generate 3D synthetic scenes with three different amounts of depth and three levels of texture complexities; and (3) Perform a visual attention test, using an eye-tracker. During the experiment, a series of still images were viewed on a 3D LCD display using passive polarized glasses technology. As the purpose of the experiment is the observation of differences in visual attention and scanpaths for different scene parameters, each observer was able to view the images freely at his own discretion.

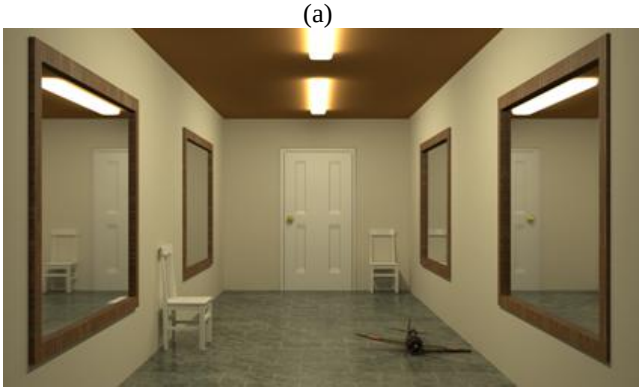
2.1. Stimulus generation

The amount of depth presented to observers is controlled by changing the virtual 3D camera parameters. The determination of camera parameters is done with the Stereo Calculator developed in Orange Labs, by proposing rules to define the shooting parameters based on the optimization of the stereoscopic distortion and the comfortable viewing zone [12]. To ensure visual comfort for watching 3DTV, Stereo Calculator uses optimization rules with the most conservative value ±0.2 diopters of depth of focus (DoF)[13, 14]. Thus, the software allows a precise calculation of the camera baseline and the convergence distance for the creation of comfortable and uncomfortable viewing conditions. The calculations can be accomplished when display parameters (image definition, size, viewing distance), camera parameters (focal length, sensor size), scene parameters (foreground, background distance of a scene, region of interest) and human visual attributes (depth of focus, inter-pupil baseline) are defined. Finally, each scene has three levels of depth: 2D, 3D comfortable and 3D uncomfortable. The necessary amount of depth was controlled by changing the camera baseline using the input parameter DoF=0.1 for comfortable viewing and DoF=0.3 for uncomfortable viewing. These thresholds were selected based on the former research done in Orange Labs [15]. For the sake of clarity DoF=0.1 is denoted as "01" in following tables and figures and DoF=0.3 as "03".

All scenes were generated and rendered using Blender software, which allows for the foreground and background distances of a scene to be measured and for the stereoscopic camera parameters to be controlled accurately. Six different scenes were selected for the experiment: "Bathroom", "Cartoon", "Hallway", "Kitchen", "Tea", and "Room"[16]. Each scene has three texture complexities: low, medium and high. Underscores after the scene name denote its texture complexity: lt – low texture complexity, mt – medium texture complexity, and ht –high texture complexity. In the experiment, low texture is the absence of a pattern on objects and low contrast (if it was possible). Simple geometrical patterns were selected for medium texture complexity and complex non-geometrical patterns were selected for high texture complexity. The contents of the scenes were mainly indoors as it is very difficult to find a suitable substitute for outdoor textures, such as leaves, grass, or sky when dealing with varying degrees of complexity. Some examples of the generated scenes are illustrated in figure 1.

(a)

(b)



(c)

(d)

(e) (f)

Figure 1. Examples of stimuli used in the experiment: (a) Bathroom_mt, (b) Cartoon_ht, (c) Hallway_lt, (d) Kitchen_ht, (e) Tea_mt, (f) Room_lt.

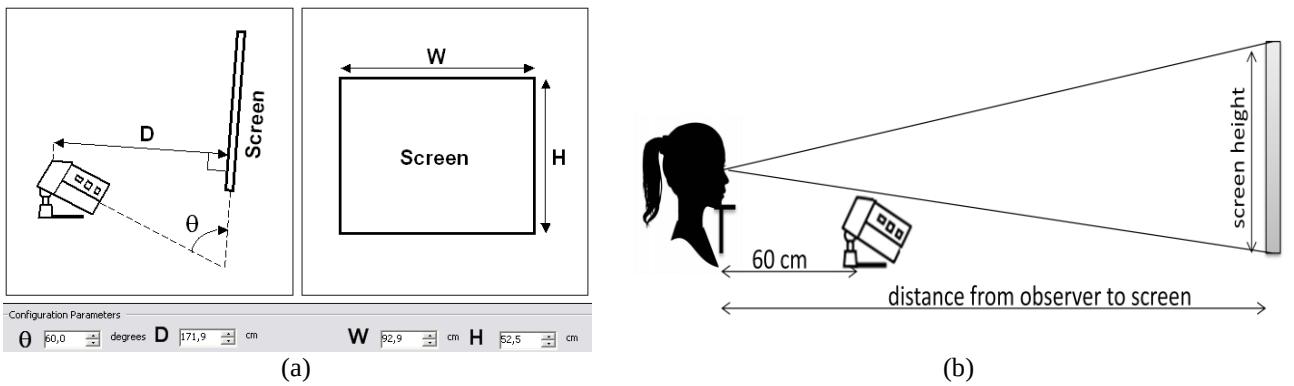
In total, we generated 54 images (6 images×3 depth levels×3 textures). Since we wanted to prevent the observers from memorizing the images and hence using top-down visual mechanisms, we formed 9 sets containing 6 images with different contents. Each set had two 2D images, 2 images with comfortable depth and 2 images with uncomfortable depth.

2.2. Experimental set-up

The psychophysical test set-up consisted of a Tobii x50 eye-tracker and 42" LG 42LW stereoscopic display with line interleaved technology. The physical dimensions of the screen are 93×52 cm. The resolution of the screen in 2D mode is 1920×1080 and in 3D stereoscopic mode 1920×540 per view. Images were displayed in a side-by-side format with a picture resolution of 1920×1080 using passive 3D glasses.

During tracking, the Tobii eye-tracker uses near infrared diodes to generate reflection patterns on the corneas of the eyes of the observer. These reflection patterns are collected by a camera and analyzed by Clear View software. Finally, it is possible to collect the gaze points on the screen, i.e., where the observer was looking. The Tobii x50 requires geometrical adjustments. For instance, the distance from the observer to the eye tracker should be around 60 cm; the eye tracker should be positioned straight in front of the stimuli and at a particular angle below the user. Head motion was restricted by using a chin rest. The hardware setup requires some measurements (Figure 2a), but once it has been configured for a particular setup, eye tracking is fully automatic. The duration of the experiment per observer was about 10 minutes.

The general scheme of the set-up of the experiment is presented in figure 2b. The distance from the observer to the screen was 4.5H (2.34 m). The distance from the observer to the eye tracker was 60 cm.



(a) (b)

Figure 2. Experimental set-up: (a) Tobii x50 configuration tool, (b) General scheme of the set-up of the experiment.

Taking into account the width of the screen (SW=93 cm) and the distance from the observer to the screen (SD=234 cm), we applied trigonometry rules and calculated the visual half angle as a basis.

