A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions

Josselin Gautier, Olivier Le Meur

A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions

J. Gautier (*), O. Le Meur(*)

(*) University of Rennes 1, Campus Universitaire de Beaulieu, 35042 Rennes, France

Abstract

The role of the binocular disparity in the deployment of visual attention is examined in this paper. To address this point, we compared eye tracking data recorded while observers viewed natural images in 2D and 3D conditions. The influence of disparity on saliency, center and depth biases is first studied. Results show that visual exploration is affected by the introduction of binocular disparity. In particular, participants tend to look first at closer areas in 3D condition, and then direct their gaze to more widespread locations. Beside this behavioral analysis, we assess the extent to which state-of-the-art models of bottom-up visual attention predict where observers looked at in both viewing conditions. To improve their ability to predict salient regions, low-level features as well as higher level foreground/background cues are examined. Results indicate that, consecutively to initial centering response, the foreground feature plays an active role in the early but also middle instants of attention deployments. Importantly, this influence is more pronounced in stereoscopic conditions. It supports the notion of a quasi-instantaneous bottom-up saliency modulated by higher figure/ground processing. Beyond depth information itself, the foreground cue might constitute an early process of “selection for action”. Finally, we propose a time-dependent computational model to predict saliency on still pictures. The proposed approach combines low-level visual features, center and depth biases. Its performance outperforms state-of-the-art models of bottom-up attention.

Keywords: eye movements, saliency model, binocular disparity, stereoscopy

1 Introduction

The human system processes the surrounding environment information through dedicated sensory organs. However, as we are not able to perceive our visual environment at a glance due to the overwhelming information from the visual world, exploratory eye movements are used to direct the fovea -responsible for detailed vision- to particular locations of the scene. This process is called overt attention. The deployment of visual attention involves two main mechanisms: a stimulus-dependent mechanism and an observer-dependent mechanism. The former, also called bottom-up attention is driven by low-level features (1), (2) while the latter, also called top-down attention, integrates high-level cognitive processes (task, prior knowledge) (3), (4).

Different computational models of overt attention have been proposed to predict where we deploy our attention on a given stimulus, through a topographic representation of visual attention: a saliency map. Most of them are based on bottom-up mechanisms and compute a predicted saliency map indicating the conspicuous parts of the incoming stimulus (1), (2).

As the depth processing is known to follow the bottom-up processing in the ventral pathway, it is interesting to assess whether the depth cue is able to improve the saliency model's performance. Indeed, for the human vision system, the problem of recovering the distance to objects and surface in a scene is ambiguous as it is inherently a light projection from a 3D-world onto a 2D retina which can be inverted in infinite number of ways. To solve this ambiguity, visual system relies on a combination of different depth cues: the monocular depth cues available from one eye, like accommodation, motion parallax, perspective, shading, etc, as well as binocular cues like convergence and binocular disparity. While most of the first cues give relative depth information on how far objects are relative to each other, the binocular cues give absolute information about the actual distance to objects. Contrary to convergence which gives distance signal with low depth resolution at short distances (up to 2 meters), the binocular disparity is useful at short and medium distances with a high discrimination of the depth thresholds (5). Depth perception thus involves a combination of multiple but possibly conflicting depth cues to estimate the 3D structure of visual scenes.

There have been different suggestions to consider either the depth or the stereo disparity as individual features of a computational model of visual attention. Maki et al. (6),(7) first proposed a computational model based on image flow, depth and motion detection. The depth is used to prioritize the targets so that the closer the objects are, the higher priorities they are given. The main limitation comes from this assumption, as the closest object is not necessarily the most salient. Ouerhani et al. (8) also included the raw depth and some depth related features into Itti's model (2). Depth was integrated as an additional feature and transformed into a conspicuity map based on center-surround mechanisms. The principle and consistency of depth integration were qualitatively illustrated. More recently Zhang et al. (9) proposed to handle the stereoscopic visual attention. The raw depth is combined with motion and static saliency map (from Itti's model). The fusion of these three attributes with arbitrary weights is then performed. It is unfortunate that neither comparison of model's performances with human observers nor stereoscopic perception consideration were given. Actually, one of the rare attempts to account for the stereoscopic perception is the stereo visual attention framework from Bruce and Tsotsos (10). The selective tuning model of Tsotsos has been extended to address the binocular rivalry occurring in stereo vision. Unfortunately the model's performance was not given.

