# Livrable D6.1 of the PERSEE project : Perceptual Assessment : Definition of the scenarios

Junle Wang, Josselin Gautier, Emilie Bosc, Jing Li, Vincent Ricordel

Projet PERSEE

´ SCHÈMAS PERCEPTUELS ET CODAGE VIDÈO 2D ET 3D [A]

n° ANR-09-BLAN-0170

Livrable **D6.1** *15/10/2011*

---

# Perceptual Assessment : Definition of the scenarios

---


| | | |
|---|---|---|
| Junle | WANG | IRCCyN |
| Josselin | GAUTIER | IRISA |
| Emilie | BOSC | INSA |
| Jing | LI | IRCCyN |
| Vincent | RICORDEL | IRCCyN |

# Contents

# 1 Introduction

In this report we describe contributions for the task 6 within the PERSEE project. The target of this task is to carry on the perceptual and subjective performance assessment of the tools developed for the project.

Here we present the definition of the scenarios of the experimental tests that have been done in link with the perceptual models of the task 1 (these models are described in the deliverable D1.2 entitled "Perceptual Modelling: Definition of the models"). We present also the tests that have been done to assess the subjective quality of synthesized contents.

The document is organized as follows:

- Section 1 is about the test to assess the impact of the blur on the depth perception;
- Section 2 presents the test to assess the depth bias;
- Section 3 presents the test to assess the subjective quality of synthesized contents (coded or not);
- Section 4 presents the test to assess the visual discomfort induced by moving stimulus.

# 2 Test to assess the impact of the blur on the depth perception

When 3D images are shown on a planar stereoscopic display, binocular disparity becomes a pre-eminent depth cue. But it induces simultaneously the conflict between accommodation and vergence, which is often considered as a main reason for visual discomfort. If we limit this visual discomfort by decreasing the disparity, the apparent depth also decreases. This psychophysical experiment was designed to quantitatively evaluate the influence of a monocular depth cue, blur, on the apparent depth of stereoscopic scenes.

We propose to decrease the (binocular) disparity of 3D presentations, and to reinforce (monocular) cues to compensate the loss of perceived depth and keep an unaltered apparent depth. We conducted a subjective experiment using a two-alternative forced choice task. Observers were required to identify the larger perceived depth in a pair of 3D images with/without blur. By fitting the result to a psychometric function, we obtained points of subjective equality in terms of disparity, from which we could compute the increase of perceived depth caused by blur.

## 2.1 Subjects

Thirty-five subjects participated in the experiment. Twelve subjects are male, twenty-three are female. The subjects did not know about the purpose of the experiment. Subjects ranged in age from 17 to 40 years. All had either normal or corrected-to-normal visual acuity. They were tested for visual acuity using a Snellen Chart, for depth acuity using a Randot Stereo Test and for color vision using Ishihara plates.All the subjects were compensated and were naive to the purpose of the experiment.

## 2.2 Apparatus

One of the limitations of previous studies was the apparatus used. CRT monitors were used and this type of displays has some drawbacks: (1)the surface containing stimuli was slightly curved, (2) the stimuli's virtual distance was affected by refraction due to the front glass plate, and (3) the screen was usually not large enough to cover a favorable field-of-view.

In our study, state-of-the-art stereoscopic display was used. Stimuli were displayed on a Samsung 22.5-inch LCD screen (figure 1), which had a resolution of 1680 * 1050 pixels, and the refresh rate was 120 Hz. Each screen pixel subtended 65.32 arcsec at a 90 cm viewing distance. The display yielded a maximum luminance of about 50 $cd/m^2$ when watched through the activated shutter glasses. Stimuli were viewed binocularly through the Nvidia active shutter glasses (Nvidia 3D Vision kit) at a distance of approximately 90 cm. The peripheral environment luminance was adjusted to about 44 $cd/m^2$. When seen through the eye-glasses, this value corresponded to about 7.5 $cd/m^2$ and thus to 15% of the screen's maximum brightness as specified by ITU-R BT.500.

(a) (b)

Figure 1: Apparatus. (a) Samsung SyncMaster 2233 monitor; (b) Nvidia 3D vision kit.

## 2.3 Stimuli

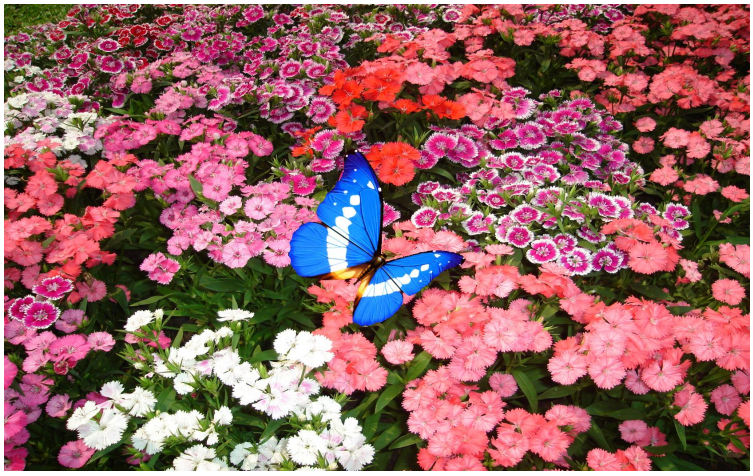
Each stimulus consisted of one single object in the foreground and a background. A 400 * 400 pixels butterfly image was used as the foreground object. This stimulus was easy to accommodate because it was spatially complex and therefore contained a wide range of spatial frequencies from low to high. The butterfly was shown with a horizontal offset between the left and the right view in order to create a disparity cue. The magnitude and direction of this offset varied to supply a variety of near or far disparity. At the 90 cm viewing distance, the background plane subtended 29.8 * 18.9 arcdeg, and the foreground object subtended 7.2 arcdeg. The background plane was a photo of a flowerbed. The background contained a great amount of textures, contained no distinct region of interest distracting the observers, and looked natural. When required, the background was spatially blurred by applying a Gaussian blur kernel with a 5-pixel radius. This amount of blur equals to the blur created by supposing that the focused object (foreground) and the defocused object (background) are at a distance of 13.7 cm in front of the screen and 19.8 cm behind the screen respectively, both of which are the limitations of the comfortable viewing zone derived from the 90 cm viewing distance. Note that all the blurred background were with the same amount of blur.

## 2.4 Design and procedure

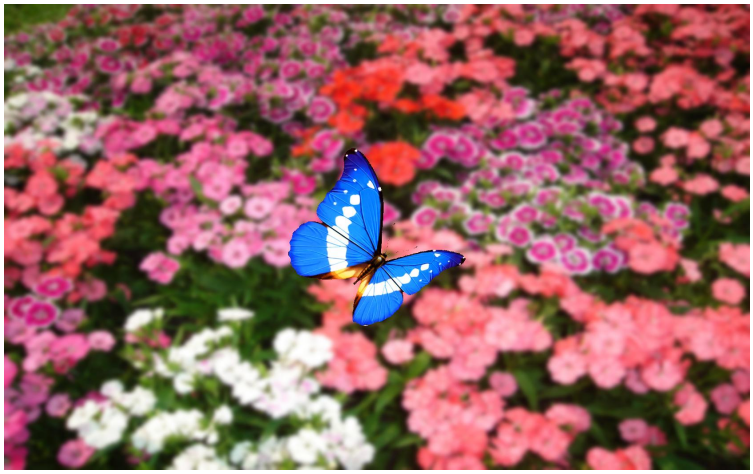
In each trial, a pair of stimuli were shown to the subjects. One stimulus contained a blurred background (BB) and a sharp foreground object, while the other stimulus contained a sharp background (SB) and also a sharp foreground object (as shown in figure 2). The backgrounds of both stimuli were positioned at the same depth. Two parameters of the stimuli were varied: the absolute position ($D_a$) of the background plane and the relative distance ($D_r$) of the foreground object to the background. The selection of absolute distance ranged from -19.7 cm to 6.6 cm in steps of 6.6 cm (negative values denote the positions behind the screen plane, and positive values denote the positions in front of the screen plane). The relative distance ranged from 0 cm to approximately 33 cm. All the positions of both background and foreground objects are selected considering the limitations of the comfortable viewing zone[13]. There are thus twenty combinations of absolute position and relative distance in total for the BB-stimuli.

Once the absolute position and the relative distance were selected for each BB-stimulus, then we paired the BB-stimulus with a set of 7 or 8 SB stimuli which are with relative distances ranging from 4 cm less than the BB stimulus to 8 cm larger than the reference stimulus. These steps were chosen considering both the depth rendering ability of the screen and the depth perception ability of the observers. One trial contains one BB-stimulus and one SB-stimulus. We had 155 trials in the experiment. This setup is shown in Figure 3, and the parameters are shown in Table 1.

In Figure 3, the blue planes represent the (foreground) depth planes in which the



(a) The left view of a Sharp-Background (SB) stimuli



(b) The left view of a Blur-Background (BB) stimuli

Figure 2: Example Stimuli

butterfly is located, the red plane represents the depth plane in which the background is located. We named the distance between background plane and the screen as Absolute Position ($D_a$), while the interval between the foreground plane and the background plane as the Relative Distance ($D_r$). Both these two distances were free parameters of the design of stimuli. The variation range of these two planes always stayed in the comfortable viewing zone. The first figure shows the BB-stimulus case, while the second shows the SB-stimulus case. Each BB-stimulus is paired with one of 7 or 8 SB-stimuli to create one trial. The selection of the parameters $D_a$ and $D_r$ for both BB-stimuli and SB-stimuli are presented in Table 1.

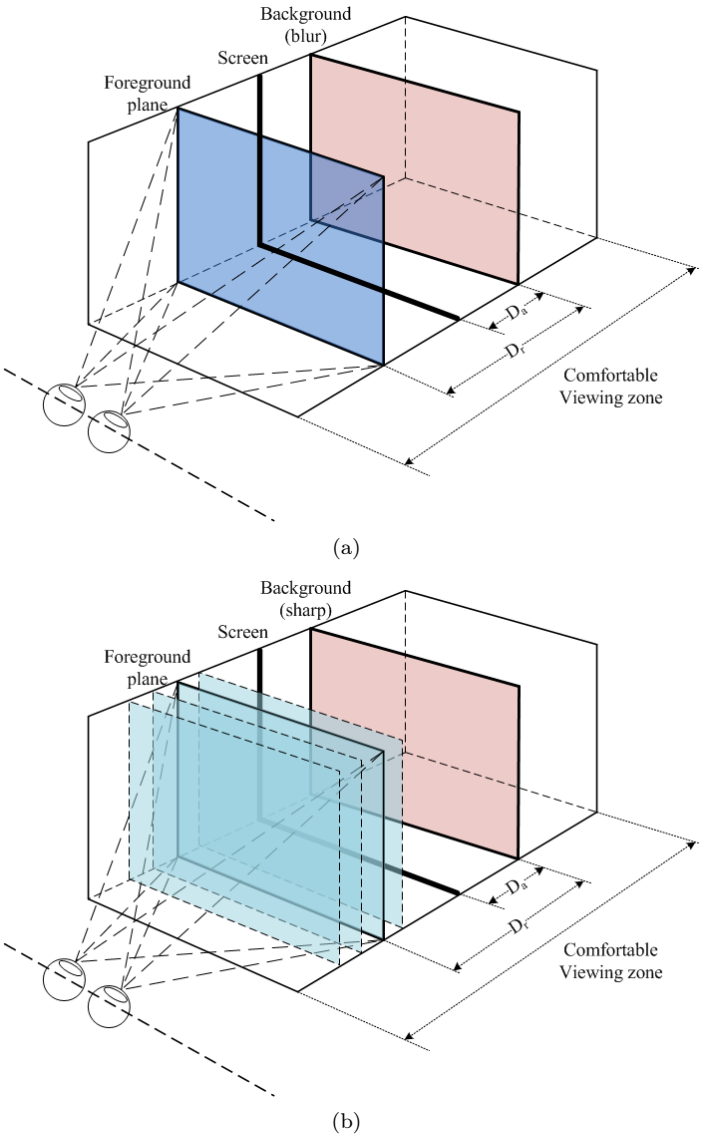



Figure 3: Schematic diagram of the experiment setup

The subjective experiment was conducted using a two-alternative forced choice task. For each trial, one of the 155 conditions was chosen randomly and a pair of stimuli were displayed. Observers were asked to look at the butterfly and then determine in the trial whether the BB stimulus contained a larger depth interval between the

| $D_a(cm)$ | $D_{r\_BB}(cm)$ | $D_{r\_SB}(cm)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -19.7 | 0 | | 0 | 1.2 | 2.3 | 3.4 | 4.5 | 5.6 | 6.6 |
| -19.7 | 6.6 | 3.4 | 6.6 | 7.6 | 8.6 | 9.6 | 10.6 | 11.5 | 13.4 |
| -19.7 | 13.2 | 11.1 | 13.2 | 14.3 | 15.6 | 16.9 | 18.1 | 19.7 | 21.7 |
| -19.7 | 19.7 | 17.3 | 19.7 | 20.9 | 22 | 23.5 | 24.9 | 26.6 | 28.3 |
| -19.7 | 26.3 | 24.9 | 26.3 | 27 | 28 | 29.3 | 30.5 | 32 | 33.4 |
| -19.7 | 32.8 | 30.5 | 32.8 | 34 | 35.1 | 36.2 | 37.5 | 38.7 | 39.9 |
| -13.2 | 0 | | 0 | 1 | 2 | 3 | 4 | 4.9 | 5.9 |
| -13.2 | 6.6 | 4.5 | 6.6 | 7.7 | 8.6 | 9.8 | 11.1 | 12.7 | 14.3 |
| -13.2 | 13.2 | 11.5 | 13.2 | 14.3 | 15.4 | 16.5 | 17.6 | 18.7 | 19.7 |
| -13.2 | 19.7 | 18.3 | 19.7 | 20.7 | 21.4 | 22.3 | 23.3 | 24.5 | 26 |
| -13.2 | 26.3 | 24.5 | 26.3 | 27.1 | 27.9 | 29 | 30.1 | 31.4 | 32.6 |
| -6.6 | 0 | | 0 | 0.9 | 1.8 | 2.7 | 3.5 | 4.4 | 5.2 |
| -6.6 | 6.6 | 3.5 | 6.6 | 8 | 9.1 | 10.2 | 11.6 | 13.4 | 15.4 |
| -6.6 | 13.2 | 12 | 13.2 | 14.4 | 15.4 | 16.3 | 17.3 | 18.5 | 19.6 |
| -6.6 | 19.7 | 16.9 | 19.7 | 20.8 | 21.6 | 22.7 | 23.8 | 25.3 | 26.8 |
| 0 | 0 | | 0 | 0.8 | 1.6 | 2.3 | 3.4 | 4.5 | 5.6 |
| 0 | 6.6 | 4.5 | 6.6 | 7.6 | 8.6 | 9.5 | 10.8 | 12.3 | 14.3 |
| 0 | 13.2 | 11.4 | 13.2 | 13.7 | 14.3 | 15.1 | 15.9 | 17.2 | 18.7 |
| 6.6 | 0 | | 0 | 1 | 2 | 2.9 | 3.9 | 4.8 | 6.3 |
| 6.6 | 6.6 | 4.2 | 6.6 | 7.1 | 8 | 8.8 | 9.9 | 11.2 | 12.9 |

Table 1: The selection of parameters $D_a$ and $D_r$ for both BB-stimuli and SB-stimuli. Note that there are only 7 possible selections for the $D_{r\_BB}$, because putting the foreground behind the background will cause trouble of fusion.

butterfly and the background than the SB stimulus. As a two monitor screen setup was technically not possible, observers were able to control the displaying of stimuli by means of a key press to switch from one stimulus to the other. When the observer switched between the stimuli, a 700 ms grey interval was shown in order to avoid memorization by the observers of the "exact" positions of the foregrounds. For each trial, the total observation time and the number of switches were not limited.

## 2.5 Result and data analysis

For each condition, some observers considered the BB-stimulus as having a larger depth interval (between the foreground and the background), while the other observers chose the SB-stimulus. We measure the proportion of 'BB-stimulus contains a larger depth interval' responses, and plot the data as a function of the disparity difference between the Dr in the BB-stimulus ($D_{r\_BB}$) and the Dr in the SB-stimulus ($D_{r\_SB}$). The cumulative Weibull function was used as the psychometric function. The disparity difference corresponding to the 50% point can be considered as the Point of Subjective Equality (PSE). When measuring the disparity difference at that point, the increase

of perceived depth is obtained. In total, by filtering out the data of 7 observers who made decisions in the test quite differently from other observers, 28 observations of each conditions were included in the computation. An example pattern of response and the fitted psychometric function is shown in Figure 4.