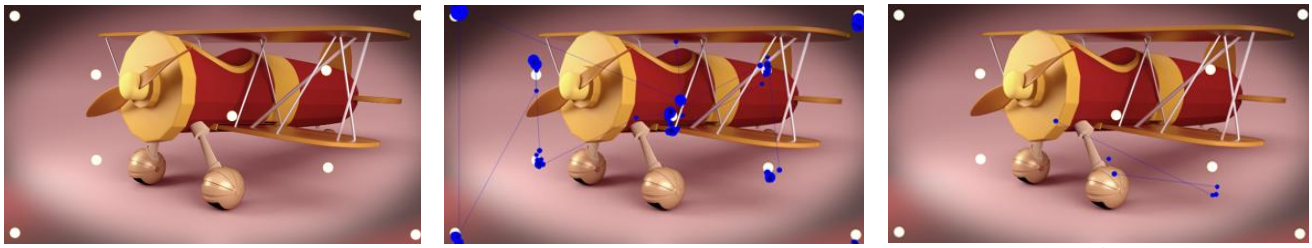
$$\tan\left(\frac{\alpha}{2}\right) = \frac{SW/2}{SD} \quad (1)$$

Therefore the number of pixels per one degree is $\frac{\text{horisontal resolution}}{\alpha} = \frac{1920}{22.47^\circ} \approx 85$ pixels.
This value was used for the calculation of saliency maps.

2.3. Experimental methodology

For each observer, the experiment consisted of five stages: a visual test, a reading of the instruction sheet, a calibration, a training period and, finally, the visual attention test. Monocular acuity, color vision, far vision test, fusion test and stereoscopic acuity of the observers were checked using an Essilor ERGOVISION machine prior to the visual attention test. The instruction sheet offered some explanations on how to behave during the calibration stage, the training stage and during the test itself. Each observer was tasked with simply observing the images freely during the test. After reading the instruction sheet, the observer was required to put on the passive polarized glasses in order to begin the calibration.

The eye tracker requires a calibration to learn the characteristics of the eyes of each observer. During the calibration stage, an observer simply looks at a dot that appears in different positions of the screen. The calibration procedure is fully automatic and takes about 30 seconds. A five-point calibration procedure was used in our experiment. Even if the software reports that the calibration was done successfully, there are still a few special circumstances in which the system has tracking difficulties, such as for people with bi-focal glasses or people with elements (eye lids, mascara, etc.) that significantly block the eye tracker camera's view of the subject's eyes. Thus, after the automatic calibration a specially designed chart was used in order to check if the device is able to track the observer's gaze correctly (Figure 3a). The calibration image contains nine white dots on the top of a picture of an airplane. Observers were instructed to focus on every white point for 3 seconds. The gaze plot indicates if the calibration process was accomplished successfully. An example of a successful calibration is presented in figure 3b. If the eye-tracker had difficulties properly recording the data for one observer, the resulting gaze plot resembled figure 3c. No observers with unsuccessful gaze plots were allowed to participate in the test.



(a) (b) (c)

Figure 3. Calibration phase: (a) Calibration chart, (b) successful gaze plot, (c) unsuccessful gaze plot.

The training was done in stereoscopic mode using three images with three levels of depth: 2D, DoF=0.1, DoF=0.3. Each image was presented for 20 seconds and separated from the subsequent one by displaying a gray screen for 5 seconds. The images were different from those used in the test. The training phase was designed to familiarize observers with the test conditions. The duration of the training was 1 min 20 s.

During the test we displayed one of the nine sets of images. The duration of the test was 2 min 40 s. 135 people (106 males and 39 females from 21 to 60 years old) participated in the test. Each image was tested on 15 observers.

3. EYE TRACKING DATA ANALYSIS

The analysis was accomplished by utilizing the fixation data of observers and saliency maps. The extra calibration process was used as a precaution: all observers who participated in the test had a successful calibration chart. Hence, all fixation data that was recorded by the Tobii eye tracker was usable. The first fixation of each stimuli presentation was not discarded, as stimuli were separated by a gray slide without a fixation cross in the center.

To test whether the introduction of disparity and texture has an effect on basic eye movement properties, we analyzed the duration of fixations and the length of saccades and compared them to state of the art results. The analysis of raw data of

visual fixations was done with software developed at IRISA by Olivier Le Meur [17, 18]. Then, to reveal a difference between exploration patterns and saliency maps for different levels of depth, AUC (Area Under Curve) and CC (linear coefficient) metrics were computed. The influence of texture complexity was determined by computing IOVC (Inter Observer Visual Congruency) criterion. More details are given in the following sections.

3.1. Analysis of exploration pattern and saliency maps

In order to compare gaze patterns for the same content in stereoscopic and non-stereoscopic conditions, many studies use the heat maps [8-10], which represent the areas fixated by observers. A heat map is a powerful way to visualize the gaze behavior of an entire group of observers for a particular image. The heat map consists of the stimuli as a background image (in our context it will be the left view) and a hotspot mask superimposed on top of it. A hotspot mask is a color scale between blue (no fixations) and red (highest number of fixations). In addition, the normalization of the heat map is done for each image using a color scale depending on fixation number, i.e., red indicates the highest number of fixations for a given image which can differ from the other images. As a consequence, it is difficult to compare heat maps for different scenes precisely.

For example, several heat maps are shown in figure 4 for the scenes “Hallway_mt”, “Kitchen_lt”, “Tea_ht”; each image on this figure shows the heat map corresponding to a viewing duration of 20 s. We only present 3 scenes here with 3 different texture complexities since the heat maps for the rest of the scenes are similar. Visual analysis of the heat maps corresponding to a viewing duration of 20 s did not reveal any differences in gaze patterns for conditions with different disparities or for different texture complexities. All nine heat maps (3 texture complexities×3 depth levels) for each scene looked qualitatively similar (Figure 5).