Another factor that significantly influences our visual deployment is the central bias. It is often assumed that this effect results from motor biases in the saccadic system or from the central distribution of image features. However, Tatler (11) showed that the central fixation bias is irrespective of observer's task or image features distribution. Be that as it may, the inclusion of the center bias in existing saliency models significantly increases the performances (12),(13).

A first proposal to consider together the relative contributions of depth information and central bias has been done recently by Vincent et al. (14). In addition to potential high-level factors like lights and sky, the

foreground and central bias contributions are quantitatively studied. Results highlighted the potential role of foreground and central bias in saliency prediction. However, the contributions of these different visual features were fixed over time.

Ho Phuoc et al.(15) followed a similar methodology, but to study over time the role of some low-level visual guiding factors. Following their statistical analysis of the evolution of feature weights, an adapted saliency model was proposed. The pooling of feature maps was based on a set of learned weights. However, as in Vincent (14), these weights were fixed over time.

In this paper, we propose to design a time-dependent computational model of visual attention in order to predict where observers look at on still pictures for 2D and 3D conditions. The section 2 presents the materials as well as the eye tracking dataset (16). Behavioral and computational studies are presented in section 3. The fourth section describes the proposed time-dependent saliency model as well as its performances.

Thus, this article aims at answering 4 questions:

- Does the binocular disparity affect spatial locations of fixated areas, center and depth biases?
- Is the predictability of state-of-the-art bottom-up saliency models affected by stereo disparity?
- How to model the center and depth biases effects as individual features?
- How to include these features into a “time-dependent model” to predict salient regions over time?

2 Materials and methods

The eye tracking dataset provided by Jansen et al. (16) is used in this paper. The experimental conditions, i.e. materials and methods to construct this database in 2D and 3D conditions, are briefly reminded here. 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 (17).

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* (16)*)*

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 this kind of display is that it does not require special eyeglasses. A tracking system adjusts the two displayed views 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 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 1.

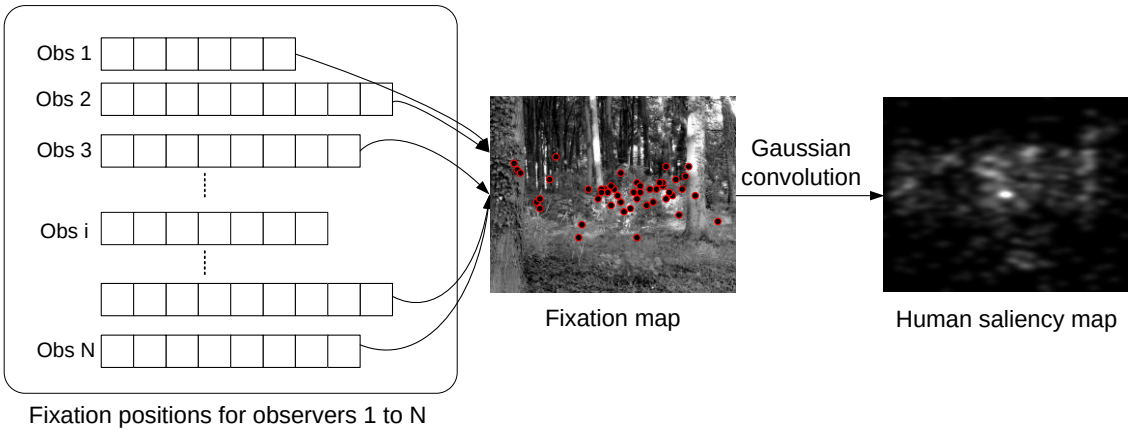

Figure 1 Illustration of the human saliency map computation from N observers

3 Behavioral and computational studies

Jansen et al. (16) 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 fixated i.e. human 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 analysis 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.