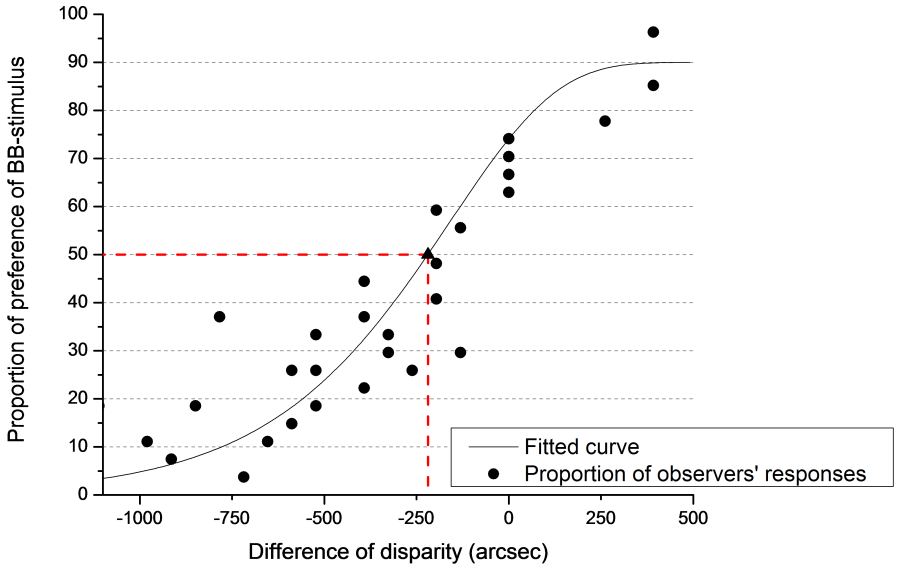



Figure 4: An example pattern of the proportion of observers' responses and the fitted psychometric function. In this trial, we consider $D_{r\_BB}$ = 6.6 cm and $D_a$ = -19.7 cm, -13.2 cm, -6.6 cm, 0 cm, 6.6 cm. An equal apparent depth is reached at -220 arcsec

# 3 Test to assess the depth bias

In the studies of 2D visual attention, eye-tracking data shows a so-called "center-bias", which means that fixations are biased towards the center of 2D still images. However, in the stereoscopic visual attention, depth is another feature having great influence on guiding eye movements. Relative little is known about the impact of depth.

We conducted a binocular eye-tracking experiment by showing synthetic stimuli on a stereoscopic display. Observers were required to do a free-viewing task through active shutter glasses. Gaze positions of both eyes were recorded for obtaining the depth of fixation. Stimuli were well designed in order to let the center-bias and depth-bias affect eye movements individually. Results showed that the number of fixations varies as a function of depth planes.

## 3.1 Stimuli

Synthetic stereoscopic stimuli were used for this experiment. The stimuli consisted in the presentation of scenes in which a background and some similar objects were deliberately displayed at different depth positions. We generated the depth by horizontally shifting the objects to simulate the binocular disparity. This was also the only depth cue we took advantage of in this experiment.

The background was a flat image consisting in white noise (figure 5 (a)), which was placed at a depth value of -20 cm (20 cm beyond the screen plane). In each scene, the objects consisted in a set of black disks of the same diameter S. They were displayed at different depth values randomly chosen among {-20, -15, -10, -5, 0, 5, 10, 15, 20} cm. Though the objects were placed at different depths (figure 5 (b)), the positions of projection of the objects on the screen plane uniformly laid on a circle centered on the screen center (figure 5 (c)). Thus, we assume that no "center-bias" was introduced in the observation.

Three parameters were varying from one scene to another: 1, the number of objects, $N \in \{5, 6, 7, 8, 9\}$; 2, the radius R of the circle on which the objects were projected on the screen plane, $R \in \{200, 250, 300\}$ pixels; 3, the size of the objects, which was represented by the diameter of the disk S varying from $\frac{\pi R}{N\sqrt{2}}$ to $\frac{2\pi R}{N\sqrt{2}}$. The range is selected in order to avoid any overlap of the objects. Derived from the combinations of this set of parameters, 118 scenes were presented to each observer. Each scene was presented for 3 seconds. Figure 6 gives the examples of the scenes.

There were three advantages of using this kind of synthetic stereoscopic stimuli to investigate the depth-bias:

- Firstly, compared to natural content, synthesis stimuli were easier to control. We could precisely allocate the position and depth of every object in the scene. This

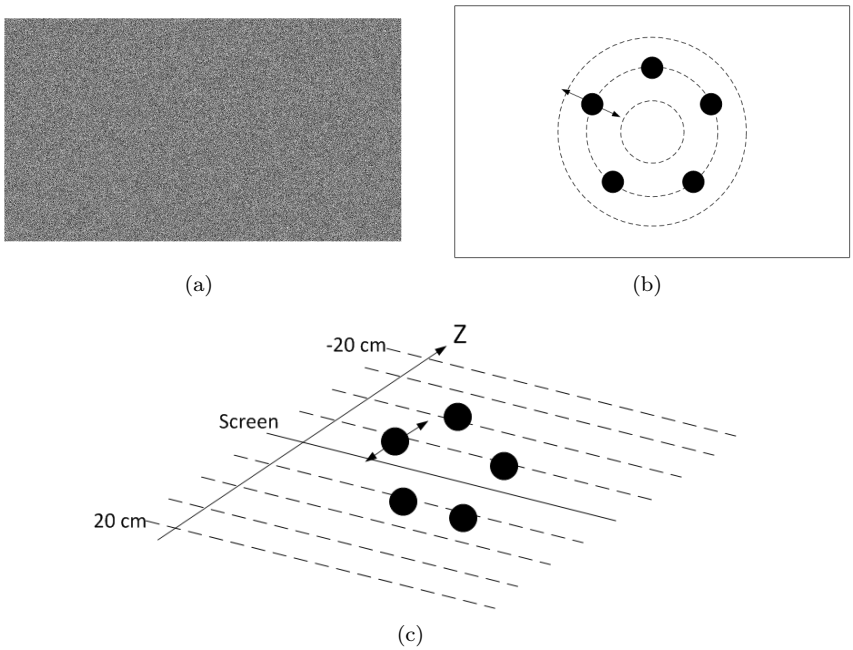



Figure 5: Composition of the stimuli. (a)The background of the stimuli. Only white-noise is contained in the background, which is positioned behind all the stimuli at -20 cm depth plane. (b)Positions of the objects' projections on the screen plane. All the projections laid uniformly on a circle center at the screen center.(c)Allocation of objects in the depth range from -20 cm to 20 cm.

accurate control of the scene could enabled us a better quantitative analysis of eyes movement.

- Secondly, even in 3D viewing, human's eye movements were affected by many bottom-up 2D visual features of the stimuli, such as color, intensity, object's size, and the center-bias. These factors could contaminate our evaluation of depth's influence on visual attention. In our experiment, for each condition, all the objects were with the constant shape, constant size, and constant distance to the center of the screen. This set up let the stimuli get rid of as many bottom-up visual attention features as possible. The white noise background and the simple allocation of objects could also avoid the as much as possible the influence of top-down mechanism in visual attention.

- Third, the complexity of scenes presented to the observers was low, which enable a shorter observation duration. The duration of eye-tracking experiments for natural content images was usually 10 seconds or more. Compared to that, the

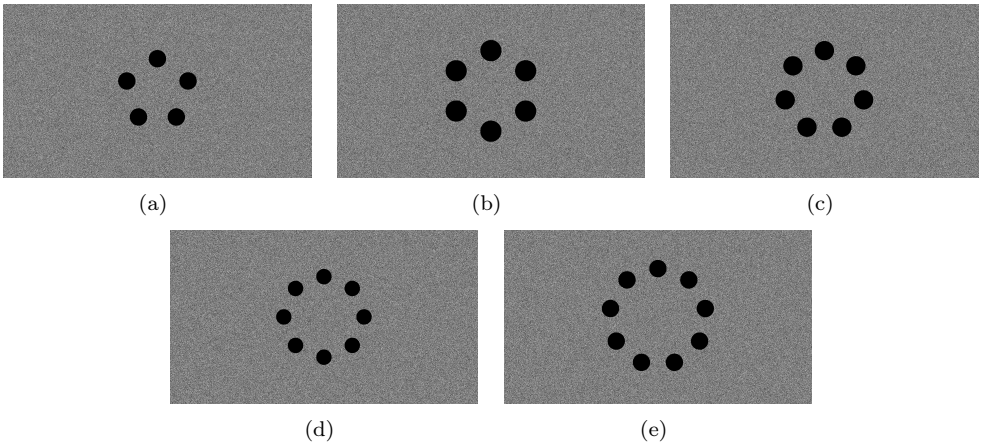



Figure 6: Examples of stimuli with different number of objects used in the eye-tracking experiment.

observation duration time in our experiment was relatively short (3 seconds for each condition), nevertheless, it was still long enough for participants to explore the scene as they want and subconsciously position their fixations of the objects. Hence, using these simple stimuli allowed experimenters to collect more data, as well as learn the evolution of depth-bias over time.

## 3.2 Participants

Twenty-seven subjects participated in the experiment. 12 subjects are male, 15 are female. The subjects ranged in age from 18 to 44 years. The mean age of the subjects was 22.8 years old. All the subjects had either normal or corrected-to-normal visual acuity, which was verified by three pretests before the start of eye-tracking experiment. Monoyer chart was used to check the acuity (subject must get the result higher than 9/10); Ishihara test was used to check the color vision (subject should be without any color troubles); and Randot stereo test was used to check the 3D acuity (subject should get the result higher than 7/10). Among the subjects, 23 of them were students, 3 were university staffs, and 1 software developer. All of them were naive to the purpose of the experiment, and were compensated for the participation of the experiment.

## 3.3 Apparatus and procedures

Stimuli were displayed on a 26-inch (552 mm * 323 mm) Panasonic BT-3DL2550 LCD screen (figure 7 (a)), which had resolution of 1920 * 1200 pixels, and the refresh rate was 60 Hz. Each screen pixel subtended 61.99 arcsec at a 93 cm viewing distance. The maximum luminance of the display was 180 cd/m2, which yielded a maximum luminance of about 60 cd/m2 when watched through the glasses. Observers viewed the

(a)

(b)

(c)

Figure 7: (a)The 26-inch Panasonic BT-3DL2550 LCD screen used in the experiment. (b)SMI

stereoscopic stimuli through a pair of passive polarized glasses at a distance of 93cm. The environment luminance was adjusted according to each observer, in order to let the pupil has an appropriate size for eye-tracking. SMI RED 500 remote eye-tracker was used to record the eye movements (figure 7 (b)).

The viewing distance corresponds to a 33.06 * 18.92 degrees field of view of the background of the stimuli, all the objects were displayed in an area within 10.32*5.91 degrees. A chin-rest was used to stabilize observer's head (figure 7 (c)), and the observers were instructed to "view anywhere on the screen as they want".

All the 118 scenes were presented in a random order. Between every two scene, a

center point was showed for 500 ms at the screen center with zero disparity. A nine-point calibration was performed at the beginning of the experiment, and repeated every twenty scenes. The quality of calibration was verified by the experimenter on another monitor. Participants could require for a rest before every calibration started.

## 3.4 Post processing of eye tracking data

The recorded eye movements were first processed by the Begaze software provided by SMI to identify fixations and filter out saccades. Each fixation was then decided if it was located on one of the objects or not. A fixation was considered to be located on an object if it was positioned on the object or within a surrounding area (10% larger than the object's size). Otherwise, the fixation was considered to be on the background. Therefore, the depth information of each fixation could be obtained. Note that only the 'on target' fixations (the fixations located on a object) were considered in the following analysis.

# 4 Existence of a depth bias on natural images

## 4.1 Experimental condition of occulometric database

The eye tracking dataset provided by Jansen et al. is used in this section [27]. We briefly remind the experimental conditions, i.e. materials and methods to construct this database in 2D and 3D conditions. Stereoscopic images were acquired with a stereo rig composed of two digital cameras. In addition, a 3D laser scanner was used to measure the depth information of these pairs of images. By projecting the acquired depth onto the images and finding the stereo correspondence, disparity maps were then generated. The detailed information relative to stereoscopic and depth acquisition can be found in [32]. The acquisition dataset is composed of 28 stereo images of forest, undistorted, cropped to 1280x1024 pixels, rectified and converted to grayscale. A set of six stimuli was then generated from these image pairs with disparity information: 2D and 3D versions of natural, pink noise and white noise images. Our study focuses only on 2D and 3D version of natural images of forest. In 2D condition two copies of the left images were displayed on an auto stereoscopic display. In 3D condition the left and right image pair was displayed stereoscopically, introducing a binocular disparity to the 2D stimuli.

*The 28 stimulus sets were split-up into 3 training, 1 position calibration and 24 main experiments sets. The training stimuli were necessary to allow the participant to become familiar with the 3D display and the stimulus types. The natural 3D image of the position calibration set was used as reference image for the participants to check their 3D percept.*(cited from Jansen et al.[27])

A 2 view auto stereoscopic 18.1" display (C-s 3D display from SeeReal technologies, Dresden, Germany) was used for stimuli presentation. The main advantage of such display is that it doesn't require special eyeglasses. A tracking system adjusts the two view display to the user position. A beam splitter in front of the LCD panel projects all odd columns to a dedicated angle of view, and all even ones to another. Then, through the tracking system, it ensures the left eye perceives always the odd columns and the right eye the even columns whatever the viewing position. A "3D" effect introducing binocular disparity is then provided by presenting a stereo image pair interlaced vertically. In 2D condition, two identical left images are vertically interlaced. The experiment involved 14 participants. Experiment was split into two sessions, one session comprising a training followed by two presentations separated by a short break. The task involved during presentation is of importance in regards to the literature on visual attention experiments. Here, instructions were given to the subjects to study carefully the images over the whole presentation time of 20s. They were also requested to press a button once they could perceive two depth layers in the image. One subject misunderstood the task and pressed the button in all images. His data were excluded from the analysis. Finally, participants were asked to fixate a cross marker with zero disparity, i.e. on the screen plane, before each stimulus presentation. The fixation corresponding to the prefixation marker was discarded, as each observer started to look at a center fixation cross before the stimuli onset and this would biased the fixation to this region at the first fixation. An "Eyelink II" head-mounted

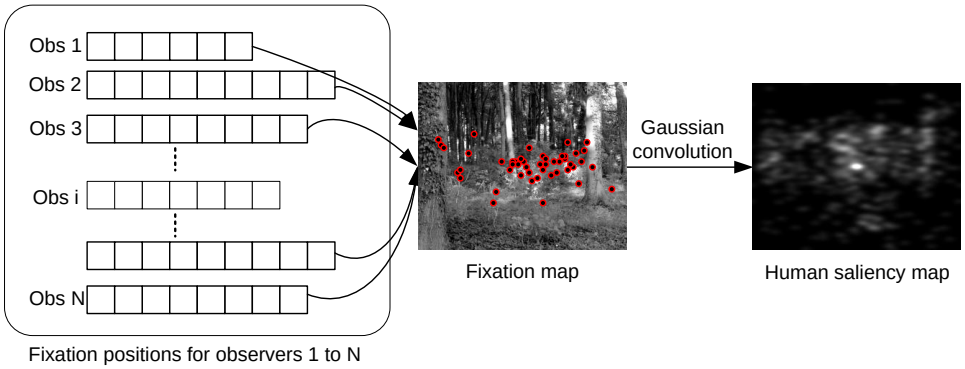



Figure 8: Illustration of the human saliency map computation from N observers

occulometer (SR Research, Osgoode, Ontario, Canada) recorded the eye movements. The eye position was tracked on both eyes, but only the left eye data were recorded; as the stimulus on this left eye was the same in 2D and 3D condition (the left image), the binocular disparity factor was isolated and observable. Observers were placed at 60 cm from the screen. The stimuli presented subtended 34.1° horizontally and 25.9° vertically. Data with an angle less than 3.75° to the monitor frame were cropped. In the following sections, either the spatial coordinates of visual fixations or ground-truth i.e. human saliency map is used. The human saliency map is obtained by convolving a 2D fixation map with a 2D Gaussian with full-width at half-maximum (FWHM) of one degree. This process is illustrated in Figure 8

## 4.2 Behavioral and computational studies

Jansen et al. [27] gave evidence that the introduction of disparity altered the basic properties of eye movement such as rate of fixation, saccade length, saccade dynamics, and fixation duration. They also showed that the presence of disparity influences the overt visual attention especially during the first seconds of viewing. Observers tend to look at closer locations at the beginning of viewing. We go further by examining four points: first we examine whether the disparity impacts the spatial locations of salient areas. Second, we investigate the mean distance between fixations and screen center, i.e. the center bias in 2D and 3D condition. The same examination is done over the depth bias in both viewing conditions. The last question is related to the disparity influence on the of state-of-the-art models performance of bottom-up visual attention.

### 4.2.1 Do salient areas depend on the presence of binocular disparity?

The area under the Receiver Operating Characteristic (ROC) curve is used to quantify the degree of similarity between 2D and 3D human saliency maps. The AUC (Area Under Curve) measure is non-parametric and is bounded by 1 and 0.5. The upper

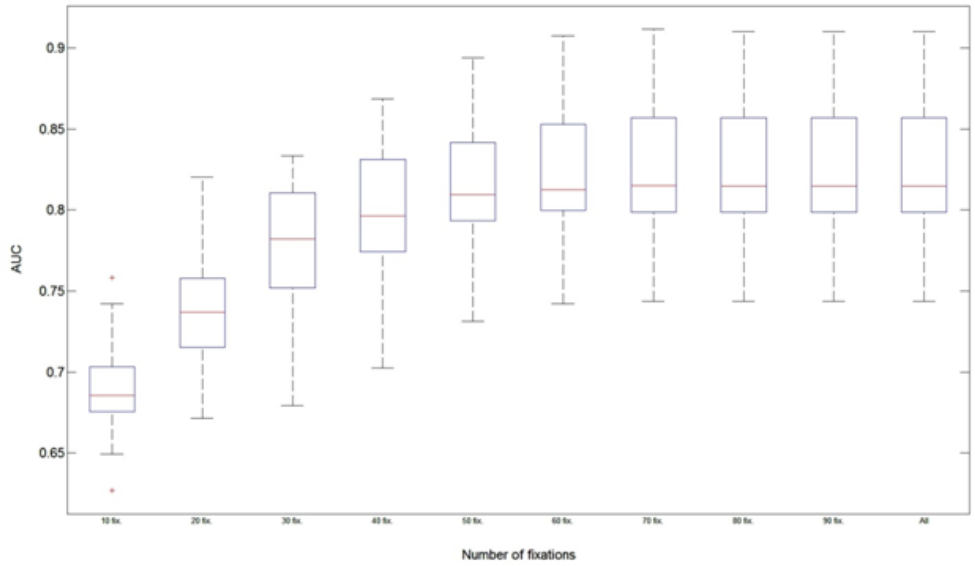



Figure 9: Boxplot of the AUC values between 2D and 3D human (experimental) saliency maps as a function of the number of fixations (the top 20% 2D salient areas are kept).

bound indicates a perfect discrimination whereas the lower bound indicates that the discrimination (or the classification) is at the chance level. The thresholded 3D saliency map is then compared to the 2D saliency map. For the 2D saliency maps taken as reference, the threshold is set in order to keep 20% of the salient areas. For 3D saliency maps, the threshold varies linearly in the range of 0 to 255. Figure 9 shows AUC values in function of the fixation rank. Over the whole viewing time (called "All" on the right-hand side of Figure 9), the AUC value is high. The median value is equal to 0.81 0.008 (mean±SEM). When analyzing only the first fixations, the similarity degree is the lowest. For instance, the similarity increases from 0.68 to 0.81 in a significant manner ($F(1, 23)=1.8$,$p<0.08$, paired $t(23)=13.73$, $p \ll 0.01$). Results suggest that the disparity influences the overt visual attention just after the stimuli onset. This influence significantly lasts up to the first 30 fixations ($F(1, 23)=0.99$,$p<0.49$), paired $t(23)=4.081.64$, $p<0.0001$).