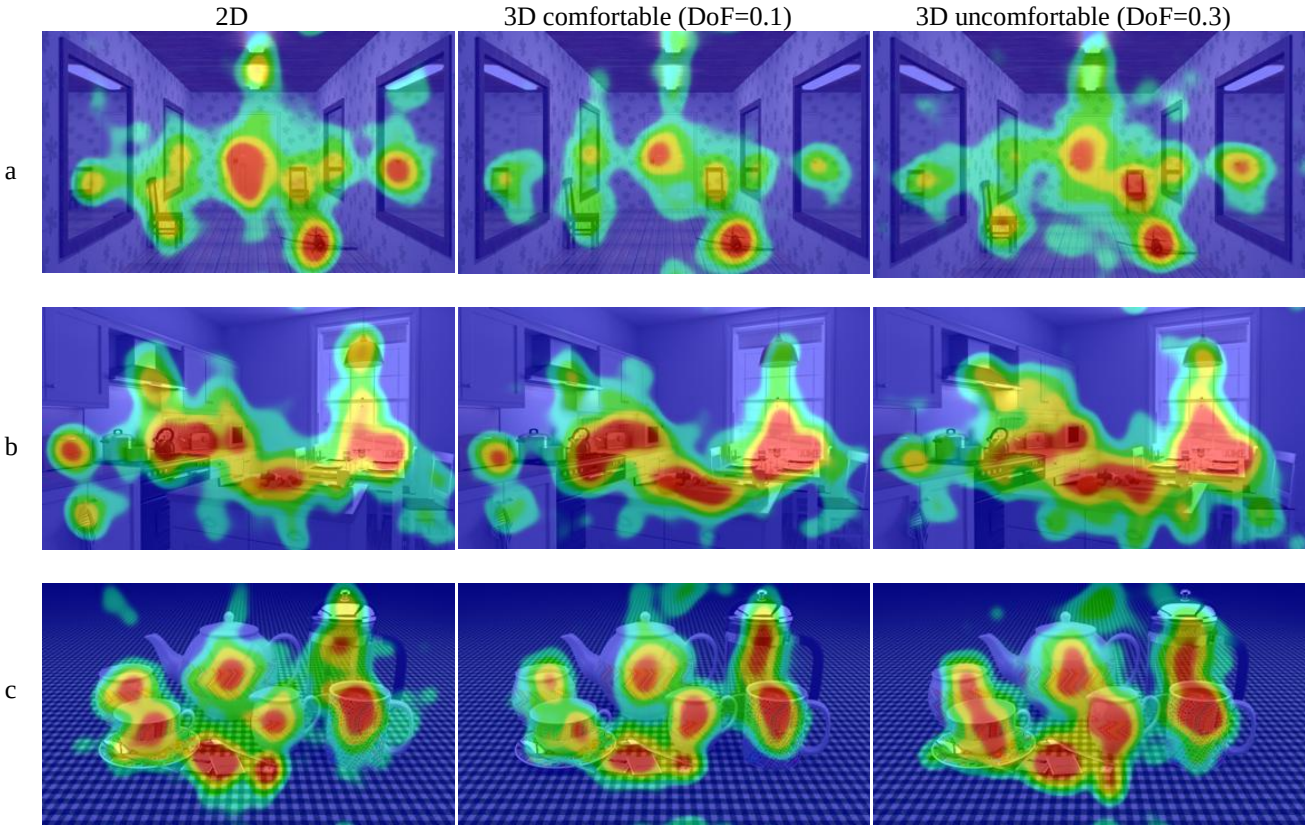


Figure 4. Heat maps for scenes (a) Hallway_mt, (b) Kitchen_lt, (c) Tea_ht

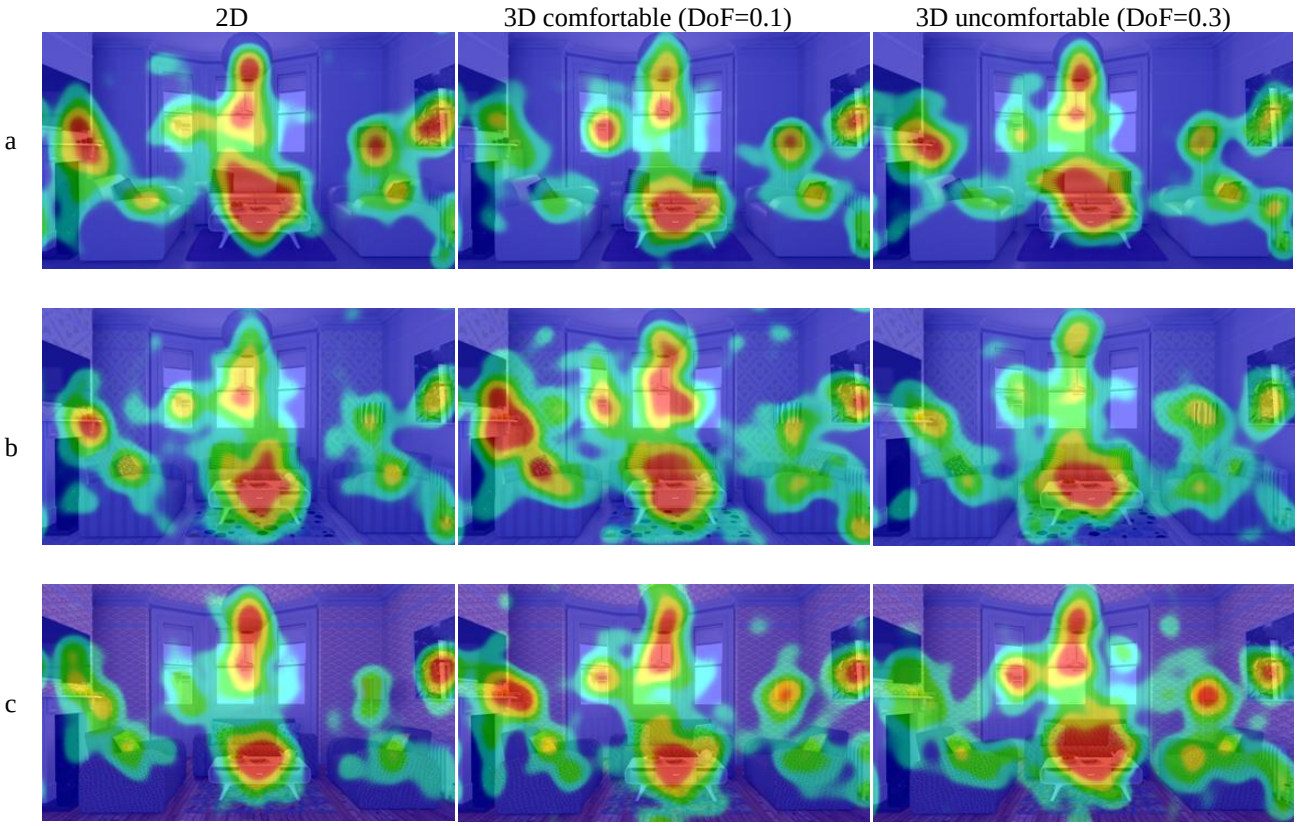


Figure 5. Heat maps for scenes (a) Room_lt, (b) Room_mt, (c) Room_ht

The correlation between the saliency maps for each scene with different depth levels was computed using the Pearson linear correlation coefficient (CC) and Area Under Curve (AUC). A higher the AUC results in a better prediction. A value of 0.5 value indicates a random performance while 1.0 denotes a perfect performance. Details of these metrics are given in [16]. The results for the AUC and CC metric are presented in table 1. The maximum for each column marked in bold.

Table 1 – AUC, CC correlation values between 2D and 3D DoF=0.1 (2D/01) SMs; between 2D and 3D DoF=0.3 (2D/03) SMs; between 3D DoF=0.1 and 3D DoF=0.3 (01/03) SMs.