3.1 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 bound indicates a perfect discrimination whereas the lower bound indicates that the discrimination (or the classification) is at the chance level. The thresholded 3D human saliency map is then compared to the 2D human saliency map. For the 2D human saliency maps taken as reference, the threshold is set in order to keep 20% of the salient areas. For 3D human saliency maps, the threshold varies linearly in the range of 0 to 255. Figure 2 shows the AUC scores between these 2D and 3D human saliency maps obtained with different viewing times (the first 10 fixations (1-10), the first 20 fixations (1-20), etc). Cumulated fixations over time allow to deal with the increase of inter-observers dispersion over time and to put emphasis on salient areas due to the re-fixation trend over time (18). The median value is equal to 0.81 ± 0.008 (mean ± SEM). When analyzing only the first fixations, the similarity degree is the lowest. The similarity increases from 0.68 to 0.77 in a

significant manner (paired t-test, $p<0.01$). Results suggest that the disparity influences the overt visual attention just after the stimuli onset.

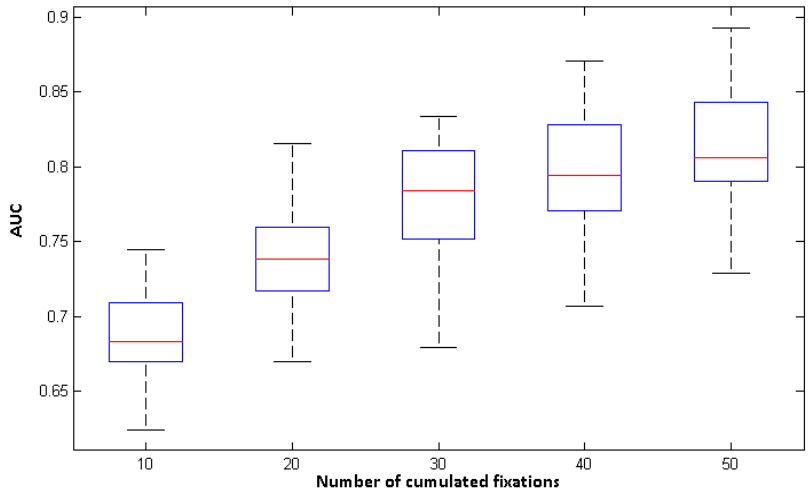

Figure 2: Boxplot of the AUC values between 2D and 3D human (experimental) saliency maps as a function of the number of cumulated fixations (the top 20% 2D salient areas are kept).

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. (16), 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 human saliency maps.

3.2 Center bias for 2D and 3D pictures

Previous studies have shown that observers tend to look more at the central regions than at the peripheral regions of a scene displayed on a screen. This tendency might be explained by a number of reasons (see for instance (11)). Recently, Bindemann (19) 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.

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 4 and Figure 5). As expected, 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 (note that this is different from the previous analysis where cumulated fixations over time were used. This is here less appropriate since the center bias is time-dependent). 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.

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 viewing time is an important but already known (11) factor, influencing the center bias. 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.