Although the method used to quantify the influence of stereo disparity on the allocation of attention is different from the work of Jansen et al. [27], we draw the same conclusion. The presence of disparity on still pictures has a time-dependent effect on our gaze. During the first seconds of viewing (enclosing the first 30 fixations), there is a significant difference between the 2D and 3D saliency maps.

### 4.2.2 Center bias for 2D and 3D pictures

Previous studies have shown that observers tend to look more at the central regions of a scene displayed on a screen than at the peripheral regions. This tendency might be explained by a number of reasons (see for instance [49]). Recently, Bindemann [6] demonstrated that the center bias is partly due to an experimental artifact stemming from the onscreen presentation of visual scenes. He also showed that this tendency was difficult to remove in a laboratory setting. Does this central bias still exist when viewing 3D scenes? This is the question we address in this section.

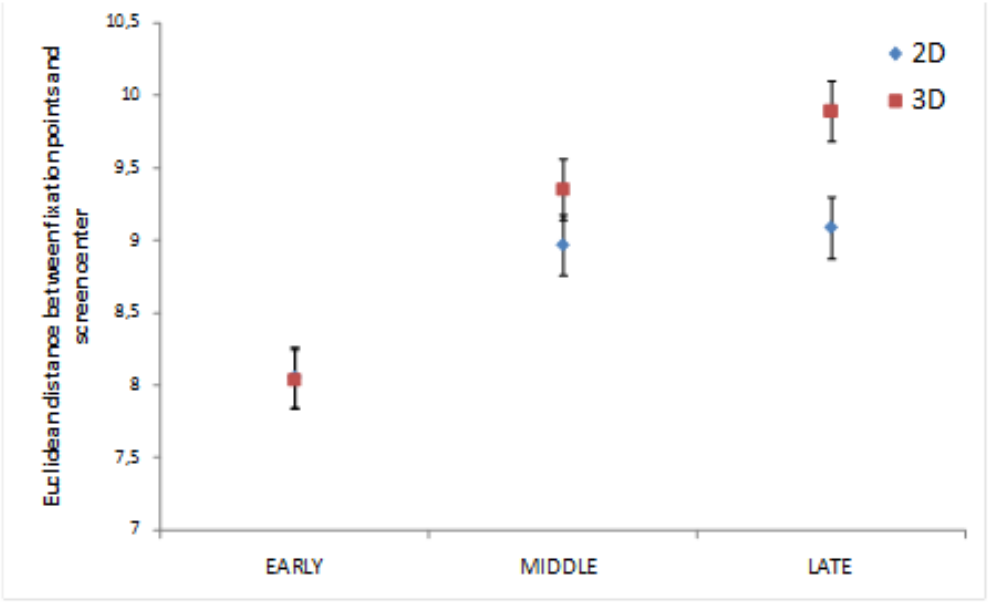



Figure 10: Average Euclidean distance between the screen center and fixation points. The error bars correspond to SEM (Standard Error of the Mean).

When analyzing the fixation distribution, the central bias is observed for both 2D and 3D conditions. The highest values of the distribution are clustered around the center of the screen (see Figure 11 and Figure 12). This bias is more pronounced just after the stimuli onset. To quantify these observations further, a 2x3 ANOVA with the factors 2D-3D (stereoscopy) and three slots of viewing times (called early, middle and late) is applied to the Euclidean distance of the visual fixations to the center of the screen. Each period is composed of ten fixations: early period consists of the first ten fixations, middle the next ten and the late period is composed of the ten fixations occurring after the middle period. A 2x3 ANOVA shows a main effect of the stereoscopy factor $F(1, 6714) = 260.44$ $p<0.001$, a main effect of time $F(2, 6714) = 143.01$ $p<0.001$ and an interaction between both $F(2, 6714) = 87.16$ $p<0.001$. First the influence of viewing time on the center bias is an already known factor. Just after

the stimuli onset, the center bias is more pronounced than after several seconds of viewing. Second there is a significant difference of the central tendency between 2D and 3D conditions and that for the three considered time periods.

Bonferroni t-tests however showed that the central tendency is not statistically significant (2D/3D) for the early periods as illustrated by Figure 3. For the middle and late periods, there is a significant difference in the central bias ($p<0.0001$ and $p\ll0.001$, respectively). The median fixation durations were 272, 272 and 276ms in 2D condition and 276, 272 and 280ms in 3D condition for early, middle and late period respectively.

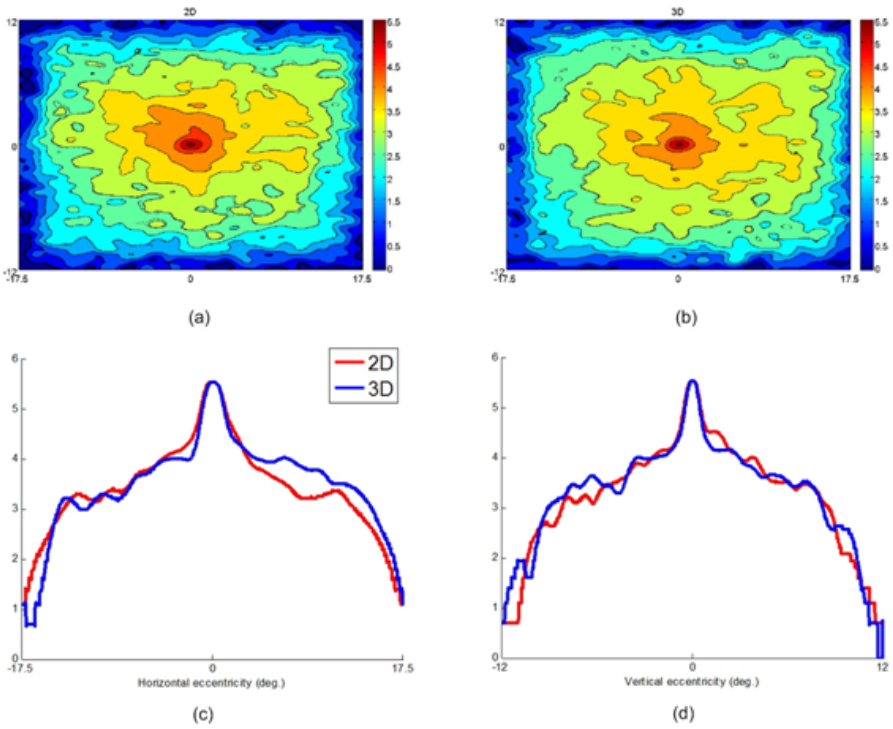



Figure 11: (a) and (b) are the distributions of fixations for 2D and 3D condition, respectively. (c) and (d) represent the horizontal and vertical cross sections through the distribution shown in (a) and (b). All the visual fixations are used to compute the distribution.

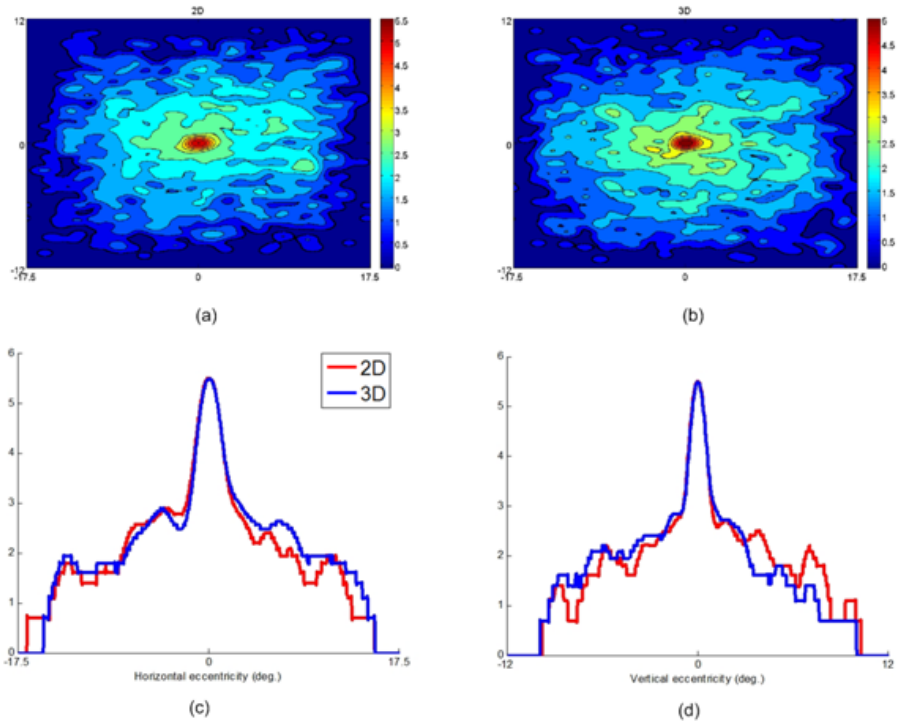



Figure 12: (a) and (b) are the distributions of fixations for 2D and 3D condition, respectively. (c) and (d) represent the horizontal and vertical cross sections through the distribution shown in (a) and (b). All the visual fixations are used to compute the distribution.

### 4.2.3 Depth bias: do we look first at closer locations?

In [27], a depth bias was found out suggesting that observers tend to look more to closer areas just after the stimulus onset than to further areas. A similar investigation is conducted here but with a different approach. Figure 13 illustrates a disparity map: the lowest values represent the closest areas whereas the furthest areas are represented by the highest ones. Importantly, the disparity maps are not normalized and are linearly dependent on the acquired depth.

Figure 13: Original picture (a) and its disparity map (black areas stand for the closest areas whereas the bright areas indicate the farthest ones).

We measured the mean disparity for each fixation point in both conditions (2D and 3D). A neighborhood of one degree of visual angle centered on fixation points is taken in order to account for the fovea size. A 2x3 ANOVA with the factors 2D-3D (stereoscopy) and three slots of viewing times (called early, middle and late) is performed to test the influence of the disparity on the gaze allocation. First the stereoscopy factor is significant $F(1, 6714) = 8.8$ $p<0.003$. The factor time is not significant $F(2,6714)=0.27$ $p<0.76$. Finally, we observed a significant interaction between both factors $F(2,6714)=4.16$ $p<0.05$. Bonferroni t-tests showed that the disparity has an influence at the beginning of the viewing (called early), ($p<0.0001$). There is no difference between 2D and 3D for the two others time periods, as illustrated by Figure 14.

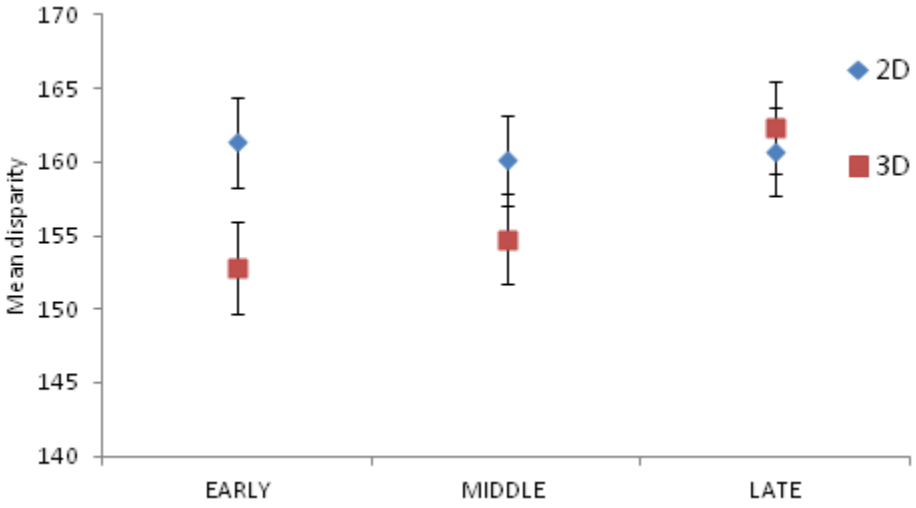



Figure 14: Mean disparity (in pixels) in function of the viewing time (early, middle and late). The error bars correspond to SEM (Standard Error of the Mean).

### 4.2.4 Conclusion

In this behavioral section based on occulometric experiments, we investigated whether the binocular disparity significantly impacts our gaze on still images. It is, especially on the first fixations. This depth cue induced by the stereoscopic condition indeed impacts our gaze strategy: in stereo condition and for the first fixations, we tend to look more at closer locations. These confirm the work of Jansen et al. [27], and support the existence of a depth bias.

# 5 Test to assess the subjective quality of 2D and 3D synthesized contents (with or without compression)

Depth-Image-Based-Rendering algorithms are used for virtual view generation, which is required in both applications. This process induces new types of artifacts. Consequently it impacts on the quality, which has to be identified considering various contexts of use. While many efforts have been dedicated to visual quality assessment in the last twenty years, some issues still remain unsolved in the context of 3DTV. Actually, DIBR is bringing new challenges from the visual quality perspectives mainly because it deals with geometric distortions, which have been barely addressed so far.

Virtual views synthesized either from decoded and distorted data or from original data, need to be assessed. The best assessment tool remains the human judgment as long as the right protocol is used. Subjective quality assessment is still delicate while

addressing new type of conditions because one has to define the optimal way to get reliable data. Tests are time-consuming and consequently one should draw big lines on how to conduct such experiment to save time and observers. Since DIBR is introducing new conditions, the right protocol to assess the visual quality with observers is still an open question. The adequate protocol might vary according to the purpose (impact of compression, DIBR techniques comparison ). Since subjective quality assessment tests are time-consuming, objective metrics have been developed and are extensively used. They are meant to predict human judgment and their reliability is based on their correlation to subjective assessment results. As, the way to conduct the subjective quality assessment protocols is already questionable, reliability of objective quality metrics among existing ones that could be useful in DIBR context, should be tested in the new conditions.

Yet, trustworthy working groups base partially their future specifications, concerning new strategies for 3D video, on the outcome of objective metrics. Considering the test conditions may rely on usual subjective and objective protocols (because of their availability), the outcome of wrong choices could result to a poor quality of experience for users. Then, new tests should be carried on to determine the reliability of subjective and objective quality assessment tools in order to exploit their results for the best.

In this study, we propose to answer two questions: First, how adapted are the used subjective assessment protocols in the case of DIBR-based rendered virtual views? Second, is there a correlation between commonly used 2D video metrics scores and subjective scores when evaluating the quality of DIBR-based rendered virtual views? Indeed, we first address the 2D conditions because it is a first step that should be studied, before including parameters such as 3D vision, that are not completely understood at the moment.

## 5.1 New artifacts related to DIBR

As explained in the introduction, DIBR is brings new types of artifact, different from those commonly encountered in video compression: most video coding standards rely on DCT, and the resulting artifacts are specific (some of them are described in [61]). Artifacts brought by DIBR are mainly geometric distortions. They are related to two causes: the accuracy of incoming data values (e.g. depth estimation accuracy) and the synthesis process strategies. Note that they are also different from stereoscopic impairments (such as cardboard effect, crosstalk, etc. as described in [33]), which occur in stereoscopic conditions (fusion of left and right views in human visual system). Synthesis process strategies mainly aimed at dealing with the critical problem in DIBR, namely the disocclusion: when generating a new viewpoint, areas that were not visible in the reference viewpoint, become visible in the new point. They are discovered. There is no available color in-formation to fill in these areas, which leads to geometric distortions. Extrapolation techniques are meant to fill the disoccluded regions.

In this section, typical DIBR artifacts are described. In most of the cases, these artifacts are located around large depth discontinuities, but they are more perceptible in case of high texture contrast between background and foreground.

*Object shifting*: a region may be slightly translated or resized, depending on the

chosen extrapolation method (if the method chooses to assign the background values to the missing areas, object may be resized), or on the encoding method (blocking artifacts in depth data result in object shifting in synthesis). Figure 15 depicts this type of artifact.

(a) Original frame (b) Synthesized frame

Figure 15: Shifting/Resizing artifacts

*Blurry regions*: This may be due to the inpainting method used to fill the disoccluded areas. It is obvious around the background/foreground transitions. These remarks are confirmed on Figure 16 around the disoccluded areas.