	AUC			CC		
	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom_lt	0,83	0,81	0,83	0,86	0,79	0,88
Bathroom_mt	0,81	0,81	0,82	0,85	0,87	0,89
Bathroom_ht	0,81	0,83	0,85	0,82	0,86	0,91
Cartoon_lt	0,87	0,87	0,88	**0,91**	0,90	0,91
Cartoon_mt	0,85	0,85	0,86	0,84	0,84	0,92
Cartoon_ht	0,85	0,85	0,83	0,89	0,90	0,88
Hallway_lt	0,87	0,86	0,86	**0,91**	**0,93**	0,86
Hallway_mt	0,85	0,86	0,85	0,85	0,85	**0,92**
Hallway_ht	0,84	0,84	0,84	0,84	0,87	0,88
Kitchen_lt	0,86	0,85	0,88	0,88	0,85	0,91
Kitchen_mt	0,84	0,84	0,84	0,89	0,82	0,88
Kitchen_ht	0,82	0,82	0,83	0,89	0,89	0,91
Tea_lt	**0,90**	**0,90**	**0,91**	0,89	0,92	0,86
Tea_mt	0,86	0,87	0,88	0,83	0,90	0,88
Tea_ht	0,88	0,87	0,89	**0,91**	0,91	**0,92**
Room_lt	0,86	0,87	0,86	0,89	0,92	0,87
Room_mt	0,83	0,81	0,80	0,87	0,85	0,76
Room_ht	0,84	0,84	0,82	0,88	0,83	0,85

As it can be seen from the table 1, the AUC values and the CC values representing the correlation between saliency maps with different disparities are very high. This suggests that there is no strong difference between the saliency maps for a viewing duration of 20 seconds, which means that depth has no obvious influence on visual attention. In spite of this, there is considerable evidence in the literature that disparity has a time dependent saliency effect [6, 10, 11]. For further analysis, we divided the observation time of 20 seconds into 5 intervals: 1-4 seconds, 4-8 seconds, 8-12 seconds, 12-16 seconds and 16-20 seconds. Saccade length, fixation duration, disparity impact and texture complexity impact was analyzed for each time interval.

3.2. Saccade length

Each saccade length was measured as the distance between locations of two fixations in degrees. Saccade length has a tendency to shorten over time and with the introduction of disparity (Figure 6). The average decrease of saccade length over time was calculated as the difference between the saccade length of the first time interval and the last one. The difference for 2D: -0.73°, 3D DoF=0.1: -0.49°, 3D DoF=0.3: -0.32°. A paired samples t-test was conducted to compare saccade length for 2D and 3D conditions. There was a significant difference $t(89)= 3.56$, $p<0.05$, $p=0.0006$ in the scores for saccade length over time for 2D and 3D comfortable conditions. In addition, we found a significant difference of $t(89)=6.45$, $p<0.05$, $p=5.7E-09$ in the scores for average saccade length over time for 2D and 3D uncomfortable conditions. Finally, a significant difference was detected between 3D comfortable and 3D uncomfortable conditions: $t(89)= 2.24$, $p<0.05$, $p=0.027$. Results for the average saccade length for each time interval is presented in figure 6. Regardless, we did not find any proof from the paired t-tests that texture has an influence on saccade length.

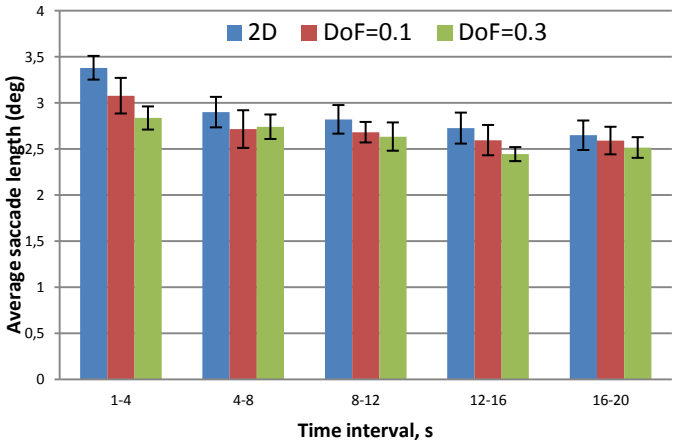


Figure 6. Influence of depth on average saccade length over time

To summarize, the average saccade length decreases constantly over time for all disparities. At the same time, saccade length decreases when bigger disparities are present. These results are in accordance with the study of Jansen *et al*. that reported that saccade length is reduced in 3D conditions and generally shortens over time.

3.3. Fixation duration

The introduction of depth into stimuli has an influence on the human visual system since binocular parallax and convergence become involved. Basically, additional time is required in order to verge eyes and fuse delivered images for the perception of depth. Hence it is expected that fixation duration would increase with disparities and with texture complexity as well.

However, the statistical analysis of fixation duration showed that there is no relation between the fixation durations and depth levels (Figure 7). Our results corroborate the findings of neither Huyanh-Thu *et al*. nor Jansen *et al*. who reported that a disparity cue shortened the median fixation duration. However, we draw attention to the fact that Jansen *et al*'s results were obtained for pink noise and white noise images; there was no effect for natural images. Huyanh-Thu *et al*. found that for video sequences the average fixation duration was shorter for 3D conditions.
The average increase in fixation duration over time was calculated as the difference between the fixation duration of the last time interval and the first one. The difference is: +30.4 ms for 2D, +27.01 ms for 3D comfortable, and +19.41 ms for

3D uncomfortable. A paired samples t-test was conducted to compare fixation durations for all corresponding disparities for the first and the last time interval. There was a significant difference for all conditions. 2D: $t(17)=-4.92$, $p<0.05$, $p=0.0001$; DoF=0.1: $t(17)=-4.48$, $p<0.05$, $p=0.0003$; DoF=0.3: $t(17)=-3.7$, $p<0.05$, $p=0.002$. Thus, in figure 7, it can be noted that the fixation duration tended to increase over time. This conclusion is supported by the data of Jansen *et al.*

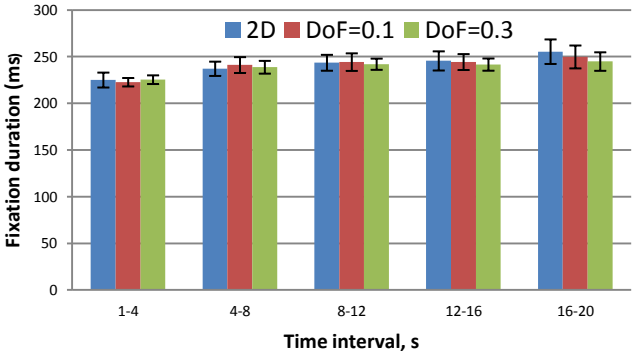


Figure 7. Influence of depth on average fixation duration over time

With a paired samples t-test, we did not find any significant influence of texture complexity on fixation duration.

3.4. Disparity effect on visual attention

In order to assess the effect of disparity on visual attention, the AUC metric was calculated between pairs of saliency maps for 2D, 3D DoF=0.1, and 3D DoF=0.3 over time. The results are presented in table 2. The maximum values for each time interval are marked in bold and the minimum values are underlined.