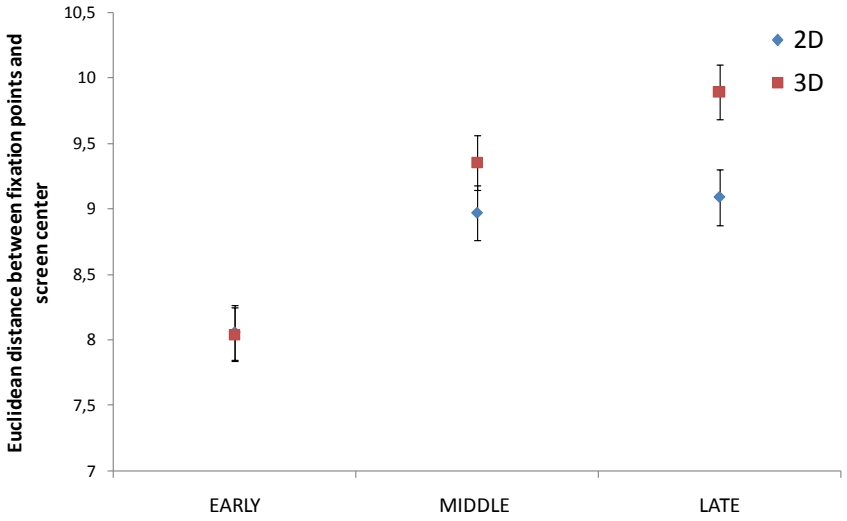

Figure 3: Average Euclidean distance between the screen center and fixation points. The error bars correspond to SEM (Standard Error of the Mean).

Bonferroni t-tests, however, showed that the central tendency between 2D and 3D conditions is not statistically significant 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 $p<<0.001$, respectively).

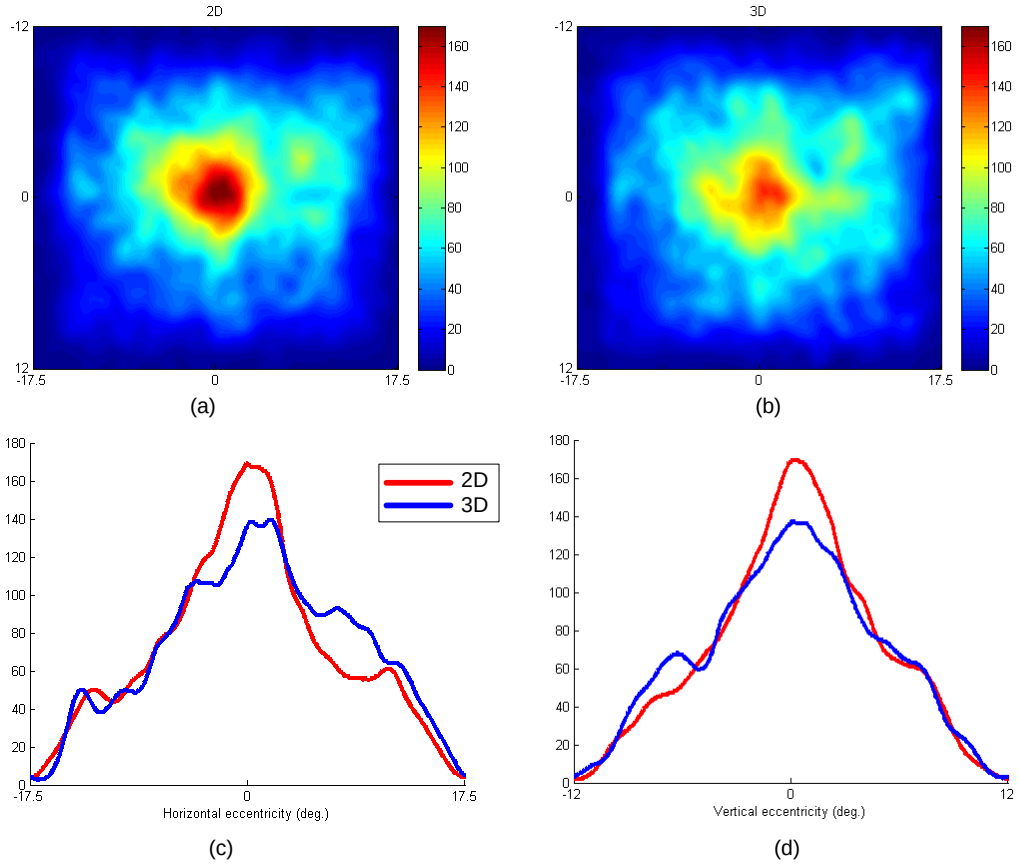

Figure 4: (a) and (b) are the distributions of fixations for 2D and 3D condition, respectively. (c) and (d) represent the horizontal and vertical cross sections through the distribution shown in (a) and (b). All the visual fixations are used to compute the distribution.