(a) Original frame (b) Synthesized frame

Figure 16: Blurring artifacts (Book Arrival)

*Texture synthesis*: inpainting methods can fail in filling complex textured areas. To overcome these limitations, a hole filling approach based on patch-based texture

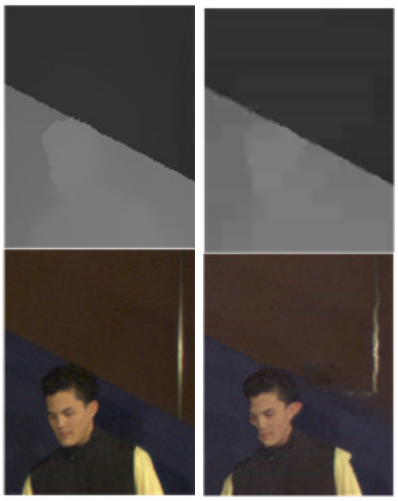
synthesis is proposed in [35].

*Flickering*: when errors occur randomly in depth data along the sequence, pixels are wrongly projected: some pixels suffer slight changes of depth, which appears as flickers in the resulting synthesized pixels. To avoid this methods such as [30] propose to acquire background knowledge along the sequence and to conse-quently improve the synthesis process.

*Tiny distortions*: in synthesized sequences, a large number of tiny geometric distortions and illumination differences are temporally constant and perceptually invisible. However, pixel-based metrics may penalize these distorted zones.

When encoding either depth data or color sequences before performing the synthesis, compression-related artifacts are combined with synthesis artifacts. Artifacts from data compression are generally scattered within the whole image, while artifacts inherent to the synthesis process are mainly located around the disoccluded areas. The combination of both type of distortion, depending on the compression method, relatively affects the synthesized view. Indeed, most of the used compression methods are 2D video codecs inspired, and are thus optimized for human perception of color. As a result, artifacts occurring especially in depth data induce severe distortions in the synthesized views. In the following, a few examples of such distortions are presented.

*Blocking artifacts*: this occurs when the compression method induces blocking artifacts in depth data. In the synthesized views, whole blocks of color image seem to be translated. Figure 17 illustrates the distortion.



(a) Original depth frame (up) and color original frame (bottom)

(b) Decoded depth frame (up) and resulting synthesized frame (bottom)

Figure 17: Blocking artifacts from depth data compression result in distorted synthesized views (Breakdancers).

*Ringing artifacts*: when ringing artifacts occur in depth data around strong discontinuities, objects edges appear distorted in the synthesized view. Figure 18 depicts this artifact.

(a) Original depth frame (up) and original color frame (bottom)

(b) Distorted depth frame (up) and resulting synthesized frame (bottom)

Figure 18: Ringing artifacts in depth data lead to distortions in the synthesized views.

## 5.2 Subjective quality assessment methodologies

In the absence of any better 3D-adapted subjective quality assessment methodologies, the evaluation of synthesized views is mostly obtained through 2D validated protocols. Different methods were developed by the ITU-R and ITU-T. The appropriate method is selected, according to one objective. Indeed the methods differ depending on the type of distortion and on the evaluation. In the case of synthesized views evaluation, one should choose the adequate subjective method. This section introduces two reliable 2D subjective quality assessment methodologies, based on the proposed classification of subjective test protocols described in [5], and on the methods described in [9]. Then requirements for 3D-adapted subjective quality assessment protocols are presented.

**Absolute categorical rating with Hidden Reference Removal (ACR-HR)** methodology [4] consists in presenting test objects (i.e. images or sequences) to observers. The objects are presented one at a time and in a random order, to the observers. Observers score the test item according to a discrete category rating scale. From the scores obtained, a differential score (DMOS for Differential Mean Opinion Score) is computed between the mean opinion scores (MOS) of each test object and its associated hidden reference. The quality scale recommended by ITU-R is depicted

| | |
|---|---|
| 5 | Excellent |
| 4 | Good |
| 3 | Fair |
| 2 | Poor |
| 1 | Bad |

Table 2: Comparison scale for ACR-HR

in Table 2. ACR-HR methodology is a single stimulus method.

The results of an ACR-HR test are obtained by averaging observers' opinion scores for each stimulus, in other words, by computing mean opinion scores (MOS). ACR-HR requires many observers to minimize the contextual effects (previously presented stimuli influence the observer opinion, i.e. presentation order influences opinion ratings). Accuracy increases with the number of participants.

**Paired comparisons (PC)** methodology [4] is an assessment protocol in which stimuli are presented by pairs to the observers: it is a double-stimulus method. The latter select the one out of the pair that best satisfies the specified judgment criterion, i.e. image quality. The results of a paired comparisons test are recorded in a matrix: each element corresponds to the frequencies a stimulus is preferred over another stimulus. These data are then converted to scale values using Thurstone-Mosteller's or Bradley-Terry's model [22]. It leads to a hypothetical perceptual continuum. The presented experiments follow Thurstone-Mosteller's model where naive observers were asked to choose the preferred item from one pair. Although the method is known to be highly accurate, it is time consuming.

The differences between ACR-HR and PC are of different types. First, with ACR-HR, even though they may be included in the stimuli, the reference sequences are not identified as such by the observers. Observers provide an absolute vote without any reference. In PC, observers only need to indicate their preference out of a pair of stimuli. Then the requested task is different: while observers assess the quality of the stimuli in ACR-HR, they just provide their preferences in PC.

The quality scale is another issue. ACR-HR scores provide knowledge on the perceived quality level of the stimuli. However the voting scale is coarse, and because of the single stimulus presentation, observers cannot remember previous stimuli and precisely evaluate small impairments. PC scores (i.e. ‚Äúpreference matrices‚Äù) are scaled to a hypothetical perceptual continuum. However, it does not provide knowledge on the quality level of the stimuli, but on the stimuli order of preferences. Moreover, PC is very well suited for small impairments, thanks to the fact that only two conditions are compared to each other. For these reasons, PC tests are often coupled with ACR-HR tests.

Another aspect concerns the complexity and the feasibility of the test: PC is simple because observers only need to provide preference in each double stimulus. However, when the number of stimuli increase, the test becomes hardly feasible as the number of

comparisons grows as , with N, the number of stimuli. In the case of video sequences assessment, a double-stimulus method such as PC involves the use of either one split-screen environment (or two full screens), with the risk of distracting the observer (as explained in [39]), or one screen but sequences are displayed one after the other, which increases the length of the test. On the other hand, the simplicity of ACR-HR allows the assessment of a larger number of stimuli. However, the results of this assessment are reliable as long as the group of participants is large enough.

## 5.3 Objective quality assessment metrics

Objective metrics are meant to predict human perception of quality of images and thus avoid spending time in subjective quality assessment tests. They are then supposed to be highly correlated with human opinion. In the absence of approved metrics for assessing synthesized views, most of the studies rely on the use of 2D validated metrics, or on adaptations of such. There are different types of objective metrics, depending on their requirement for reference images.

**Full reference methods (FR)**: these methods require references images. Most of the existing metrics rely on FR methods.

**Reduced reference methods (RR)**: these methods require only elements of the reference images.

**No-reference methods (NR)**: these methods do not require reference images. They mostly rely on Human Visual System models to predict human opinion of the quality. Also, a prior knowledge on the expected artifacts highly improves the design of such methods.

A widely used FR metric, is Peak Signal to Noise Ratio (PSNR), because of its simplicity. It measures the signal fidelity of a distorted image compared to a reference. It is based on the measure of the Mean Squared Error (MSE). However because of the pixel-based approach of such a method, the amount of distorted pixels is depicted, but the perceptual quality is not: PSNR does not take into account the visual masking phenomenon. Thus, even if an error is not perceptible, it contributes to the decrease of the quality score. Indeed, studies (such as [7]) showed that in the case of synthesized views, PSNR is not reliable, especially when comparing two images with low PSNR scores.

As an alternative to pixel-based methods, Universal Quality Index UQI [56] is a perceptual-like metric. It models the image distortion by a combination of three factors: loss of correlation, luminance distortion, and contrast distortion.

PSNR-HVS [18], based on PSNR and UQI, is meant to take into account the Hu-man Visual System (HVS) properties.

PSNR-HVSM [40] is based on PSNR but takes into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions.

Single-scale Structural SIMilarity (SSIM) [57] is considered as an extension of UQI. It combines image structural information: mean, variance, covariance of pixels, for a single local patch. The blocksize depends on the viewer distance to the screen.Multi-scale SSIM (MSSIM) is the average SSIM scores of all patches of the image Visual

Signal to Noise Ratio (VSNR) [12] is also a perceptual-like metric: it is based on a visual detection of distortion criterion, helped by CSF.

Weighted Signal to Noise Ratio (WSNR) that uses a weighting function adapted to HVS denotes a weighted Signal to Noise Ratio, as applied in [16]. Information Fidelity Criterion (IFC) [46] uses a distortion model to evaluate the information shared between the reference image and the degraded image. This method has been improved by the introduction of a HVS model. The method is called Visual Information Fidelity (VIF) [45]. VIFP is a pixel-based version of VIF. Noise quality measure (NQM) quantifies the injected noise in the tested im-age.

Video Structural Similarity Measure (V-SSIM) [58] is a FR video quality metric which uses structural distortion as an estimate of perceived visual distortion. At the patch level, SSIM score is a weighted function of SSIM of the different component of the image (i.e. luminance, and chromas). At the frame level, SSIM score is a weighted function of patches, SSIM scores (based on the darkness of the patch). Finally at the sequence level, VSSIM score is a weighted function of frames, SSIM scores (based on the motion).

Video Quality Metric (VQM) was proposed by Pinson and Wolf in [39]. It is a RR video metric that measures perceptual effects of numerous video distortions. It includes a calibration step (to correct spatial/temporal shift, contrast, and brightness according to the reference video sequence), an analysis of perceptual features. VQM score combines all the perceptual calculated parameters.

Perceptual Video Quality Measure (PVQM) [23] is meant to detect perceptible distortions in video sequences. Different indicators are used. First, an edge-based indicator allows the detection of distorted edges in the images. Second, a motion-based indicator analyses two successive frames. Third, a color-based indicator de-tects non-saturated colors. Each indicator is pooled separately ACR-HRoss the video and incorporated in a weighting function to obtain the final score.

Moving Pictures Quality Metric (MPQM) [53] uses a HVS model. In particular it takes into account the masking phenomenon and the contrast sensitivity.

Motion-based Video Integrity Evaluation (MOVIE) is a FR video metric that uses several steps before computing the quality score. It includes the decomposition of both reference and distorted video by using a multi-scale spatio-temporal Gabor filter-bank. A SSIM-like method is used for the spatial quality analysis. An optical flow calculation is used for the motion analysis. Spatial and temporal quality indi-cators determine the final score.

Only a few commonly used algorithms (in the 2D context) have been described above. There exist many other algorithms for visual quality assessment that are not covered here.

## 5.4 Experimental material

Standardized methodologies for subjective multimedia quality assessment, such as Paired Comparisons (PC) and Absolute Categorical Rating (ACR-HR), have proved their efficiency regarding the quality evaluation of 2D conventional images. Then, a simple assumption is that the two aforementioned methodologies should be suitable

for evaluating the quality of images synthesized from DIBR al-gorithms in 2D conditions. The hypothesis is studied in the following experimental protocol. Seven DIBR algorithms processed three test sequences to generate, for each one, four different viewpoints.

These seven DIBR algorithms are referenced from A1 to A7:

- A1: based on Fehn [20], where the depth map is pre-processed by a low-pass filter. Borders are cropped, and then an interpolation is processed to reach the original size.
- A2: based on Fehn [20]. Borders are inpainted by the method proposed by Telea [50].
- A3: Tanimoto *et al.* [47], it is the recently adopted reference software for the experiments in the 3D Video group of MPEG.
- A4: Mueller *et al.* [34], proposed a hole filling method aided by depth information.
- A5: Ndjiki-Nya *et al.* [4], the hole filling method is a patch-based texture synthesis.
- A6: Koeppel *et al.* [5], uses depth temporal information to improve the synthesis in the disoccluded areas.
- A7: corresponds to the unfilled sequences (i.e. with holes).

The sequences are Book Arrival (1024×768, 16 cameras with 6.5cm spacing), Lovebird1 (1024×768, 12 cameras with 3.5 cm spacing) and Newspaper (1024×768, 9 cameras with 5 cm spacing). The test was conducted in an ITU conforming test environment. ACR-HR and Paired comparisons were used to collect perceived quality scores. Paired comparisons were run only for still images evaluation. The stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [7].

## 5.5 Analysis of subjective scores on still images and video sequences

This section consists of a case study whose goal is to answer the question: are usual requirements for subjective evaluation protocols still appropriate for assess-ing 3D synthesized views? A first approach is to be independent from the 3D, that is to say both the stereopsis and the 3D display whose technology is still a major factor of visual quality degradation, as explained in previous sections. Thus, the case study presented in this section focuses on the quality evaluation of DIBR-based synthesized views in 2D conditions. Besides, these conditions are plausible in a Free Viewpoint Video (FVV) application.

### 5.5.1 Results on still images

Watching a still image synthesized from DIBR methods is a plausible scenario in FVV, and it can also be considered as preliminary results for synthesized video sequences quality assessment. Thus, still images quality deserves to be evaluated. First experiments were conducted only over ‚Äúkey‚Äù frames, due to the complexity of PC tests when number of items increases, and the length of both protocols. That is to say that for each of the three reference sequences, only one frame was selected. For a given reference video sequence, each one of the seven DIBR algorithms generated four intermediate viewpoints (that is 84 synthesized sequences in total). ACR-HR was performed over the whole set of selected frames. For PC, each pair consists of two of the selected frames, synthesized with two different DIBR algorithms. Then, for the twelve synthesized sequences, twelve 7×7 preference matrices were processed, for PC test. Figure 19 shows regions of the synthesized frames with the different DIBR algorithms. Forty-three naive observers partici-pated in this test.

The seven DIBR algorithms are ranked according to the obtained ACR-HR and PC scores, as depicted in Table 3. This table indicates that the rankings obtained by both testing method are consistent. For both type of test, first line gives the MOS score and second line gives the rankings of the algorithms, obtained through the MOS scores.

| | *A1* | *A2* | *A3* | *A4* | *A5* | *A6* | *A7* |
|---|---|---|---|---|---|---|---|
| ACR-HR | 2.388 | 2.234 | 1.994 | 2.250 | 2.345 | 2.169 | 1.126 |
| Rank order | 1 | 4 | 6 | 3 | 2 | 5 | 7 |
| PC | 1.038 | 0.508 | 0.207 | 0.531 | 0.936 | 0.45 | -2.055 |
| Rank order | 1 | 4 | 6 | 3 | 2 | 5 | 7 |

Table 3: Rankings of algorithms according to subjective scores

In Table 3, although the algorithms can be ranked from the scaled scores, there is no information concerning the statistical significance of the quality difference of two stimuli (one more preferred than another one). Then statistical analyses have been conducted over the subjective measurements: a student's t-test has been performed over ACR-HR scores, and over PC scores for each algorithm. This provides knowledge on the statistical equivalence of the algorithms. Table 3 and Table 4 show the results of the statistical tests over ACR-HR and PC values respectively. In both tables, the number in parentheses indicates the minimum required number of observers that allows statistical distinction (VQEG recommends 24 participants as a minimum [3], values in bold are higher than 24 in the table).

A first analysis of these two tables indicates that the PC test leads to clear-cut decisions, compared to ACR-HR test: indeed, the distributions of the algorithms are statistically distinguished with less than 24 participants in 17 cases with PC (only 11 cases with ACR-HR). In one case (between A2 and A5), less than 24 participants are required with PC, and more than 43 participants are required to establish the statistical difference with ACR-HR. The latter case can be explained by the fact that the visual quality of the synthesized images (and thus, some distortions) may seem very similar for non-expert observers. This makes the ACR-HR test more tough for observers. These results indicate that it seems more difficult to assess the quality of

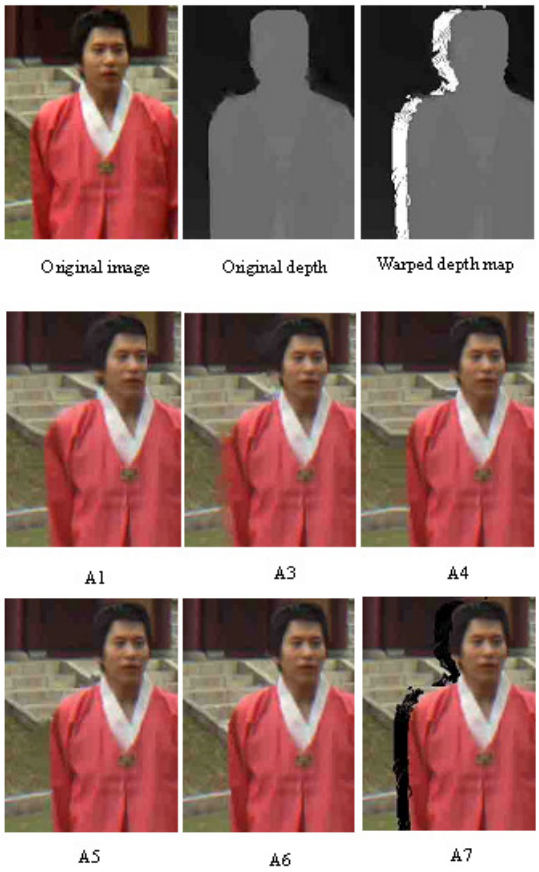



Figure 19: DIBR-based synthesized frames of the "Lovebird1" sequence

synthesized views than in other contexts (for instance, quality assessment of images distorted through compression). Indeed, the results with ACR-HR test, in Table 3, confirm this idea: in most of the cases, more than 24 participants (or even more than 43) are required to distinguish the classes (Remember that A7 is the synthesis with holes around the disoccluded areas). However, as seen with rankings results above, methodologies give consistent results: when algorithms distinctions are stable, they are the same with both methodologies.