Table 2 – AUC values between 2D and 3D DoF=0.1 (2D/01); between 2D and 3D DoF=0.3 (2D/03); between 3D DoF=0.1 and 3D DoF=0.3 (01/03).

	1-4, s			4-8, s			8-12, s			12-16, s			16-20, s		
AUC	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom_lt	0,90	0,88	0,88	0,75	0,74	0,79	0,79	0,78	0,82	0,76	0,73	0,71	0,75	0,73	0,75
Bathroom_mt	0,84	0,82	0,87	0,79	0,80	0,77	0,78	0,73	0,77	0,71	0,76	0,74	0,72	0,76	0,75
Bathroom_ht	0,86	0,85	0,86	0,77	0,74	0,81	0,77	0,79	0,81	0,77	0,76	0,83	0,80	0,80	0,80
Cartoon_lt	0,86	0,87	0,85	0,85	0,87	0,84	0,86	0,84	0,85	0,84	0,84	0,88	0,82	0,82	0,85
Cartoon_mt	0,88	0,87	0,78	0,80	0,79	0,81	0,76	0,82	0,81	0,79	0,78	0,89	0,79	0,81	0,82
Cartoon_ht	0,86	0,85	0,82	0,81	0,82	0,82	0,82	0,83	0,79	0,79	0,81	0,80	0,81	0,76	0,79
Hallway_lt	0,88	0,88	0,89	0,85	0,87	0,87	0,81	0,78	0,79	0,83	0,81	0,82	0,80	0,83	0,83
Hallway_mt	0,86	0,85	0,83	0,79	0,80	0,83	0,79	0,76	0,83	0,78	0,78	0,80	0,81	0,86	0,72
Hallway_ht	0,81	0,77	0,85	0,81	0,80	0,83	0,79	0,83	0,78	0,79	0,77	0,86	0,75	0,78	0,78
Kitchen_lt	0,80	0,73	0,86	0,85	0,81	0,86	0,85	0,83	0,85	0,84	0,79	0,84	0,79	0,80	0,81
Kitchen_mt	0,85	0,81	0,87	0,80	0,81	0,81	0,76	0,80	0,73	0,81	0,80	0,80	0,81	0,79	0,78
Kitchen_ht	0,79	0,75	0,81	0,77	0,82	0,82	0,75	0,82	0,77	0,81	0,78	0,74	0,75	0,65	0,71
Tea_lt	0,90	**0,91**	0,91	0,87	0,83	0,85	**0,88**	0,87	0,86	0,87	**0,91**	0,89	0,86	0,87	**0,87**
Tea_mt	0,84	0,84	0,88	0,79	0,76	0,84	0,85	0,87	0,86	0,86	0,84	0,81	0,87	0,84	0,82
Tea_ht	0,89	0,88	0,88	0,82	0,83	**0,91**	0,87	0,86	0,86	0,86	0,83	0,80	0,79	0,82	0,83
Room_lt	0,84	0,81	0,84	0,83	0,87	0,84	0,82	0,81	0,81	0,84	0,82	0,82	0,80	0,80	0,79
Room_mt	0,85	0,84	0,79	0,76	0,73	0,77	0,79	0,83	0,75	0,75	0,70	0,69	0,79	0,84	0,78
Room_ht	0,88	0,86	0,83	0,79	0,79	0,80	0,83	0,77	0,76	0,79	0,74	0,65	0,71	0,72	0,72
Average	0,85	0,84	0,85	0,80	0,80	0,82	0,81	0,81	0,81	0,80	0,79	0,80	0,79	0,79	0,79
Confidence int, ±	0,01	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,03	0,02	0,02	0,02

At each time interval, the maximum AUC value belongs to the scene “Tea”. This result reflects the particularity of the scene: this scene contained the least number of objects, they were located in the center of the scene at the foreground, and there was no pronounced background (the tablecloth was moving away with depth to infinity). The depth distortion was not as noticeable for DoF=0.3 as for others scenes because the objects were at the foreground. The minimum values belong to the scene “Kitchen” (1-4 s, 8-12 s, 16-20 s) and to the scene “Room” (4-8 s, 12-16 s). This result is also not

surprising since these scenes contained a lot of different objects, along with a depth distortion for DoF=0.3 that was very pronounced.

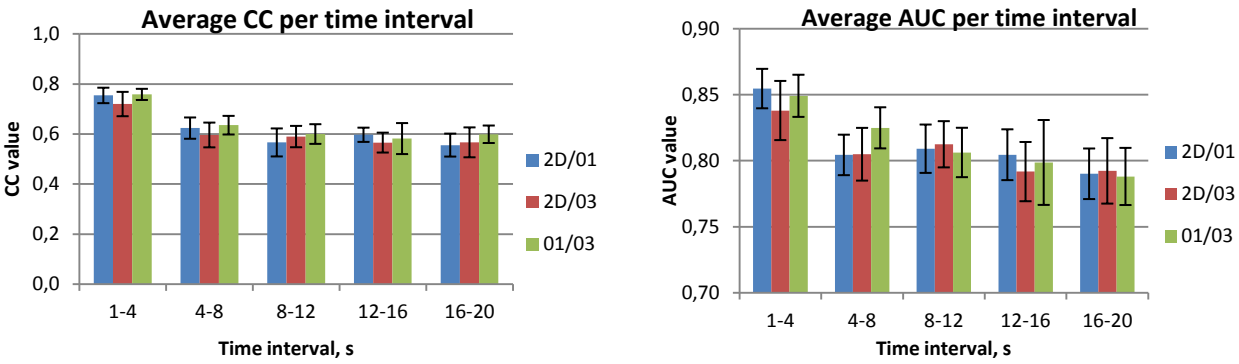
A paired samples t-test was conducted to compare the AUC values for 2D and 3D conditions. There was a significant difference in the scores for 2D/01 and 2D/03 conditions, $t(17) = 2.11$, $p<0.05$, $p=0.004$ for the period of time 1-4 seconds. For the rest of the time periods, the differences were insignificant. For the period of time 4-8 seconds, there was a significant difference in the scores for 2D/01 and 01/03 conditions, $t(17)=-3.51$, $p<0.05$, $p=0.003$ and in the scores for 2D/03 and 01/03 conditions, $t(17)=-2.48$, $p<0.05$, $p=0.024$. The differences were insignificant for the rest of the time periods for both conditions.

The way images are observed during the first time interval 1-4 s is the most similar to the observation of videos where the frames change one after the other. Basically, there is no time to explore all the parts of the complex scenes and attention is attracted by salient regions. Besides, at this time period the way images are observed can be influenced by the central bias [11, 19, 20]. The demonstrated values of the AUC reflect the fact that the saliency maps are very similar for all the scenes. The minimum value during 4 first seconds is 0.73 for one of the most complex scenes ("Kitchen") while the average value is 0.85. In our opinion, the high AUC values could indicate that disparity plays a very subtle role in the selection of salient features of a scene. But this hypothesis is not fully supported by the CC values (min 0.47; avg 0.74) which are presented in table 3 for comparison. Thereby we are lacking a standard (a method or a metric), which allows for the evaluation of depth effect and then comparison of results within different studies.