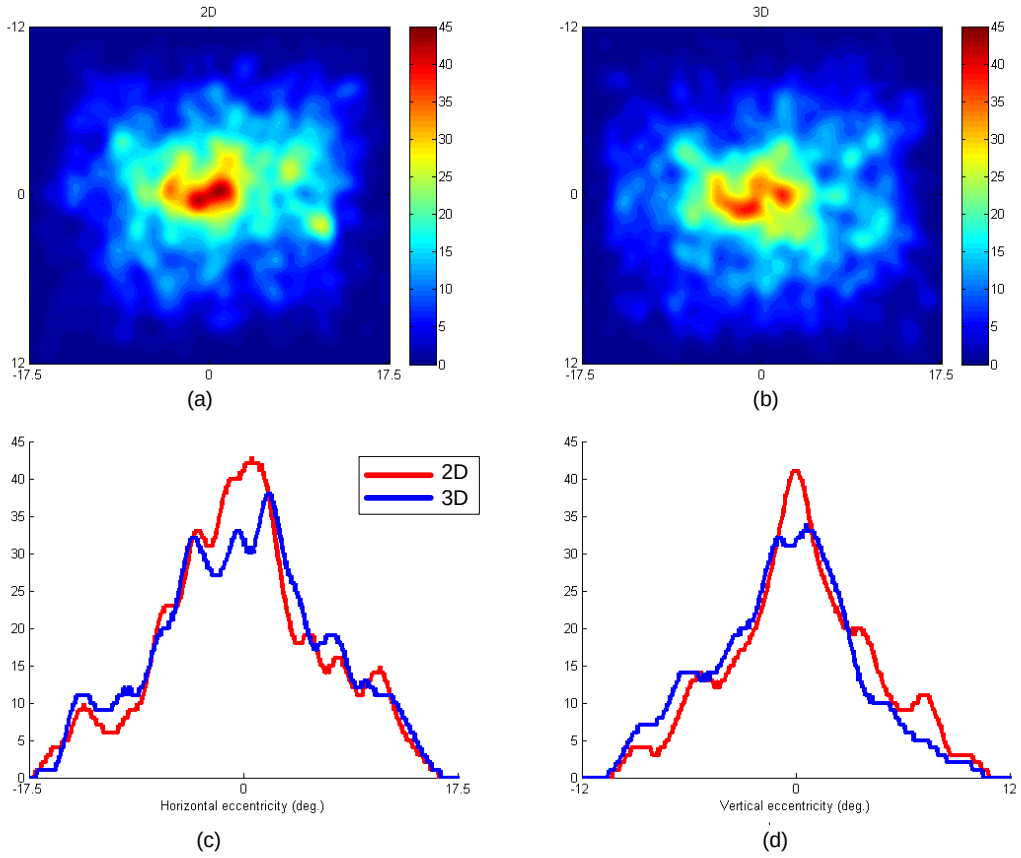

Figure 5: Same as Figure 4 but for the first 10 fixation points.

3.3 Depth bias: do we look first at closer locations?

In (16), a depth bias was found out suggesting that observers tend to look more to closer areas just after the stimulus onset than to farther areas. A similar investigation is conducted here but with a different approach. Figure 6(b) illustrates a disparity map: the lowest values represent the closest areas whereas the farthest areas are represented by the highest ones. Importantly, the disparity maps are not normalized and are linearly dependent on the acquired depth.

(a) (b)

Figure 6: Original picture (a) and its disparity map (black areas stand for the closest areas whereas the bright areas indicate the farthest ones).

The mean disparity is measured 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, a significant interaction is observed 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 7.