Finally, these experiments show that fewer participants are required for a PC test than for an ACR-HR test. However, as stated before, PC tests, while efficient, are feasible only with a limited number of items to be compared. Another problem, pointed out by these experiments, concerns the assessment of similar items: with both methods, 43 participants were not always sufficient to obtain a stable and reliable decision. Results suggest that observers had difficulties assessing the different types of artefacts.

| | $A1$ | $A2$ | $A3$ | $A4$ | $A5$ | $A6$ | $A7$ |
|---|---|---|---|---|---|---|---|
| $A1$ | | ↑**(32)** | ↑(<24) | ↑**(32)** | o **(>43)** | ↑**(30)** | ↑(<24) |
| $A2$ | ↓**(32)** | | ↑(<24) | o **(>43)** | o **(>43)** | o **(>43)** | ↑(<24) |
| $A3$ | ↓(<24) | ↓(<24) | | ↓(<24) | ↓(<24) | ↓(<24) | ↑(<24) |
| $A4$ | ↓**(32)** | o**(>43)** | ↑(<24) | | o**(>43)** | o**(>43)** | ↑(<24) |
| $A5$ | o**(>43)** | o**(>43)** | ↑(<24) | o**(>43)** | | ↑**(28)** | ↑(<24) |
| $A6$ | ↓**(30)** | o**(>43)** | ↑(<24) | o **(>43)** | ↓**(28)** | | ↑(<24) |
| $A7$ | ↓(<24) | ↓(<24) | ↓(<24) | ↓ (<24) | ↓(<24) | ↓(<24) | |

Table 4: Results of Student's t-test with ACR-HR results. Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.

| | $A1$ | $A2$ | $A3$ | $A4$ | $A5$ | $A6$ | $A7$ |
|---|---|---|---|---|---|---|---|
| $A1$ | | ↑(<24) | ↑(<24) | ↑(<24) | ↑(<24) | ↑(<24) | ↑(<24) |
| $A2$ | ↓(<24) | | ↑**(28)** | o(<24) | ↓(<24) | o**(>43)** | ↑(<24) |
| $A3$ | ↓(<24) | ↓**(28)** | | ↓(<24) | ↓(<24) | ↓(<24) | ↑(<24) |
| $A4$ | ↓(<24) | o**(>43)** | ↑(<24) | | ↓(<24) | ↑**(43)** | ↑(<24) |
| $A5$ | ↓(<24) | ↑(<24) | ↑(<24) | ↑(<24) | | ↑(<24) | ↑(<24) |
| $A6$ | ↓(<24) | o**(>43)** | ↑(<24) | ↓**(<43)** | ↓(<24) | | ↑(<24) |
| $A7$ | ↓(<24) | ↓(<24) | ↓(<24) | ↓(<24) | ↓(<24) | ↓(<24) | |

Table 5: Results of Student's t-test with Paired comparisons results. Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "less than 24" observers participate.

As a conclusion, this first analysis, involving still images quality assessment, reveals that more than 24 participants may be necessary for these types of test. PC gives clear-cut decisions, due to the mode of assessment (preference) while algorithm's statistical distinctions with ACR-HR are slightly less accurate. However, ACR-HR and PC are complementary: when assessing similar items, like in this case study, PC can provide a ranking, while ACR-HR gives the overall perceived quality of the items.

### 5.5.2 Results on video sequences

In the case of video sequences, only ACR-HR test was conducted, as mentioned before. PC test with video sequences would have required either two screens, or switching between items. In the case of the use of two screens, it involves the risk of missing frames of the tested sequences, because one cannot watch simulta-neously two different video sequences. In the case of the switch, it would have in-creased considerably the length of the test. The test concerns the 84 sequences synthesized from the seven DIBR algo-rithms. Thirty-two naive observers participated in this test. Table 5 shows the algorithms ranking from the obtained subjective scores. The ranking order differs from the one obtained with ACR-HR test in the still image context and the MOS values slightly vary.

| | $A1$ | $A2$ | $A3$ | $A4$ | $A5$ | $A6$ | $A7$ |
|---|---|---|---|---|---|---|---|
| ACR-HR | 2.70 | 2.41 | 2.14 | 2.03 | 1.96 | 2.13 | 1.28 |
| Rank order | 1 | 2 | 3 | 5 | 6 | 4 | 7 |

Table 6: Rankings of algorithms according to subjective scores

| | *A1* | *A2* | *A3* | *A4* | *A5* | *A6* | *A7* |
|---|---|---|---|---|---|---|---|
| *A1* | | ↑(7) | ↑(3) | ↑(3) | ↑(2) | ↑(3) | ↑(1) |
| *A2* | ↓(7) | | ↑(2) | ↑(2) | ↑(1) | ↑(2) | ↑(1) |
| *A3* | ↓(3) | ↓(2) | | O(>32) | ↑(9) | O(>32) | ↑(1) |
| *A4* | ↓(3) | ↓(2) | O(>32) | | O(>32) | O(>32) | ↑(1) |
| *A5* | ↓(2) | ↓(1) | ↓(9) | O>32) | | ↓(15) | ↑(1) |
| *A6* | ↓(3) | ↓(2) | O>32) | O(>32) | ↑(15) | | ↑(1) |
| *A7* | ↓(1) | ↓(1) | ↓(1) | ↓(1) | ↓(1) | ↓(1) | |

Table 7: Results of Student's t-test with ACR-HR results. Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line1 is statistically superior to column 2. Distinction is stable when 7 observers participate.

And, still, although the values allow the ranking of the algorithms, they do not directly provide knowledge on the statistical equivalence of the results. Table 6 depicts the results of the Student's t-test processed with the values. Compared to ACR-HR test with still images (section 5.5.1), distinctions between algorithms seem to be more obvious. Statistical significance of the difference between the algorithms, based on the ACR-HR scores, exists and seems clearer in the case of the video se-quences than in the case of still images. This can be explained by the exhibition time of the video sequences: watching the whole video, observers can refine their judgment, compared to still images. Note that the same algorithms were not statis-tically differentiated: A4, A3, A5 and A6.

As a conclusion, ACR-HR test with video sequences give clearer statistical differences between the algorithms than ACR-HR test with still images. This suggests that new elements allow the observers to make a decision: existence of flickering, exhibition time, etc. Results of student's test with still images are confirmed with video sequences.

## 5.6 Analysis of the objective scores on still images and video sequences

A second case study concerns the reliability of usual objective metrics. The latter is questioned regarding metrics ability to accurately assess the quality of views synthesized from DIBR algorithms. To answer this question, the performances of the seven synthesis methods are evaluated in the following experiments. The same experimental material as in Section 1.5.1 was used. The objective measurements were realized over the 84 synthesized views by the means of MetriX MuX Visual Quality Assessment Package [1] except for two metrics: VQM and VSSIM. VQM were available at [2]; VSSIM was imple-mented by the authors, according to [58]. The reference was the original acquired image. For video sequences, still image metrics were applied on each frames of the sequences and then averaged by the number of frames.

### 5.6.1 Results on still images

The results of this subsection concerns the measurements conducted over the same selected ‚Äúkey‚Äù frames as in section 5.5. The whole set of objective metrics give the same trends. Table 8 provides correlation coefficients between obtained objective scores. It reveals that they are highly correlated. Note the high correlation scores between pixel-based and more perceptual-like metrics such as PSNR and SSIM (83.9%). The first step consists in comparing the objective scores with the subjective scores (in section 5.5). The consistency between objective and subjective measures is evaluated by calculating the correlation coefficients for the whole fitted measured points. The coefficients are presented in Table 9. In the results of our test, none of the tested metric reaches 50% of human judgment. This reveals that contrary to the received opinion, the objective tested metrics, whose efficiency has been proved for the quality assessment of 2D conventional media, do not reliably predict human appreciation in the case of synthesized views. Table 10 presents the rankings of the algorithms, obtained from the objective scores. Rankings from subjective scores are mentioned for comparison. They present a noticeable difference concerning the ranking order of A1: judged as the best algorithm out of the seven by the subjective scores, it is ranked as the last by the whole set of objective metrics. Another comment refers to the assessment of A6: often judged as the best algorithm, it is judged as one of the worst algorithms through the subjective tests. The ensuing assumption is that objective metrics detect and penalize non-annoying artifacts.

| | PSNR | SSIM | MSSIM | VSNR | VIF | VIFP | UQI | IFC | NQM | WSNR | PSNR $_{hsvm}$ | PSNR $_{hsv}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR | | 83.9 | 79.6 | 87.3 | 77.0 | 70.6 | 53.6 | 71.6 | 95.2 | 98.2 | 99.2 | 99.0 |
| SSIM | 83.9 | | 96.7 | 93.9 | 93.4 | 92.4 | 81.5 | 92.9 | 84.9 | 83.7 | 83.2 | 83.5 |
| MSSIM | 79.6 | 96.7 | | 89.7 | 88.8 | 90.2 | 86.3 | 89.4 | 85.6 | 81.1 | 77.9 | 78.3 |
| VSNR | 87.3 | 93.9 | 89.7 | | 87.9 | 83.3 | 71.9 | 84.0 | 85.3 | 85.5 | 86.1 | 85.8 |
| VIF | 77.0 | 93.4 | 88.8 | 87.9 | | 97.5 | 75.2 | 98.7 | 74.4 | 78.1 | 79.4 | 80.2 |
| VIFP | 70.6 | 92.4 | 90.2 | 83.3 | 97.5 | | 85.9 | 99.2 | 73.6 | 75.0 | 72.2 | 72.9 |
| UQI | 53.6 | 81.5 | 86.3 | 71.9 | 75.2 | 85.9 | | 81.9 | 70.2 | 61.8 | 50.9 | 50.8 |
| IFC | 71.6 | 92.9 | 89.4 | 84.0 | 98.7 | 99.2 | 81.9 | | 72.8 | 74.4 | 73.5 | 74.4 |
| NQM | 95.2 | 84.9 | 85.6 | 85.3 | 74.4 | 73.6 | 70.2 | 72.8 | | 97.1 | 92.3 | 91.8 |
| WSNR | 98.2 | 83.7 | 81.1 | 85.5 | 78.1 | 75.0 | 61.8 | 74.4 | 97.1 | | 97.4 | 97.1 |
| PSNR $_{hsvm}$ | 99.2 | 83.2 | 77.9 | 86.1 | 79.4 | 72.2 | 50.9 | 73.5 | 92.3 | 97.4 | | 99.9 |
| PSNR $_{hsv}$ | 99.0 | 83.5 | 78.3 | 85.8 | 80.2 | 72.9 | 50.8 | 74.4 | 91.8 | 97.1 | 99.9 | |

Table 8: Correlation coefficients between objective metrics in percentage.

| | PSNR | SSIM | MSSIM | WSNR | VIF | VIFP | UQI | IFC | NQM | WSNR | PSNR $_{HVSM}$ | PSNR $_{HVS}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CC$_{MOS}$ | 38.6 | 21.9 | 16.1 | 25.8 | 19.3 | 19.2 | 20.2 | 19.0 | 38.6 | 42.3 | 38.1 | 37.3 |
| CC$_{PC}$ | 40.0 | 23.8 | 34.9 | 19.7 | 16.2 | 22.0 | 32.9 | 20.1 | 37.8 | 36.9 | 42.2 | 41.9 |

Table 9: Correlation coefficients between MOS and objective scores in percentage.

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
|---|---|---|---|---|---|---|---|
| MOS | 2.388 | 2.234 | 1.994 | 2.250 | 2.345 | 2.169 | 1.126 |
| Rank order | 1 | 4 | 6 | 3 | 2 | 5 | 7 |
| PC | 1.4038 | 0.5081 | 0.2073 | 0.5311 | 0.9363 | 0.4540 | -2.0547 |
| Rank order | 1 | 4 | 6 | 3 | 2 | 5 | 7 |
| PSNR | 18.752 | 24.998 | 23.180 | 26.117 | 26.171 | 26.177 | 20.307 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| SSIM | 0.638 | 0.843 | 0.786 | 0.859 | 0.859 | 0.858 | 0.821 |
| Rank order | 7 | 4 | 6 | 1 | 1 | 3 | 5 |
| MSSIM | 0.648 | 0.932 | 0.826 | 0.950 | 0.949 | 0.949 | 0.883 |
| Rank order | 7 | 4 | 6 | 1 | 2 | 2 | 5 |
| VSNR | 13.135 | 20.530 | 18.901 | 22.004 | 22.247 | 22.195 | 21.055 |
| Rank order | 7 | 5 | 6 | 3 | 1 | 2 | 4 |
| VIF | 0.124 | 0.394 | 0.314 | 0.425 | 0.425 | 0.426 | 0.397 |
| Rank order | 7 | 5 | 6 | 2 | 2 | 1 | 4 |
| VIFP | 0.147 | 0.416 | 0.344 | 0.448 | 0.448 | 0.448 | 0.420 |
| Rank order | 7 | 5 | 6 | 1 | 1 | 1 | 4 |
| UQI | 0.237 | 0.556 | 0.474 | 0.577 | 0.576 | 0.577 | 0.558 |
| Rank order | 7 | 5 | 6 | 1 | 3 | 1 | 4 |
| IFC | 0.757 | 2.420 | 1.959 | 2.587 | 2.586 | 2.591 | 2.423 |
| Rank order | 7 | 5 | 6 | 2 | 3 | 1 | 4 |
| NQM | 8.713 | 16.334 | 13.645 | 17.074 | 17.198 | 17.201 | 10.291 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| WSNR | 13.817 | 20.593 | 18.517 | 21.597 | 21.697 | 21.716 | 15.588 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| PSNR HSVM | 13.772 | 19.959 | 18.362 | 21.428 | 21.458 | 21.491 | 15.714 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| PSNR HSV | 13.530 | 19.512 | 17.953 | 20.938 | 20.958 | 20.987 | 15.407 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |

Table 10: Rankings according to measurements

### 5.6.2 Results on video sequences

The results of this subsection concern the measurements conducted over the entire synthesized sequences. As in the case of still images studied in the previous section, the rankings of the objective metrics (Table 11) are consistent with each other: the correlation coefficients between objective metrics are very close from the figures depicted in Table 8, and so they are not presented here. As with still images, the difference between the subjective-test-based ranking and the ranking from the objective scores is noticeable. Again, the algorithm judged as the worst by the objective measurements, is the one preferred by the observers.

Table 12 presents the correlation coefficients between objective scores and subjective scores, based on the whole set of measured points. None of the tested objective metric

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
|---|---|---|---|---|---|---|---|
| MOS | 2.70 | 2.41 | 2.14 | 2.03 | 1.96 | 2.13 | 1.28 |
| Rank order | 1 | 2 | 3 | 5 | 6 | 4 | 7 |
| PC | 19.02 | 24.99 | 23.227 | 25.994 | 26.035 | 26.04 | 20.89 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| PSNR | 0.648 | 0.844 | 0.786 | 0.859 | 0.859 | 0.859 | 0.824 |
| Rank order | 7 | 4 | 6 | 1 | 1 | 1 | 5 |
| SSIM | 0.664 | 0.932 | 0.825 | 0.948 | 0.948 | 0.948 | 0.888 |
| Rank order | 7 | 4 | 6 | 1 | 1 | 1 | 5 |
| MSSIM | 13.14 | 20.41 | 18.75 | 21.786 | 21.965 | 21.968 | 20.73 |
| Rank order | 7 | 5 | 6 | 3 | 2 | 1 | 4 |
| VSNR | 0.129 | 0.393 | 0.313 | 0.423 | 0.423 | 0.424 | 0.396 |
| Rank order | 7 | 5 | 6 | 2 | 2 | 1 | 4 |
| VIF | 0.153 | 0.415 | 0.342 | 0.446 | 0.446 | 0.446 | 0.419 |
| Rank order | 7 | 5 | 6 | 1 | 1 | 1 | 4 |
| VIFP | 0.359 | 0.664 | 0.58 | 0.598 | 0.598 | 0.598 | 0.667 |
| Rank order | 7 | 5 | 6 | 3 | 3 | 3 | 1 |
| UQI | 0.779 | 2.399 | 1.926 | 2.562 | 2.562 | 2.564 | 2.404 |
| Rank order | 7 | 5 | 6 | 2 | 2 | 1 | 4 |
| IFC | 8.66 | 15.933 | 13.415 | 16.635 | 16.739 | 16.739 | 10.63 |
| Rank order | 7 | 4 | 5 | 3 | 1 | 1 | 6 |
| WSNR | 14.41 | 20.85 | 18.853 | 21.76 | 21.839 | 21.844 | 16.46 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| PSNR HSVM | 13.99 | 19.37 | 18.361 | 21.278 | 21.318 | 21.326 | 16.23 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| PSNR HSV | 13.74 | 19.52 | 17.958 | 20.795 | 20.823 | 20.833 | 15.91 |
| Rank order | 7 | 4 | 5 | 3 | 2 | 1 | 6 |
| VSSIM | 0.662 | 0.879 | 0.809 | 0.899 | 0.898 | 0.893 | 0.854 |
| Rank | 7 | 4 | 6 | 1 | 2 | 3 | 5 |
| VQM | 0.888 | 0.623 | 0.581 | 0.572 | 0.556 | 0.557 | 0.652 |
| Rank order | 7 | 5 | 4 | 3 | 1 | 2 | 6 |

Table 11: Rankings according to measurements

reaches 50% of the subjective scores. The metric obtaining the higher correlation coefficient is VSNR, with 47.3%. Figure 20 shows the same obtained correlation scores, with resulting ranking of tested metrics. It is easily observed that the top metrics are perceptual-like metrics (they include psychophysical ap-proaches).