Table 3 – CC values between 2D and 3D DoF=0.1 (2D/01); between 2D and 3D DoF=0.3 (2D/03); between 3D DoF=0.1 and 3D DoF=0.3 (01/03).

	1-4, s			4-8, s			8-12, s			12-16, s			16-20, s		
CC	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom_lt	0,79	0,77	0,78	0,54	0,43	0,67	0,56	0,57	0,63	0,58	0,44	0,37	0,59	0,48	0,61
Bathroom_mt	0,76	0,78	0,85	0,51	0,54	0,64	0,53	0,50	0,58	0,52	0,58	0,59	0,38	0,59	0,53
Bathroom_ht	0,78	0,76	0,79	0,55	0,56	0,60	0,64	0,61	0,65	0,53	0,60	0,54	0,60	0,46	0,66
Cartoon_lt	0,77	0,79	0,76	0,81	0,76	0,77	0,72	0,64	0,72	0,71	0,70	0,78	0,75	0,77	0,72
Cartoon_mt	0,75	0,76	0,72	0,65	0,56	0,65	0,54	0,68	0,69	0,52	0,50	0,81	0,58	0,56	0,64
Cartoon_ht	0,70	0,81	0,69	0,48	0,44	0,59	0,61	0,65	0,61	0,64	0,67	0,73	0,57	0,47	0,57
Hallway_lt	0,86	0,83	0,86	0,72	0,77	0,73	0,57	0,61	0,64	0,60	0,62	0,58	0,57	0,72	0,60
Hallway_mt	0,78	0,81	0,75	0,65	0,55	0,63	0,43	0,34	0,67	0,61	0,55	0,59	0,61	0,76	0,62
Hallway_ht	0,71	0,73	0,79	0,60	0,46	0,54	0,52	0,58	0,67	0,61	0,54	0,65	0,42	0,57	0,51
Kitchen_lt	0,65	0,51	0,73	0,77	0,50	0,71	0,66	0,58	0,67	0,64	0,48	0,63	0,50	0,50	0,61
Kitchen_mt	0,79	0,65	0,75	0,63	0,62	0,72	0,36	0,53	0,41	0,61	0,52	0,66	0,54	0,38	0,64
Kitchen_ht	0,61	0,47	0,66	0,53	0,58	0,55	0,34	0,71	0,48	0,58	0,54	0,41	0,61	0,29	0,49
Tea_lt	0,83	0,80	0,75	0,66	0,65	0,46	0,71	0,72	0,60	0,54	0,77	0,54	0,72	0,67	0,73
Tea_mt	0,65	0,64	0,74	0,55	0,65	0,69	0,58	0,58	0,61	0,59	0,54	0,62	0,50	0,60	0,57
Tea_ht	0,78	0,73	0,78	0,74	0,77	0,70	0,70	0,60	0,65	0,66	0,58	0,63	0,62	0,63	0,69
Room_lt	0,76	0,69	0,76	0,56	0,70	0,63	0,64	0,58	0,56	0,73	0,60	0,62	0,57	0,64	0,63
Room_mt	0,78	0,61	0,75	0,65	0,56	0,63	0,39	0,67	0,44	0,56	0,51	0,41	0,43	0,65	0,51
Room_ht	0,83	0,80	0,76	0,62	0,63	0,53	0,69	0,46	0,54	0,52	0,45	0,33	0,44	0,48	0,48
Average	0,75	0,72	0,76	0,62	0,60	0,64	0,57	0,59	0,60	0,60	0,57	0,58	0,56	0,57	0,60
Confidence int, ±	0,03	0,05	0,02	0,04	0,05	0,04	0,06	0,04	0,04	0,03	0,04	0,06	0,05	0,06	0,03

It is likely that during the first time interval all observers are attracted by the most salient features of a scene, but in later time periods, the way the image is observed may differ from one observer to another. This hypothesis is supported by the AUC and CC values. In figure 8, the AUC and CC average values are presented over time. The highest correlation values in both cases were obtained for the first time interval. The paired t-test showed that the difference between the first time interval (1-4 s) and the second time interval (4-8 s) is significant for CC values as well as for the AUC. The results of the paired t-test are presented in table 4.



(a) (b)

Figure 8. Influence of depth on average fixation duration over time (a) CC scores, (b) AUC scores.

Table 4. Paired t-test: difference in AUC and CC scores for the first (1-4 s) and the second (4-8 s) time interval; t(17), $p<0.05$.

p	CC	AUC
2D/01	2,32E-05	0,00014
2D/03	0,001724	0,045636
01/03	6,7E-06	0,011549

The results presented in figure 8 are in accordance with the study of Tatler *et al.* [21], who found that the consistency between visual fixations of different subjects is high just after the stimulus is displayed but progressively decreases over time. It is likely that just after the stimulus is displayed, our attention is mostly controlled by bottom-up mechanisms, whereas top-down mechanisms become more influential after several seconds of viewing. The second factor is content dependent.

For the viewing duration of 20 s, the average AUC and CC values are presented in table 5. These results indicate that for such long durations of time, depth levels did not have an obvious influence on saliency maps – AUC and CC values are very high and similar for every condition. Values for every scene for the viewing duration of 20 seconds were presented above in table 1.

Table 5 – Average AUC and CC values for 20 s between 2D and 3D with DoF=0.1, between 2D and 3D with DoF=0.3 and DoF=0.1 and DoF=0.3.

	2D/01	2D/03	01/03
CC	0,87±0,01	0,87±0,02	0,88±0,02
AUC	0,85±0,01	0,85±0,01	0,85±0,01

Based on an analysis of the heat maps, we did not notice an influence of depth levels on selection of salient features in a scene after 4 seconds for 5 scenes out of 6. Only for the scene "Tea" during the time interval 4-8 s, some gaze points appeared at the background of the image for DoF=0.3 (Figure 8). We suppose this happened later than in previous cases due to the location of the objects: they were located in the foreground and there was no pronounced background. In the first time period (1-4 s), depth was less pronounced than the center biases and the saliency of the objects.

As in the work of Ramasamy *et al.*, one of our scenes contained a deep hallway. For the first 4 seconds, our results correlate with their work: the gaze points were more spread in 2D conditions and more concentrated at the far end in 3D conditions independent of DoF. But after 4 seconds, the gaze points became more spread and the heat maps looked similar to non-stereoscopic conditions. For scenes like "Cartoon" or "Kitchen", a similar visual behavior has been noticed but is less pronounced. A possible reason is the presence of greater number of objects in the scenes. Nevertheless, clear evidence of this tendency was not revealed for the rest of the scenes. Thus, it is possible that the saliency of the objects in the case of the "Bathroom" and "Room" scenes plays a more important role than the presence of depth. To summarize, we did not observe any particular relation between the depth and the spread of the gaze points.

Nevertheless, we believe that an analysis of heat maps is not fully reliable because a non-normalized color scale between scenes hampers their comparison.