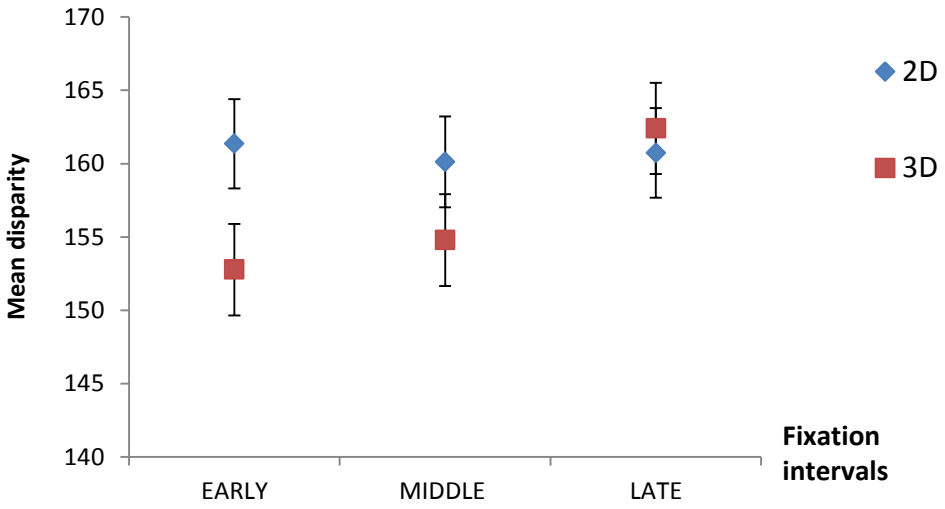

Figure 7: Mean disparity (in pixels) in function of the viewing time (early, middle and late). The error bars correspond to SEM (Standard Error of the Mean).

3.4 Can computational models of visual attention predict salient regions of 3D images?

As shown in previous sections, the introduction of binocular disparity significantly impacts our gaze on still images especially on the first fixations. Indeed, the last analysis (see section 3.3) indicates that the disparity induced by the stereoscopic condition effectively impacts the visual deployment: in stereo condition, we tend to look at closer locations on the first fixations.

Beyond the impact of binocular disparity on eye movement properties, it is interesting to assess the extent to which computational models of visual attention are able to predict where observers look at in stereoscopic conditions.

3.4.1 State-of-the-art models

In this study, three state-of-the-art models are used, two belonging to the biological inspired models and one to the statistical models:

- The model of Itti (2) was among the first to propose a method to compute topographic saliency map from a color image. From the input image, some early visual spatial features (color, intensity, orientation) are extracted, filtered out and then normalized and fused together to generate a final saliency map.
- A second model, Bruce and Tsotsos's model (20) is based on the assumption that a rare event is probably more salient than a non rare event. Saliency is then obtained by using the self-information of image's patches.
- Le Meur et al. (21) proposed an extension of Itti's model by adding human perception properties such as Contrast Sensitivity Function (CSF), hierarchical decomposition and visual masking mechanisms.

Figure 8 illustrates for a given picture the saliency maps computed by these three saliency models.