To conclude, performances of objective metrics, with respect to subjective scores, are different in the case of video sequences than in the case of still images. Correlation coefficients between objective and subjective scores were higher in the case of video sequences. However, human opinion also differed in the case of video sequences. In the case of video sequences, perceptual-like metrics were the most correlated to subjective

| | PSNR | SSIM | MSSIM | VSNR | VIF | VIFP | UQI | IFC | NQM | WSNR | PSNR HVSM | PSNR HVS | VSSIM | VQM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CCMOS | 34.5 | 45.2 | 27 | 47.3 | 43.9 | 46.9 | 20.2 | 45.6 | 36.6 | 32.9 | 34.5 | 33.9 | 33 | 33.6 |

Table 12: Correlation coefficients between objective and subjective scores in percentage.

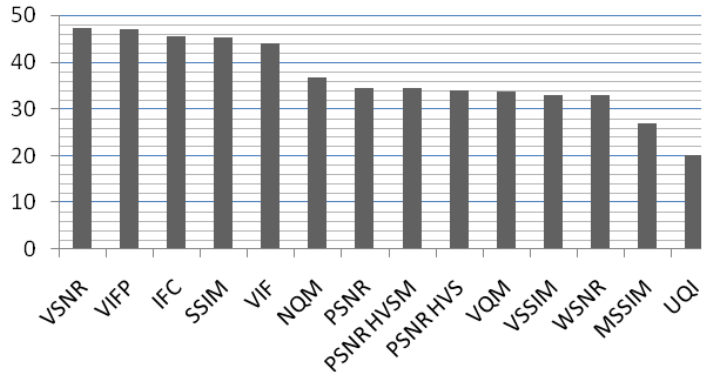



Figure 20: Ranking of used metrics according to their correlation to human judgment.

scores (also in video conditions). However, in both conditions, none of the tested metrics did not reach 50% of human judgment.

## 5.7 Future trends

### 5.7.1 Subjective methodologies of quality assessment

ACR-HR and PC are known for their efficiency in 2D conditions. However, one may need to assess the quality of 3D media in 3D conditions. Defining a new subjective video quality assessment framework is a tough task, knowing the new the complexity brought by 3D media. The difficulty of 3D images quality evaluation, compared to 2D conventional images, is now more considered. Seuntiens [44] introduced new parameters to assess in addition to image quality, which are naturalness, presence and visual experience. Thus, a multi-dimensional quality indicator may allow a reliable assessment of 3DTV media. However, it may be difficult to define such terms in the context of a subjective quality assessment protocol, and there is no standardized protocol considering these aspects yet. ITU-R BT. 1438 recommendation [26] describes subjective assessment of stereoscopic television pictures and the methods are described in [9]. Chen *et al.* [13] revisited the question of subjective video quality assessment protocols for 3DTV. This work points out the complexity of 3D media quality as-sessment. Chen *et al.* proposed to reconsider several conditions in this context, such as the viewing conditions (viewing distance, monitor resolution), the test ma-terial (depth

rendering according to the chosen 3D display), viewing duration, etc. In the following, some of the requirements proposed by Chen *et al.* in [13] are mentioned:

- General viewing conditions: First the luminance and contrast ratio is considered, because of the crosstalk involves by 3DTV screens, and because of the used glasses (as active as polarized glasses cause reduction of luminance). Second, the resolution of depth as to be defined. Third, the viewing distance recommended by ITU standards may differ according to the used 3D display. Moreover, as the authors of the study claim it, depth perception should be considered as a new parameter to evaluate the Pre-ferred Viewing Distance, such as human visual acuity or picture resolution.
- Source signals: the video format issue is mentioned. It refers to the numerous 3D representations (namely ‚ÄúLayer Depth Video‚Äù (LDV), ‚ÄúMulti-view Video-plus-Depth‚Äù (MVD), or ‚Äúvideo plus depth‚Äù (2D+Z)) whose reconstruction or conversion lead to different types of artifacts.
- Test methods: as mentioned previously, new aspect have to be considered (naturalness, presence, visual experience), and visual comfort as well. The latter refers to the visual fatigue that should be measured to help in a standardization process.
- Observers: an adapted protocol should involve the measurement of viewers stereopsis ability, first. Second, the authors mention that the required number of participants may differ from 2D. Then further experiments should define this number.
- Test duration and results analysis: the duration of the test is still to be determined, taking into account the visual comfort. The analysis of the results refers to the definition of a criterion for incoherent viewer rejection and to the analysis of the assessed parameter (depth, image quality, etc.)

### 5.7.2 Objective quality metrics

Most of the proposed metrics for assessing 3D media are in-spired from these algorithms. Previous studies ([59], [51]) already considered the reliability of usual objective metrics. However, often, experimental protocols involve depth and/or color compression, different 3D displays, and different 3D representations (2D+Z, stereoscopic video, MVD, etc...). Then human opinion is compared to decoded data objective scores. The experimental protocols often as-sess at the same time both compression distortion and synthesis distortion, without distinction. The most recent proposed 3D quality metrics propose to take into account the new modes brought by 3D. Among the proposed metrics, numerous target stereoscopic video, for instance, but not views synthesized from DIBR. Then they will not be referred to in this section.

Ekmekcioglu *et al.* [19] proposed a depth-based perceptual quality metric. It is a FR video metric that uses a weighting function based on depth data at the target viewpoint, and a temporal consistency function.

Zhao and Yu [63] proposed a FR metric, Peak Signal to Perceptible Temporal Noise Ratio. The metric evaluates quality of synthesized sequences by measuring the perceptible temporal noise within these impaired sequences.

Perceptual Quality Metric (PQM) [29] is proposed by Joveluro *et al.* Although the authors assess the quality of decoded 3D data (2D+Z), the metric is applied on left and right views synthesized with a DIBR algorithm (namely [20]). Thus, the method can be cited in this section. The quality score is a weighted function of the contrast and the luminance differences between both reference and distorted view.

Yasakethu *et al.* [60] proposed an adapted VQM for measuring 3D Video quality. It uses 2D color information and depth information. In [60], the metric is evaluated through left and right view (rendered from 2D+Z encoded data), and compared to subjective scores obtained by using an autostereoscopic display. Results show higher correlation than simple VQM.

## 5.8 Conclusion

This section proposed a reflection considering both subjective quality assessment protocols and objective quality assessment methods reliability in the context of DIBR-based media. Typical distortions related to DIBR were introduced. They are geometric distortions and mainly located around the disoccluded areas. When compression-related distortions and synthesis-related distortions are combined, the errors are generally scattered in the whole image, increasing visual annoyance. Two case studies were presented answering the two questions relating, first to the suitability of two efficient subjective protocols (in 2D), and second, to the re-liability of commonly used objective metrics. Experiments considered commonly used methods for assessing conventional images, as subjectively as objectively, to assess DIBR-based synthesized images, from seven different algorithms.

Concerning the suitability of the tested subjective protocols, the number of participants required for establishing a statistical difference between the algorithms was almost the double of the number required by VQEG (24), which reinforce Chen *et al.* requirements [13]. Both methodologies agreed on the performances ranking of the view synthesis algorithms. Experiments also showed that the observers opinion was not as stable when assessing still images as when assessing video sequences, with ACR-HR. PC gave stable results with fewer participants than ACR-HR, in the case of still images. Both methodologies have their advantages and drawbacks and they are complementary: assigning an absolute rating to distortions such as synthesized views ones seemed a tough task to observers, although it provides knowledge on the perceived quality of the set. Small impairements are better evaluated with PC.

Concerning the reliability of the tested objective metrics, the results showed that objective metrics did not correlate the observers opinion. Objective measures did not reach 50% of human judgment and they were all correlated with each other. The results suggest that tiny distortions are penalized by the objective me-trics when not perceptible by humans. Then, objective metrics inform on the exis-tence of distortions but not on their visible annoyance. Using the tested metrics is not sufficient for assessing virtual synthesized views.

The simple experiments that have been presented in this section reveal that the reliability of the tested objective metrics is uncertain when assessing intermediate synthesized views, in the tested conditions. Yet, reckoned organizations plan to base partially their future decisions, concerning new strategies for 3D video, on the outcome of such objective metrics. New standards have to been developed considering the new aspects brought by DIBR: location and type of artifacts, degree of annoyance of artifacts.

# 6 Test to assess the visual discomfort induced by stimulus movement

The aim of the study is to learn the effects of disparity and planar motion on visual discomfort. In the experiment, the paired comparison method was used. Two separate subjective experiments were conducted on experts and non-experts observers. The detailed experimental setup are introduced in the following part.

## 6.1 Experimental design

It is often accepted that 60-70 minutes of arc is the comfort threshold for static disparities. Meanwhile, some researchers also use depth of focus (DOF) to calculate the comfortable viewing zone [14], which refers to the range of distances in image space within which an image appears in sharp focus and is given in terms of diopters (D) (a value of $\pm$0.2D is suggested). To investigate how planar motion affects the visual discomfort at different disparity levels, five binocular disparity levels for the foreground were chosen in this experiment. Three of the angular disparity levels were within the comfortable viewing zone [14], two were outside it. These can be expressed in degrees of visual angle [25], as shown in Fig.21. The binocular angular disparity can be calculated by the following equations 1 and 2, $\phi_A$ and $\phi_B$ are binocular angular disparities for A and B. Note that the positive value represents the crossed disparity, such as the point A; the negative value represents the uncrossed disparity, such as the point B.

In this study, the five angular disparity levels were 0, $\pm$0.65, and $\pm$1.3 degree (+ means crossed, - means uncrossed), assuming that the interpupillary distance was 65 mm and the viewing distance was about 90 cm. A background was placed at a fixed position which was behind the screen at a distance of about 46.28cm, with the angular disparity of -1.4 degree. Fig. 22 shows the disparities used in the stimuli and their relationship with comfortable viewing zone. Three velocity levels which represented slow, medium and fast were used in the experiment.

$$\phi_A = \beta - \alpha \tag{1}$$

$$\phi_B = \gamma - \alpha \tag{2}$$

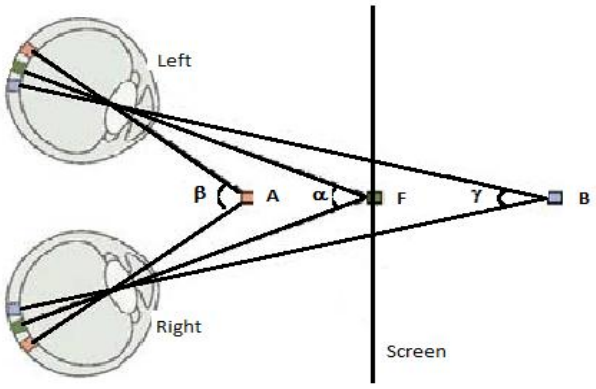



Figure 21: The definition of the binocular angular disparity, where F is the fixation point.

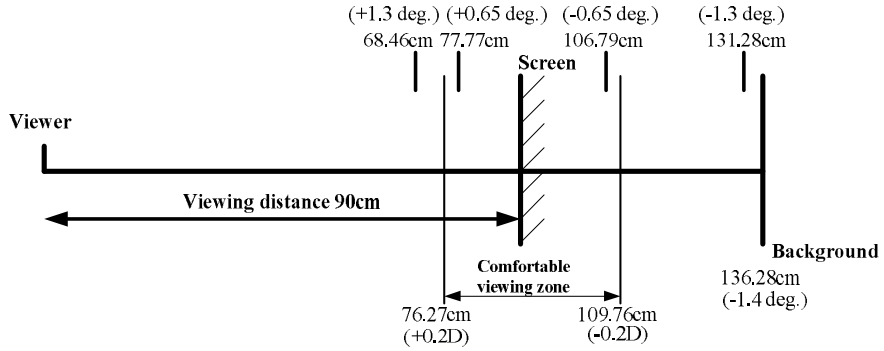



Figure 22: The relationship of the foreground and the background position and the comfortable viewing zone.

## 6.2 Stimuli

To avoid the influence of other factors on visual discomfort, we used computer-generated stereoscopic sequences for precise control. The stereoscopic sequences consisted of a left-view and a right-view image which were generated by the MATLAB psychtoolbox [8] [38]. Each image contained a foreground and a background. A black Maltese cross with 480×480 pixels was used as the foreground object as it contained both high and low spatial frequency components. This was supposed to limit the influence of one particular spatial frequency in the experiment. The Maltese cross moved along a trajectory which was a circle with center point at the center of the screen, and a radius of 300 pixels. The motion direction was anti-clockwise. The reason to choose a circle as the trajectory was that it could avoid the step impulse that came from a sudden change of the motion direction, which may cause unexplained effects of visual discomfort. As the trajectory was a circle, the velocity was expressed in degree/s. The three velocity levels were 71.8, 179.5 and 287.2 degree/s, which represented slow, medium and fast, respectively. The background was generated by adding salt&pepper noise on a black image, and then filtered by a circular averaging filter. The reason that we used this kind of image as the common background of all stimuli was that it could preclude all of the monocular cues on stereopsis. Additionally, to give the viewers a reference of the trajectory a black circle which was the same as the moving track of the object was placed on the background. So, for viewers, the stimuli appeared to be composed of two parts: the salt&pepper-like background with a black circle on it, and a moving Maltese cross on a depth plane with a certain velocity. As there were 3 levels of velocity and 5 levels of angular disparity, there would be totally 15 stimuli for the experiment. An example of the stimuli is shown in Fig.23, in which the foreground object is placed in front of the screen with an angular disparity of 1.3 degree.

## 6.3 Apparatus

The stereoscopic sequences were displayed on a Dell Alienware AW2310 23-inch 3-D LCD screen (1920×1080 full HD resolution, 120Hz), which featured 0.265-mm dot pitch. The display was adjusted for a peak luminance of 50 $cd/m^2$ when viewed with the active shutter glasses. The graphics card of the PC was an NVIDIA Quadro FX 3800. Stimuli were viewed binocularly through the NVIDIA active shutter glasses (NVIDIA 3D vision kit) at a distance of about 90 cm, which was approximately three times of the picture height. The peripheral environment luminance was adjusted to about 44 $cd/m^2$. When seen through the eye-glasses, this value corresponded to about 7.5 $cm/m^2$ and thus to 15 % of the screen's peak brightness as specified by ITU-R BT.500 [10].

## 6.4 Viewers

We conducted two separate experiments for experts-only viewers and naive observers respectively. Ten experts in 3-D perception, coding, quality assessment and subjective experiments participated in the experiment. Eight experts are male, two are female.

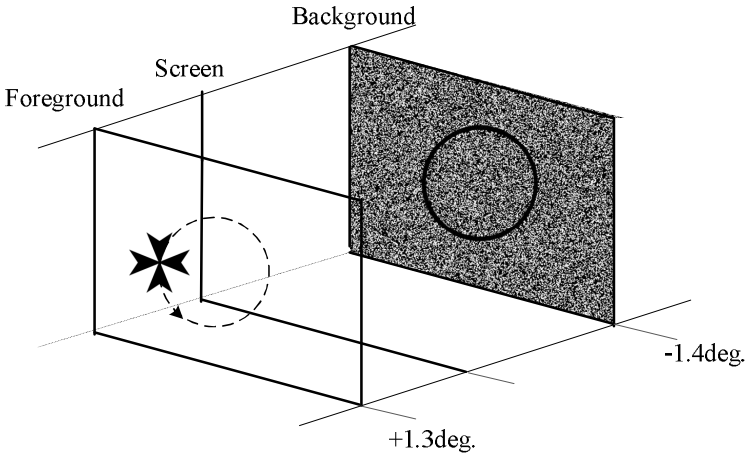



Figure 23: An example of a stereoscopic image in the experiment. The foreground object is moving at the depth plane with a disparity of 1.3 degree. The background is placed at the depth plane with a disparity of -1.4 degree. The motion direction of the Maltese cross is anti-clockwise.

Their ages ranged from 24 to 43 years old. Forty-five naive viewers who are not directly concerned with 3-D related research as part of their normal work, and are not experienced assessors participated in the second experiment. Twenty-one are male, twenty-four are female. Their ages ranged from 18 to 44 years old with average age 24.

All of the viewers have either normal or corrected-to-normal visual acuity. The visual acuity test was conducted with a Snellen Chart for both far and near vision. The Randot Stereo Test was applied for stereo vision acuity check, and Ishihara plates were used for color vision test. All of the viewers passed the pre-experiment vision check.

## 6.5 Assessment method

In our study, the paired-comparison method was chosen as it is a well-known method in the field of psychophysics [21] [42]. In the experiment, the viewers watched a pair of stimuli at one trial, and then they were asked to select the one which made them feel more uncomfortable. For experts-only test, a total of 210 pairs were presented in each individual subjective experiment. This number corresponds to the presentation of all combinations of 15 stimuli except for equal presentation on the left and right side. In particular, it contains the same condition with the stimulus order inversed as the first onset of a stimulus might have caused a bias on the feeling of visual dis-

comfort. The presentation order of the whole 210 pairs was randomly permuted for each viewer. For naive observers, A total of $\begin{pmatrix} 15 \\ 2 \end{pmatrix}$ =105 pairs were presented in each individual subjective experiment. The presentation order of stimuli in one paired comparison was different for odd numbered and even numbered observers. For example, observers with even numbers will watch stimulus A first, then stimulus B. For odd numbered observers, this order is inversed. This is used to balance the presentation order. The presentation order for voting the whole 105 paired comparisons was randomly permuted for each viewer.

## 6.6 Procedure

The subjective experiment contained a training session and a test session. In the training session, there were five pairs of stimuli. At the beginning, the viewers were told that they will watch a series of stereoscopic motion images. They were asked not to stare at the moving object all the time, but watch the whole screen of the stereoscopic sequence under test. Then, they should select the one which made them feel more uncomfortable, concerning e.g., eye strain, headache,etc. As it was not technically possible to display a pair of stimuli on two screens, the viewers had to use two keys to switch between the pair of stimuli on one screen. There was a minimum time limit of 5 seconds for the display of stimuli, which means each observer had to watch each of the stimuli at least 5 seconds before making their decision by pressing a specified button. After the explanation of the experiment, the viewers were asked to do the test by themselves. During the training session, all questions of the viewers were answered. We ensured that after the training session, all of the viewers knew about the process and task of this experiment clearly.