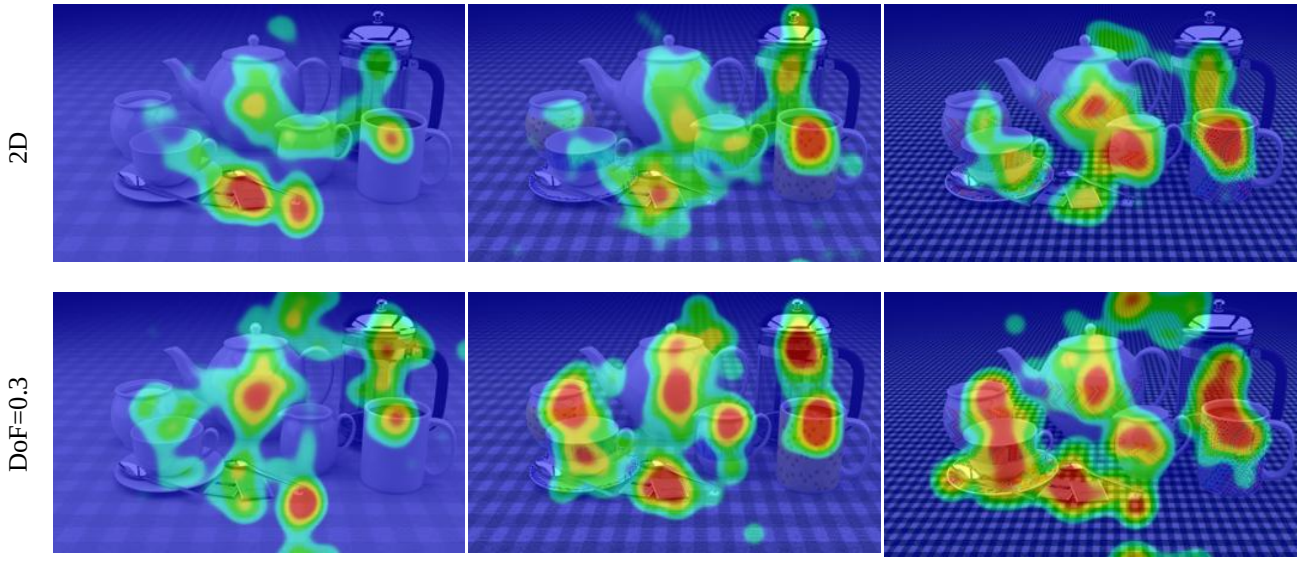


Figure 8. The first row SMs for 1-4 s for "Tea" with low, medium, high texture complexity from left to right. The second row SMs for 4-8 s for "Tea" with low, medium, high texture from left to right.

3.5. Texture effect on visual attention

In this section, we investigate whether texture complexity has an impact on visual attention or not. Thus, we calculated inter-observer visual congruency (IOVC) , which reflects the visual dispersion between observers or the consistency of overt attention (eye movement) while observers are watching the same visual scene [22]. As was already mentioned in the previous section, visual attention is controlled by low-level visual features most probably just after each stimulus is displayed. After several seconds top-down processes begin, which are content dependent. As a consequence, a stimulus composed of salient areas would presumably attract our visual attention, leading to high congruency. The presence of particular features, such as human faces, people, or animals, tends to increase the consistency between observers. On the other hand, the congruency tends to decrease with scene complexity.

The calculation of IOVC was done with all the fixation data for the viewing duration of 20 s. The computed results are presented in table 6. A paired samples t-test was conducted to compare inter-observer visual congruency for different texture complexities. There was a significant difference $t(17)= 1.74$, $p<0.05$, $p=0.002$ in the scores for congruency for LT and MT complexity. A significant difference was found in the scores for congruency for LT and HT complexity: $t(17)= 1.74$, $p<0.05$, $p=0.004$. Nevertheless, no significant difference was detected between MT and HT ($p<0.05$, $p=0.08$). For each scene high and medium texture complexities were selected by experts without using any metric. This could be the possible reason similar results were obtained for both medium and high texture complexities.

Table 6 – IOVC for low (LT), medium (MT), high (HT) texture complexities.

2D	LT	MT	HT
Bathroom	0,63	0,63	0,69
Cartoon	0,76	0,74	0,68
Hallway	0,73	0,73	0,69
Kitchen	0,76	0,70	0,68
Tea	0,83	0,76	0,78
Room	0,75	0,69	0,70

3D DoF=0.1	LT	MT	HT
Bathroom	0,70	0,66	0,67
Cartoon	0,76	0,72	0,65
Hallway	0,74	0,73	0,72
Kitchen	0,78	0,69	0,70
Tea	0,84	0,79	0,80
Room	0,71	0,65	0,67

3D DoF=0.3	LT	MT	HT
Bathroom	0,63	0,68	0,64
Cartoon	0,70	0,74	0,73
Hallway	0,73	0,67	0,73
Kitchen	0,72	0,66	0,71
Tea	0,80	0,82	0,79
Room	0,74	0,66	0,63

In figure 9, fragments of the scene "Kitchen" and "Bathroom" are presented with three texture complexities. It can be seen that the cupboard in the front of the "Kitchen" scene (figure 9a) does not attract attention, whereas when there is some pattern on top of the cupboard, its doors become salient (figure 9b, 9c). A similar situation happened with the

"Bathroom scene". (figure 9 d-f). Based on the analysis of the heat maps and the paired t-test of IOVC values, we can deduce that the selected areas of interest depend on the texture. Therefore, the resulting saliency maps might differ.