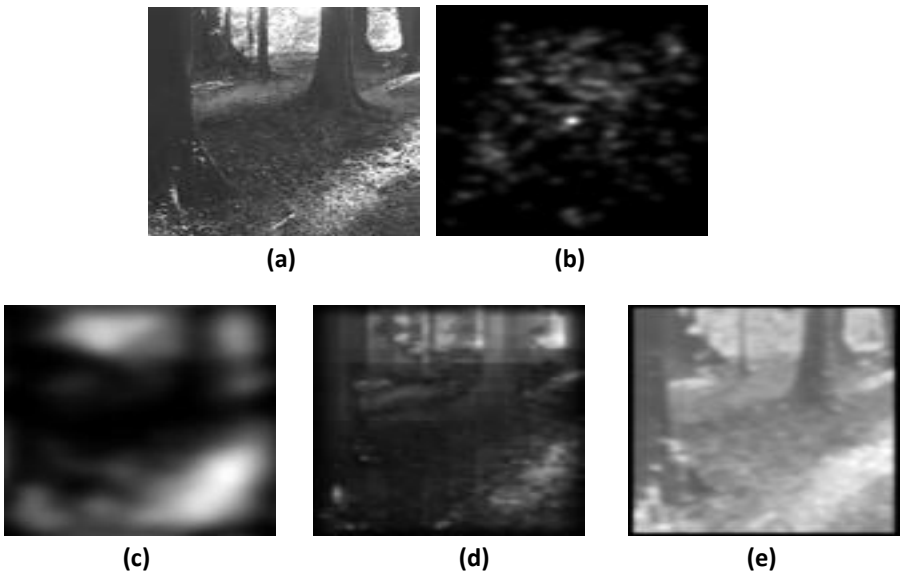

Figure 8 (a) Original left luminance image '4', corresponding experimental (b), and predicted Itti (c), Le Meur (d) and Bruce (e) saliency maps.

3.4.2 Performance over time slots in 2D and 3D conditions

As in section 3.2, the AUC (Area Under ROC Curve) is used, but to quantify the degree of similarity between human saliency maps (either in monoscopic i.e. "2D" or in stereoscopic i.e. "3D" conditions) and predicted saliency maps (as illustrated on Figure 8).

As previous analyses showed that the impact of depth is time-dependent, it is interesting to test the accuracy of model's prediction on the same time periods. Figure 9 illustrates the performance of the models on three independent slots of viewing times.

By considering the human saliency maps on the first 10, 20, 30 and 40 fixations, the only model showing a mean saliency prediction in 2D condition significantly different, and higher than in 3D conditions ($t(95) = 2.41$, $p<0.05$, $p=0.0088$) is the Itti's model.

By considering each fixation slot separately (i.e. we analyze the statistical difference for the first 10 fixations on the 24 AUC values, or for the 20 following etc), none of the models presents a mean saliency prediction in 2D condition significantly different than in 3D. One reason might be due to the small population involved in the test (24 pictures).