In the main test session, the task and procedure were the same as the training session except 210 pairs of stimuli for experts-only test and 105 pairs for naive observers were compared. As the duration of the whole test was different due to the individual difference of each viewer, and to avoid visual fatigue caused by long time watching affecting the experimental results, the viewers were asked to have a 10 minutes break after half of the test samples.

It should be noted that in the experts-only test, a total of 15 data sets were acquired as 5 experts participated twice in the experiment but on a different day.

# 7 A time-dependent visual attention model in stereoscopic condition, combining center and depth bias

## 7.1 Introduction

Recent studies [49], [64] have shown the importance and the influence of the “external biases” in the deployment of the pre-attentive visual attention. In itself, the degree

to which visual attention is driven by stimulus dependant properties or task-and-observer dependant factors is an open debate [37],[48],[52],[15]. But considering their interactions and impacts over time might be crucial to improve the predictability of existing saliency models [49], [48].

## 7.2 Statistical analysis

Following the temporal behavioral study, we proposed to rely on the center and depth biases as potentially guiding factors to existing visual attention models. In order to quantitatively evaluate the contribution of these factors, we followed a similar approach to Vincent's et al. one [55]. A statistical model of the fixation density function f(x,t) is expressed in term of an additive mixture of different features or modes, each associated to a given probability or weight. Then, each mode consists of an a priori guiding factor over all scenes. The density function is defined over all spatial fixation positions represented by the bi-dimensional variable x so that:

$$f(x,t) = \sum_{k=1}^{K} p_k(t)\phi_k(x) \tag{3}$$

where $K$ is the number of features, $\phi_k(x)$ the probability density function for each feature $k$ and $p_k(t)$ the contribution or weight of feature $k$ with the constraint that $\sum_{k=1}^{K} p_k = 1$ for a given time $t$. The statistical analysis aims at separating the contribution of the bottom-up saliency feature (itself based on low-level features) from additional features observed in the previous sections. To perform this analysis, each fixation is used separately to characterize the temporal evolution of contribution weights $p_k(t)$. An "Expectation-Maximization" (EM) method estimates the weights in order to maximize the global likelihood of the parametric model [17]. Before explaining this method, we describe the center and depth modeling.

## 7.3 Model of the center bias

The strongest bias underlined by laboratory experiments is the central bias. This bias is likely an integral feature of visual perception experiments accounting for an important proportion of human eye guidance, as proposed by [6]. However, the extent to which this potential laboratory artifact is an inherent feature of strategy of human vision remains an open subject. Tatler [49] studied the central bias over time and observer's task. He gave evidence that the central fixation tendency persists throughout the viewing in free viewing condition, while rapidly dissipated in a search task. Indeed from the third fixation, the central bias is hardly noticeable. In our case of depth-layer detection task, the observers were asked to press a button as soon as they distinguished at least two depth layers in the image. Whatever the images, observations show a strong central fixation tendency on the earliest fixations followed by a sparser fixation distribution. As in the case of search task in [49], there is little evidence for a central fixation bias from the third fixation. Considering the results of

the literature and our observations, the central bias is modeled by a single 2D Gaussian. The use of a single Gaussian filter is empirically justified by the convergence property of the fixation distribution [62]. As proposed in [24], the parameters of the Gaussian function are predefined and are not estimated during the learning. On the present dataset, this choice is justified by the strong central fixation distribution on the first fixation that goes into fast spreading and then tends to converge to a fix size. A fixation-dependent estimation of the parameters would have fit the whole spread fixation distributions. The central bias is then modeled by a time-independent bidimensionnal Gaussian function, centred at the screen center as $N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ the covariance matrix and with $\sigma_x^2$ and $\sigma_y^2$ the variance. We fit the bidimensional Gaussian to the fixation distribution on the first fixation only. Whatever the viewing conditions (2D or 3D), the fixation distributions are similarly centered and Gaussian distributed ($\sigma_{x2D} = 4.7°, \sigma_{y2D} = 2.5°, \sigma_{x3D} = 4.3°, \sigma_{y3D} = 2.3°$)

## 7.4 Model of the depth bias

Results presented in section "Existence of a depth bias on natural images" show that the perceived mean depth depends on the viewing conditions. At the beginning of viewing (early stage), the mean depth is significantly lower in 3D condition than in 2D condition. Observers show a tendency to fixate more the closest locations at the beginning of visualization than the farthest ones. How the depth cues interact to modulate the visual attention is an open issue. In particular, the figure/ground organization [43], that can be understood as an element of the edge interpretation depth cue [36], drives the visual attention pre-attentively [41]. This supports our choice of figure-ground organization implementation by a classification of depth maps in individual foreground and background maps. These maps have been thresholded at half the depth value through a sigmoid function, such that pixels values smaller and higher than 128 rapidly cancel out on background and foreground respectively. Background values are inversed such that the farther a point is in the background, the more it contributes to the background feature. At the opposite end, the closer a pixel is to the foreground, the more it contributes to foreground feature. Two resulting foreground and background map are illustrated on Figure 24 (a).

## 7.5 Proposed model

The proposed model aims at predicting where we look at in 2D and 3D conditions. The prediction is based on a linear combination of low-level visual features, center and depth biases. However, other contributions much more complex than those mentioned above likely occur over time. For instance, top-down process could interact with them, especially in the late time of fixation. To deal with this issue, an additional feature map whose fixation occurs at all locations with same probability is then used to model the influence of other factors such as prior knowledge, prior experience, etc. Obviously the contribution of the uniform map has to be as low as possible meaning that other features (low-level saliency map, center and depth biases) are the most important to predict where we look at. In summary five feature maps are used as illustrated in Figure 24(a):

- A first one is obtained by using one of the state-of-the-art bottom-up models (Itti, Bruce and Le Meur). This represents the "low-level saliency map";
- one for the central fixation bias;
- two related to the depth cue, i.e. the foreground and background maps;
- a uniform distribution map

Low-level saliency and the foreground and background features are dependent on the visual content. The center and uniform map represent higher-level cues. They are fixed over time and identical for all stimuli. The additive mixture model is then given by:

$$f(x,t) = p_{sm}(t)\phi_{sm}(x) + p_{cb}(t)\phi_{cb}(x) + p_{fg}(t)\phi_{fg}(x) + p_{bg}(t)\phi_{bg}(x) + p_{un}(t)\phi_{un}(x) \quad (4)$$

with $\phi_{sm}$ the saliency maps of one of the 3 models, $\phi_{cb}$ the central Gaussian function, $\phi_{fg}$ and $\phi_{bg}$ the foreground and background map respectively and $\phi_{un}$ the uniform density function. Each feature is homogeneous to a probability density function. $\phi_{sm}, p_{cb}, p_{fg}, p_{bg}$ and $p_{un}$ are the time-dependent weights to be estimated, their sum being equal to unity. Figures 24(a) and (b) give an illustration of the involved features. The following pseudo-code describes the EM algorithm. The weights $p_k^{(m)}(t)$ are the only parameters estimated for each iteration $m$. In practice, a fix number $M$ of 50 iterations is a good tradeoff between estimation quality and complexity.

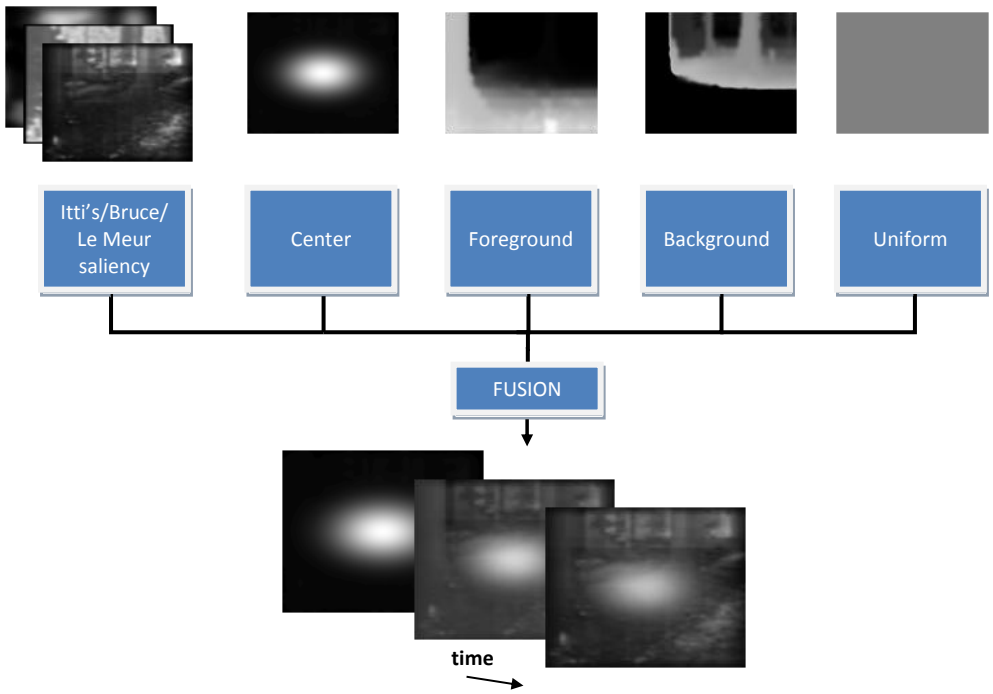



Figure 24: (a) Upper Row: Illustration of Itti's saliency map obtained from an image, the center bias in 2D condition,the corresponding foreground and background feature maps. (b) Middle row: Description of the proposed time-dependent model. (c) Lower Row: Illustration of the resulting time-dependent saliency map for the first, 10th and 20th fixation in 2D condition (when Itti's model is used to predict the bottom-up saliency map).

With $t_k = \{sm, cb, fg, bg, un\}$
initialization of the weights $p_k^{(0)}(t) = 1/K \quad \forall k$;
**for** *each fixation rank from 1 to 25* **do**
  **for** *each each iteration* $m = 1..M$ **do**
    **for** *each each feature* $k = 1..K$ **do**
      **for** *each each fixation* $i = 1..N$ **do**
        Expectation step: Given a current estimate of the parameters $p_k(t)$, an estimation of the missing probabilities $t_k$ is computed:
        $t_{i,k}^{(m)} = P\{x_i \text{ comes from the feature } k\}$
        $t_{i,k}^{(m)} = \frac{p_k^{(m-1)} \phi_k(x_i)}{\sum_{l=1}^{K} p_l^{(m-1)} \phi_l(x_i)}$
      Maximization step: The parameters $p_k^{(m)}(t)$ are updated for the iteration $m$:
      $p_k^{(m)}(t) = \frac{\sum_{i=1}^{N} t_{i,k}^{(m)}}{N}$

## 7.6 Results of the statistical analysis

The temporal contributions of the proposed features to visual attention are evaluated. The EM-based mixture model was run on half of the image dataset at each fixation rank (from the first to $25^{th}$ fixation): each fixation per observer is projected on all the feature maps associated with a given stimulus image. There are 14 participants and consequently at most 14 fixations per fixation rank per image. The EM algorithm gives at convergence an estimation of the mixture weights maximizing the linear additive combination of different features with respect to the original human fixation distribution. The process is repeated at each fixation rank, and with fixations in 2D and 3D conditions. The temporal contributions of all the visual guiding factors are illustrated on Figure **??**:

The best predictor for both viewing conditions is the predicted saliency map (from Itti's model and called Sm on Figure **??**). As expected, the central fixation bias shows a strong contribution on the two first fixations but rapidly drops to an intermediate level between saliency ($Sm$) and other contributions. The contribution of the center bias ($Cb$) is significantly (paired t-test, $p<0.001$) more important in 3D condition than 2D condition, while the foreground (Fg) is significantly (paired t-test, $p<0.001$) more important in 3D condition than in 2D. Indeed the center bias is partially compensated first by the high foreground contribution from the $3^{rd}$ to the $18^{th}$ fixation, second by the progressive saliency increase. Finally, the background and uniform contributions remain steadily low in the 2D case, but increase progressively in the late period in 3D condition.

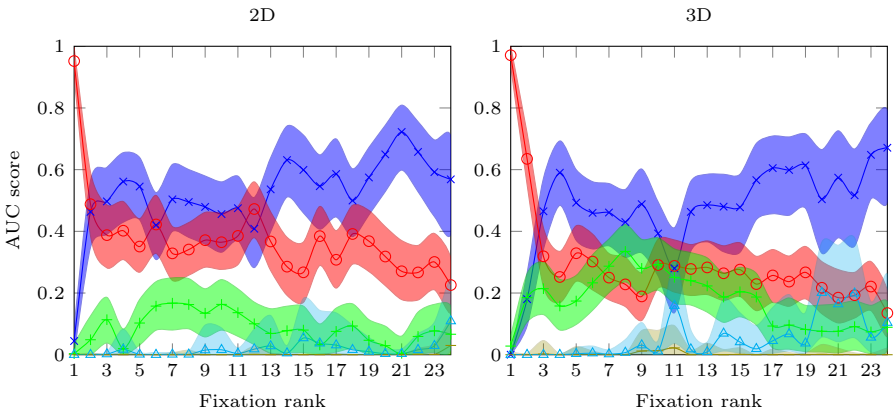



Figure 25: Temporal contributions (weights) of 5 features on 2D (left) and 3D (right) fixations to eye movements as a function of the fixation rank. Low-level saliency feature ("Sm") here comes from Itti's model. The error areas at 95% are computed by a "bootstrap" estimate (1000 replications).

## 7.7 Discussion

The temporal analysis gives a clear indication of what might guide the visual exploration on a fixation per fixation basis. We have considered different plausible features linearly combined with time-dependent weights. The temporal evolution of central bias, foreground and low-level saliency is highlighted. According to our observation, the central bias is strong and paramount on first fixation, and decreases to a stable level from the third fixation. As shown by Tatler's experiments [49] and in accordance with [24], the central fixation point at the beginning of visualization is very probably not due to the central fixation marker before stimuli onset, but to a systematic tendency to recenter the eye to the screen center. Indeed, it is shown that this tendency exists even with a marker positioned randomly within a circle of 10° radius from screen center [49]. Also, in these central bias observations and Tatler's findings (in search task), center bias was not evident from the third fixation. In our context, the contribution of center feature from third fixation is effectively lower but not negligible. The binocular disparity introduction promotes the foreground feature up to the $17^{th}$ fixations. Results suggest that foreground helps to predict salient areas in 2D condition but its contribution is much more important in stereo condition. This is coherent with our previous conclusions (cf. section 4.2.3). It is known that different depth cues interact to drive the visual attention preattentively. Among the depth cues, some are monoscopic and other stereoscopic like the binocular disparity. Our results show that a depth-related feature like the foreground contributes to predict salient areas in monoscopic conditions, because depth can be inferred from many monoscopic depth cues (like accommodation, motion parallax, familiar size, edge interpretation, shading etc.). But our results also show that the binocular disparity greatly increases

the contribution of foreground to visual attention deployment and indeed might participate to the figure/ground organization. At the opposite, the background feature does not contribute to visual attention deployment, or when it does (from the 23 and $19^{th}$ fixation in 2D and 3D conditions respectively), it is combined with a contribution of uniform distribution. We could expect that observers tend to direct their gaze globally to background plane after viewing the foreground area at the very beginning of viewing. This is not the case: fixations can occur in the background, but observers don't show a common tendency of looking at the background from a certain fixation rank. Finally, the contribution of the uniform distribution term remains low up to the "late" time of visualization. It models the influence of other high - level factors possibly due to top-down mechanisms that are not accounted by our proposed factors. Results show these factors contribute few to temporal saliency construction on the 20 first fixations. Afterwards, the uniform distribution contribution increases over time suggesting that the existing features are not sufficient to explain the eye movements. The temporal analysis is also reiterated with the low-level saliency map of Bruce and Le Meur models. Results are very similar. In the following section, we use the learnt time-dependent weights to predict where observers look at. Performance of the time-dependent saliency models is evaluated on the remaining half image dataset. The performance analysis is carried out from the first to the $19^{th}$ fixations, a time slot for which the contribution of uniform distribution is stable and low in all conditions.

## 7.8 Time-dependent saliency model

In the previous section, we have learnt through an EM algorithm the linear combination of five visual guiding factors matching the ground-truth visual saliency. The following step consists in using these weights to compute a saliency map taking into account the low-level visual features, the depth and the center bias. The same additive pooling of equation (4) is used. For each fixation, the learned weights vary, leading to a time-dependent adapted saliency map. The time-dependent saliency model is then compared to corresponding original saliency model in 2D and 3D conditions. Three methods are evaluated performed in both 2D and 3D conditions:

- The original saliency model: the saliency map is the output of state-of-the-art models.
- The equally weighted model: the final saliency map is the average of the five feature maps. The weights $p_k(t)$ are not time-dependent and are set to $1/K$, where $K$ is equal to 5 in our study.
- The time-dependent saliency model: the time-dependent saliency map is the linear combination (cf. formula (4)) using the learned and time-dependent weights $p_k(t)$.

In the second and third case, each feature is at first normalized as discrete probability density functions, (so that the sum of the whole values is equal to one) before all features are weighted and summed. Thereafter, we used two comparison metrics to

assess the performance of saliency models, i.e. their quality of fixation prediction. Again, the ROC analysis is used. However, two saliency maps were compared in section 4.2.1. Here, to assess the performance for each fixation rank, the analysis is performed between a distribution of human fixations and a predicted saliency map. Then for each couple *"image x fixation"* (with each participant's fixation for a given fixation rank), an AUC value is obtained. Results are then averaged over all test pool images for a given fixation rank.