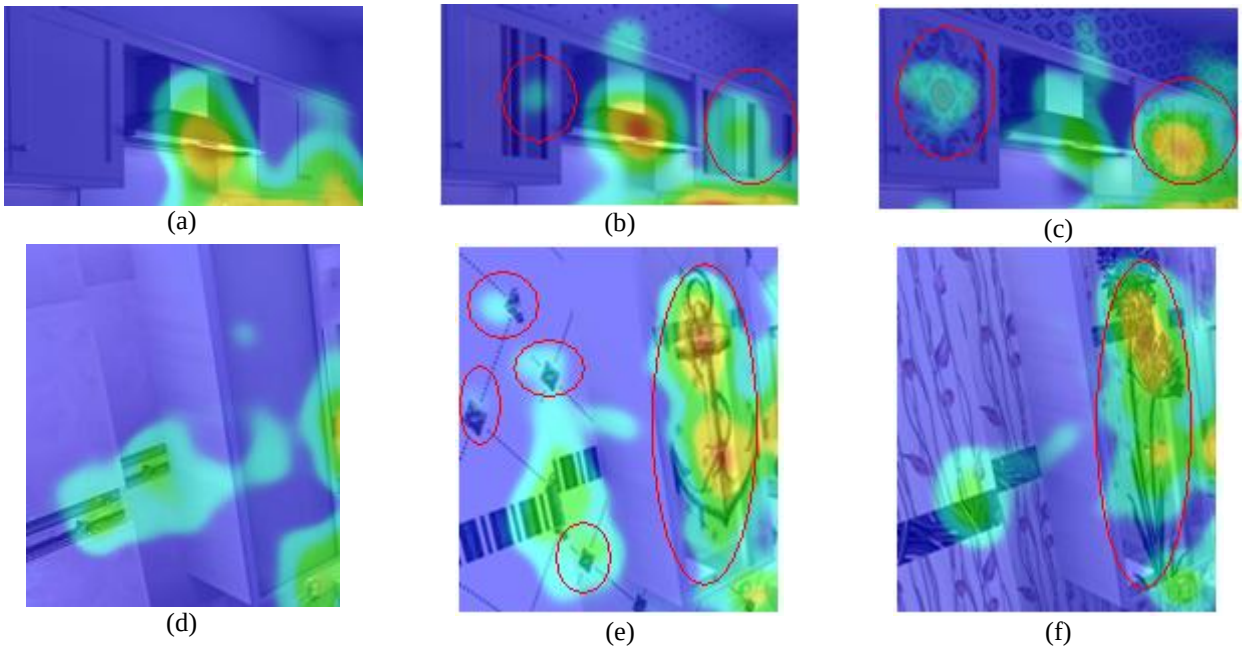


Figure 9. The first row shows fragments of heat maps for "Kitchen" with low, medium, high texture from left to right. The first row fragments of heat maps are from "Bathroom" with low, medium, high texture from left to right.

4. CONCLUSIONS

We conducted an eye-tracking experiment using 2D and 3D images with different DoF. This study aimed at estimating influence of depth, comfort/discomfort and texture complexity on visual attention. One feature of the study was that all stereoscopic content was with uncrossed disparity, i.e., all objects were behind the display plane. We evaluated the effect of the depth on basic eye movement properties. The average saccade length decreased constantly over time for all disparities. At the same time, average saccade length decreased with bigger disparities as well. These results are in accordance with the study of Jansen *et al*. The analysis of fixation duration showed that there is no relation between the fixation durations and disparities. Our results are in opposition to former studies, which reported that disparity cues shortened the median fixation duration.

As described in the background section many works have reported spread of gaze points over a scene. Anyhow still there is no common conclusion whether depth changes position of gaze points. For all the scenes, we found that the gaze points were denser and centered in the middle of a scene during the first 4 seconds, but were spread over the entire scene for the other time intervals. We did not find any strong evidence that depth has an influence on the spread of gaze points. This is in accordance with the conclusion of Huynh-Thu *et al*.

Based on the AUC and CC scores we could assume that the saliency of an object plays a more important role than the depth. We did not find any strong evidence that indicated the influence of disparity. Even when an analysis was performed for the first time interval (1-4 s), the AUC and CC scores remained very high. On the other hand, a visual analysis of the heat maps showed that there is an influence of disparity but it is not possible to conclude whether this is significant or not. All paired t-tests that aimed to find differences between comfortable (DoF =0.1) and uncomfortable (DoF=0.3) conditions were not significant. Among the reasons that might explain this result is the test methodology. The entire test for each observer lasted 2 min 40 s, and the uncomfortable condition only lasted for 40 seconds. Thus, the visual system was not stressed. After the experiment several observers reported that they experienced discomfort in some cases, but they were still looking at the background because they had never experienced depth distortion and were curious to observe such contents. Therefore, it would be necessary to stress visual systems before the test and then repeat

the experiment with stimuli for DoF=0.3. Another possible reason we did not find the disparities to have any pronounced influence is the absence of a method that is sensitive to disparities, which could reveal differences in saliency maps for different DoF. We believe that it is important to define a method or a metric which would allow for a comparison of results between different studies that investigate the effect of disparity on visual attention.

Finally, we calculated inter-observer visual congruency to assess the influence of texture complexity on visual attention; we discovered that there was a significant difference for low texture complexity and medium texture complexity, as well as for low texture complexity and high texture complexity. Based on the analysis of heat maps and paired t-tests of IOVC values, we deduced that selected areas of interest depend on the texture complexities.

REFERENCES

[1] Posner, M., "Orienting of attention," *The Quarterly Journal of Experimental Psychology*, vol. 32, pp. 3-25, (1980).

[2] Koch, C., Fau - Ullman, S., and Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitry," 19860916 DCOM- 19860916, (1985).

[3] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 1254-1259, (1998).

[4] Yarbus, A. L., "Eye movements and vision," *Plenum*, (1967).

[5] Wexler, M. and Ouarti, N., "Depth Affects Where We Look," *Current Biology*, vol. 18, pp. 1872-1876, 2008.

[6] Jansen, L., Onat, S., and König, P., "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, (2009).

[7] Wismeijer, D. A., Erkelens, C. J., van Ee, R., and Wexler, M., "Depth cue combination in spontaneous eye movements," *Journal of Vision*, vol. 10, (2010).

[8] Ramasamy, C., Donald, H. H., Andrew, T. D., and Brian, D., "Using eye tracking to analyze stereoscopic filmmaking," in *SIGGRAPH '09: Posters*, ed. New Orleans, Louisiana: ACM, pp. 1-1, (2009).

[9] Hakkinen, J., Kawai, T., Takatalo, J., Mitsuya, R., and Nyman, G., "What do people look at when they watch stereoscopic movies?," San Jose, California, USA, pp. 75240E-10, (2010).

[10] Huynh-Thu, Q. and Schiatti, L., "Examination of 3D visual attention in stereoscopic video content," pp. 78650J-78650J, (2011).

[11] Gautier, J. and Le Meur, O., "A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions," *Cognitive Computation*, vol. 4, pp. 141-156, 2012-06-01, (2012).

[12] Chen, W., Fournier, J., Barkowsky, M., and Le Callet, P., "New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone," San Francisco, California, USA, pp. 78631O-13, (2011).

[13] Yano, S., Emoto, M., and Mitsuhashi, T., "Two factors in visual fatigue caused by stereoscopic HDTV images," *Displays*, vol. 25, pp. 141-150, (2004).

[14] Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S., "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, (2008).

[15] Chen, W., Fournier, J., Barkowsky, M., and Le Callet, P., "Quality of experience model for 3DTV," Burlingame, California, USA, (2012), pp. 82881P-9.

[16] Bobal57, Robo3dguy, Jonfreer, and Jay-Artist, "Original design of "Bathroom", "Cartoon", "Hallway", "Kitchen", "Tea", "Room". ," ed: http://www.blendswap.com, (2011-2012).

[17] Le Meur, O. and Baccino, T., "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, pp. 1-16, (2012).

[18] Le Meur, O., "Fixation analysis software," in http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html#soft, ed. IRISA, Rennes 1, France, (2012).

[19] Tatler, B. W., "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, November 19, (2007).

[20] Judd, T., Ehinger, K., Durand, F., and Torralba, A., "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, (2009), pp. 2106-2113.

[21] Tatler Bw Fau - Baddeley, R. J., Baddeley Rj Fau - Gilchrist, I. D., and Gilchrist, I. D., "Visual correlates of fixation selection: effects of scale and time," 20041228 DCOM- 20050524.

[22] Le Meur, O., Baccino, T., and Roumy, A., "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," presented at the Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA, (2011).