The AUC values of original Itti's model fixation per fixation are plotted in Figure 26 and compared to the performances of the time-dependent model. For reference, the AUC value between Itti's model and the first 19 cumulated fixations, as it is usually computed, is also plotted (light blue horizontal line). Results show a constant gain of performance over time and emphasize the importance of time in the computational modeling of visual attention.

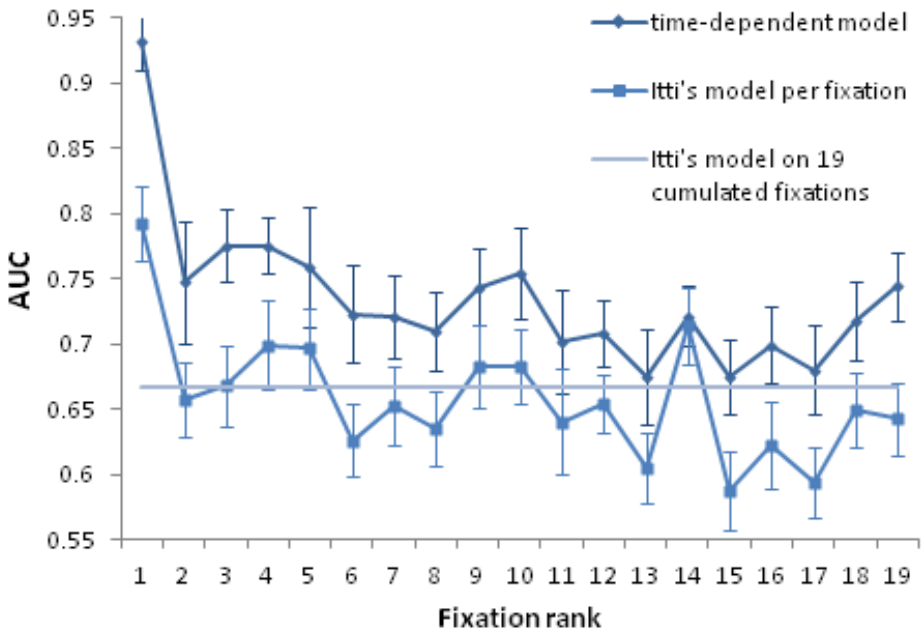



Figure 26: Temporal evolution of the performance of the time-dependent model based on Itti's, versus the Itti's model per fixation, and versus the Itti's model on 19 cumulated fixations.

The "Normalized Scanpath Saliency" (NSS) is also used to assess the performance of the normalized predicted saliency maps at the fixation positions. A NSS value is given for each couple "image x fixation/participant". Results are also averaged over all participants and all images for each fixation rank. Finally, the Figure 27 illustrates the NSS and AUC performance for the 3 state-of-the-art and the proposed models, in 2D and 3D conditions, averaged over time.

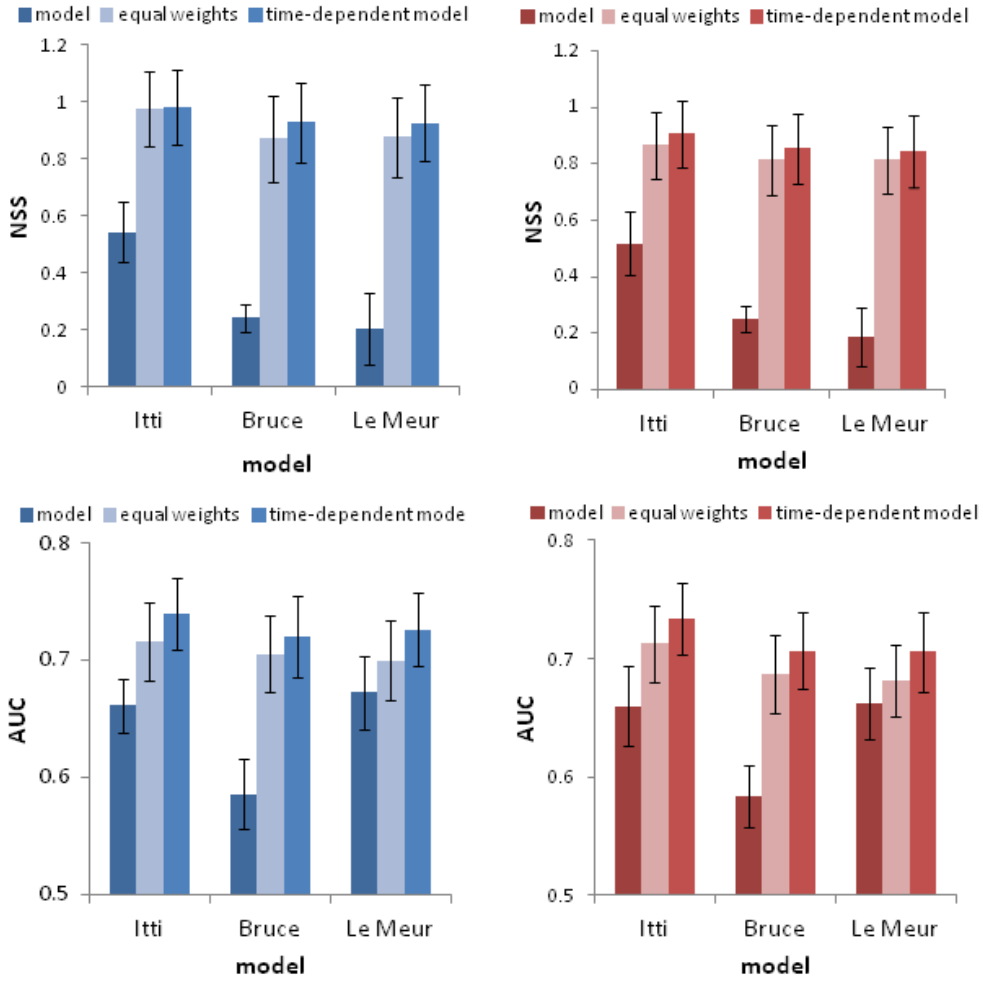



Figure 27: Comparison between 6 saliency models in 2D (left) and 3D conditions (right). Upper row: the NSS criterion, lower row the AUC criterion. The error bars correspond to the SEM. NS corresponds to Non-Significant. When the term NS is not indicated, results are significantly different ($p<0.05$).

First we note that results are all much higher than the chance level (0 for NSS and 0.5 for AUC). Not surprisingly, models including the 4 visual features low-level saliency, center bias, foreground and background (plus the uniform feature) significantly outperform existing models for both metrics. The differences are all statistically significant (paired t-test, $p<0.05$) for both criterion in both conditions and for all saliency models (except in two cases marked "NS" on the Figure 27). Itti's based time-dependent model ranks first, with a NSS score of 0.98 in 2D and 0.91 in 3D condition, and an AUC of 0.74 in 2D and 0.73 in 3D conditions. The final proposed method has greatly improved but also balances the performance between models, for NSS and AUC values. While the model using uniform weights without time adaptation leads to significant improvement, the time-dependent weighting increases even more the performance.

### 7.8.1 Discussion

The proposed approach based on time-dependent weighting improves the performance of existing attention models based on low-level visual features. The experimental dataset contained a reduced number (24) of images with different attributes of orientations, depth and contrast. The learning of the weights by EM algorithm was performed on half of this dataset, and the test of models on the remaining half images. By integrating different external and higher level feature contributions to three different existing models, the relevance of the saliency map has been increased in all viewing conditions and over time. There are however two limitations.

First of all, luminance only stimuli have been used for experiments. Even if colour might be a weak contributor to attention deployment relatively to luminance, it is however known that saliency models including color features improved their predictability [28]. From these statements and because low-level saliency models were run without color component, we can argue the contributions of low-level saliency features could be more important [31].

A second limitation is due to the content of the image itself. Natural scenes of forest were only presented to participants. Thus the depth perception, and foreground contribution in particular, might be influenced by the content of the scene itself, as well as by its geometry. A scene containing a single close object might induce a stronger foreground contribution on the early and middle period. However these remarks don't involve a reconsideration of our framework. Even if the importance of low-level saliency and foreground features might be modulated, the consideration of a pooling of low-level saliency with foreground and central feature is plausible and proved to be efficient on this dataset of images. Importantly, the foreground feature might contribute significantly more to visual deployment when binocular disparity was presented to observers. Indeed binocular disparity constitutes an additional binocular depth cue to existing monocular ones to infer the depth from 2D retinal images. In the presence of this cue, not only do observers look closer in the first fixation instants. The findings also show that the foreground itself constitutes a good predictor and a plausible visual feature that participate to a second stage of the bottom-up visual attention.

## 7.9 Conclusion

Following the observations on external center and depth biases on natural image in section 4.1, some corresponding features are proposed. Low-level saliency, center, foreground and background visual guiding factors are integrated into a time-dependent statistical parametric model. These parameters are learnt from an experimental eye fixation dataset. The temporal evolution of these features underlines some successive contributions of center, then foreground feature with a constant implication of low-level visual saliency (from the third fixation). The strong contribution of foreground feature, reinforced in the presence of "natural" binocular disparity, makes the foreground a reliable saliency predictor in the early and middle time. Then, foreground integration constitutes a simple but biologically plausible way to incorporate a complex mechanism of figure/ground discrimination for figure selection as processed in V2 area [41]. Systematic recentring tendency and following foreground selection are dedicated processes that might play an active role in the first instants of the human visual attention construction. Finally, an adapted time-dependent saliency model based on an additive mixture and the pooling of 5 features is proposed. This model significantly outperforms three state-of-the-art models. Nevertheless, the additive pooling in itself in the integration of high level visual features is a strong hypothesis. As mentioned by [24] in the case of low-level feature combination, this hypothesis is very simple with regards to the complexity of visual attention construction[54], and with regards to other computational proposals of fusion [11]. However, it constitutes an attempt of integrating V1 low-level feature with external and higher-level features that are known to occur later along the ventral pathway. Importantly, this adaptive methodology is applied at a stage where bottom-up and top-down factors are known to interact. Final results highlight the importance of a temporal consideration of individual visual features, which are known to be process specifically over time in the visual system. Integrating different features independently over time into a time-dependent saliency model is a coherent but also plausible way to model the visual attention.

# References

[1] MetriX MuX home page. http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[2] Video quality research. http://www.its.bldrdoc.gov/vqm/.

[3] VQEG 3DTV Group. http://www.its.bldrdoc.gov/vqeg/projects/3dtv/.

[4] International Telecommunication Union (ITU) Radiocommunication Assembly. Subjective video quality assessment methods for multimedia applications, 2008.

[5] M. Barkowsky. *Subjective and Objective Video Quality Measurement in Low-Bitrate Multimedia Scenarios.* Citeseer, 2009.

[6] M. Bindemann. Scene and screen center bias early eye movements in scene viewing. *Vision research*, 2010.

[7] Emilie Bosc, Muriel Pressigout, and Luce Morin. Focus on visual rendering quality through content-based depth map coding. In *Proceedings of Picture Coding Symposium (PCS)*, Nagoya, Japan, 2010.

[8] D.H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

[9] ITU-R BT. *500, Methodology for the subjective assessment of the quality of television pictures*. November, 1993.

[10] ITU-R Recommendation BT.500. Methodology for the subjective assessment of the quality of television pictures. 1974-2004.

[11] C. Chamaret, J. C. Chevet, and O. Le Meur. Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1077–1080, 2010.

[12] D. M Chandler and S. S Hemami. VSNR: a wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284–2298, 2007.

[13] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. 2010.

[14] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3dtv. *Fifth International Workshop on Video Processing and Quality Metrics (VPQM)*, 2010.

[15] V. Cutsuridis. A cognitive model of saliency, attention, and picture scanning. *Cognitive Computation*, 1(4):292–299, 2009.

[16] N. Damera-Venkata, T. D Kite, W. S Geisler, B. L Evans, and A. C Bovik. Image quality assessment based on a degradation model. *Image Processing, IEEE Transactions on*, 9(4):636–650, 2002.

[17] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[18] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli. New full-reference quality metrics based on HVS. In *CD-ROM Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, Scottsdale, USA, 2006.

[19] E. Ekmekcioglu, S. T. Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz. Depth based perceptual quality assessment for synthesized camera viewpoints. In *Proc. of Second International Conference on User Centric Media, UCMedia 2010*, Palma de Mallorca, September 2010.

[20] C. Fehn. Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, page 93,Äì104, 2004.

[21] G.A. Gescheider. *Psychophysics: the fundamentals*, chapter Chapter 9. Lawrence Erlbaum Associates, Inc., 1997.

[22] J. C Handley. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In *ISand TS PICS Conference*, pages 108–112, 2001.

[23] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. De Caluwe, S. Kohler, R. H. Koenen, and S. Rihs. PVQM-A perceptual video quality measure. *Signal processing: Image communication*, 17(10):781,Äì798, 2002.

[24] T. Ho-Phuoc, N. Guyader, and A. Guerin-Dugue. A functional and statistical Bottom-Up saliency model to reveal the relative contributions of Low-Level visual guiding factors. *Cognitive Computation*, 2(4):344–359, 2010.

[25] N Holliman. 3d display systems. *Science*, 38(8):31–36, 2010.

[26] ITU. Subjective assessment of stereoscopic television pictures. In *Recommandation ITU-R BT. 1438*. 2000.

[27] L. Jansen, S. Onat, and P. Konig. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1), 2009.

[28] T. Jost, N. Ouerhani, R. Wartburg, R. Muri, and H. Hugli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005.

[29] P. Joveluro, H. Malekmohamadi, W. A. Fernando, and A. M. Kondoz. Perceptual video quality metric for 3D video quality assessment. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, page 1,Äì4, 2010.

[30] M. Koppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand. Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering. In *Proc. IEEE International Conference on Image Processing (ICIP)*, Hong Kong, China, September 2010.

[31] O. Le Meur and J. C Chevet. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. *Image Processing, IEEE Transactions on*, 19(11):2801–2813, 2010.

[32] R. Martin, J. Steger, K. Lingemann, A. Nachter, J. Hertzberg, and P. Konig. Assessing stereo matching algorithms using ground-truth disparity maps of natural scenes. In *Proceedings of the 7th Meeting of the German Neuroscience Society/31th Gottingen Neurobiology Conference, Neuroforum 2007*, 2007.

[33] M Meesters, W Ijsselsteijn, and P Seuntiens. A survey of perceptual evaluations and requirements of three dimensional TV. *IEEE Transactions on Circuits And Systems for Video Technology*, 14(3):381–391, March 2004.

[34] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View synthesis for advanced 3D video systems. *EURASIP Journal on Image and Video Processing*, 2008.

[35] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand. Depth image based rendering with advanced texture synthesis. In *Proc. IEEE International Conference on Multimedia & Expo (ICME)*, Singapore, July 2010.

[36] S. Palmer. *Vision: From photons to phenomenology.* Cambridge, MA: MIT Press, 2000.

[37] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[38] D.G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10(4):437–442, 1997.

[39] M. H Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312,Äì322, 2004.

[40] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin. On between-coefficient contrast masking of DCT basis functions. In *CD-ROM Proc. of the Third International Workshop on Video Processing and Quality Metrics*, volume 4, 2007.

[41] F. T Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature neuroscience*, 10(11):1492–1499, 2007.

[42] R. Rajae-Joordens and J. Engel. Paired comparisons in visual perception studies using small sample sizes. *Displays*, 26(1):1–7, 2005.

[43] E. Rubin. *Visuell wahrgenommene figuren: Studien in psychologischer analyse.* Gyldendalske boghandel, 1921.

[44] Pieter Seuntiens. *Visual Experience of 3D TV.* Doctoral thesis, Eindhoven University of Technology, 2006.

[45] H. R Sheikh and A. C Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430,Äì444, 2006.

[46] H. R Sheikh, A. C Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *Image Processing, IEEE Transactions on*, 14(12):2117–2128, 2005.

[47] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori. Reference softwares for depth estimation and view synthesis. April 2008.

[48] B. W Tatler, R. J Baddeley, and I. D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005.

[49] B.W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[50] A. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics, GPU, and Game Tools*, 9(1):23–34, 2004.

[51] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Mueller. Quality assessment of 3D video in rate allocation experiments. In *IEEE Int. Symposium on Consumer Electronics (14-16 April, Algarve, Portugal)*, 2008.

[52] G. Underwood. Cognitive processes in eye guidance: algorithms for attention in image processing. *Cognitive Computation*, 1(1):64–76, 2009.

[53] C. Van, E. Lambrecht, and O. Verscheure. Perceptual quality measure using a Spatio-Temporal model of the human visual system. 1996.

[54] R. VanRullen. Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris*, 97(2-3):365–377, 2003.

[55] B. T Vincent, R. Baddeley, A. Correani, T. Troscianko, and U. Leonards. Do we look at lights? using mixture modelling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition*, 17(6):856–879, 2009.

[56] Z. Wang and A. C Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, 2002.

[57] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[58] Zhou Wang, Ligang Lu, and Alan Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, February 2004.

[59] S. L. P. Yasakethu, C. Hewage, W. Fernando, and A. Kondoz. Quality analysis for 3D video using 2D video quality models. *Consumer Electronics, IEEE Transactions on*, 54(4):1969–1976, 2008.

[60] S. L. P. Yasakethu, S. T Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz. A compound depth and image quality metric for measuring the effects of packet loss on 3D video. In *Proc. of 17th International Conference on Digital Signal Processing*, Corfu, Greece, July 2011.

[61] M. Yuen and H. R. Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70(3):247,Äì278, 1998.

[62] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3), 2011.

[63] Y. Zhao and L. Yu. A perceptual metric for evaluating quality of synthesized sequences in 3DV system. In *Proceedings of SPIE*, volume 7744, page 77440X, 2010.

[64] L. Zhaoping, N. Guyader, and A. Lewis. Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of vision*, 9(11), 2009.