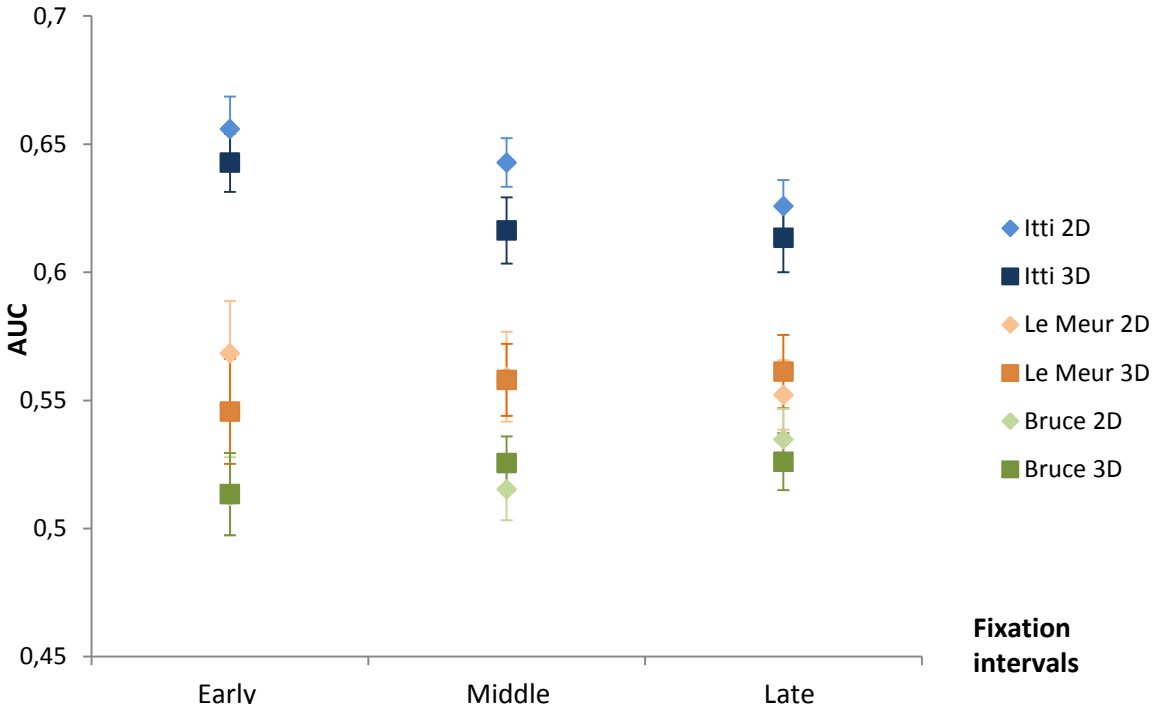

Figure 9 AUC values between predicted (i.e. model) saliency maps and 2D or 3D human saliency maps taken as reference (the top 30% salient areas are used) on the first 10, 20 and 30 fixation intervals, respectively early, middle and late.

Based on previous observations, we would expect that, due to the depth bias occurring on the first fixations, these models would show a loss of accuracy in 3D conditions on the first fixations in term of saliency prediction. Moreover the performance gap between conditions would tend to reduce with the viewing time. Results are however contrasted: Itti's model results confirm the aforementioned hypothesis, (but the difference is not statistically significant) while the models of Le Meur and Bruce show no clear distinction.

Finally, in order to validate the fact that predicted saliency maps do not contain low-order features that would artificially increase the AUC values of each model, the predicted saliency maps are block-randomize for each model by block size of 64 by 64 pixels. Mean AUC values for randomize saliency maps versus human 2D or 3D saliency maps confirm that the degree of similarity between randomized saliency maps and human ones is at the chance level. (Itti : 0.497 in 2D and 0.501 in 3D, Le Meur : 0.489 and 0.490 in 3D, Bruce : 0.496 and 0.499 in 3D).

3.5 Conclusion

In the behavioral first part of this study, 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. (16), and support the existence of a *depth bias*.

The ability of existing models to predict where observers look at in stereoscopic condition is relatively lower, especially for Itti and Le Meur models. Indeed these models show a decrease of their performances in 3D condition. To improve their performance, these models are extended by taking into account the time-dependent depth bias. This is described in next sections.

4 Time-dependent saliency model

Recent studies (11), (22) 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 (23), (24), (25), (26). But considering their interactions and impacts over time is crucial to improve the predictability of existing saliency models (11), (24).

4.1 Statistical analysis

Following the temporal behavioral study, we include 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 followe a similar approach to Vincent's et al. one (14).

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:

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

where *K* is the number of features, ϕ_k the probability density function for each feature *k* and $p_k(t)$ the contribution or weight of feature *k* with the constraint that $\sum_{k=1}^{K} p_k(t) = 1$, 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 (27). Before explaining this method, we describe the center and depth modeling.

4.1.1 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 (19). However, the extent to which this potential laboratory artifact is an inherent feature of strategy of human vision remains an open subject. Tatler (11) 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 (11), 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 (13). As proposed in (15), 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.

The central bias is then modeled by a time-independent bi-dimensional Gaussian function, centered at the screen center as $N(0,\sum_t)$, with $\sum_t = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ the covariance matrix and with σ_x^2 and σ_y^2 the variance. We fit the bi-dimensional 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 ($\sigma_{x2D} = 4.7°, \sigma_{y2D} = 2.5°, \sigma_{x3D} = 4.3°, \sigma_{y3D} = 2.3°$).

4.1.2 Model of the depth

Results presented in section 3.3 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 (28), that can be understood as an element of the edge interpretation depth cue (29), drives the visual attention pre-attentively (30). This supports our choice of figure-ground organization implementation by a segregation 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 maps are illustrated on Figure 10 (a).

4.1.3 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 (see equation (1)). 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 significant to predict where we look at.

In summary five feature maps are used as illustrated in Figure 10 (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”;
- one for the central fixation bias;
- two related to the depth cue, i.e. the foreground and background features;
- a uniform distribution feature

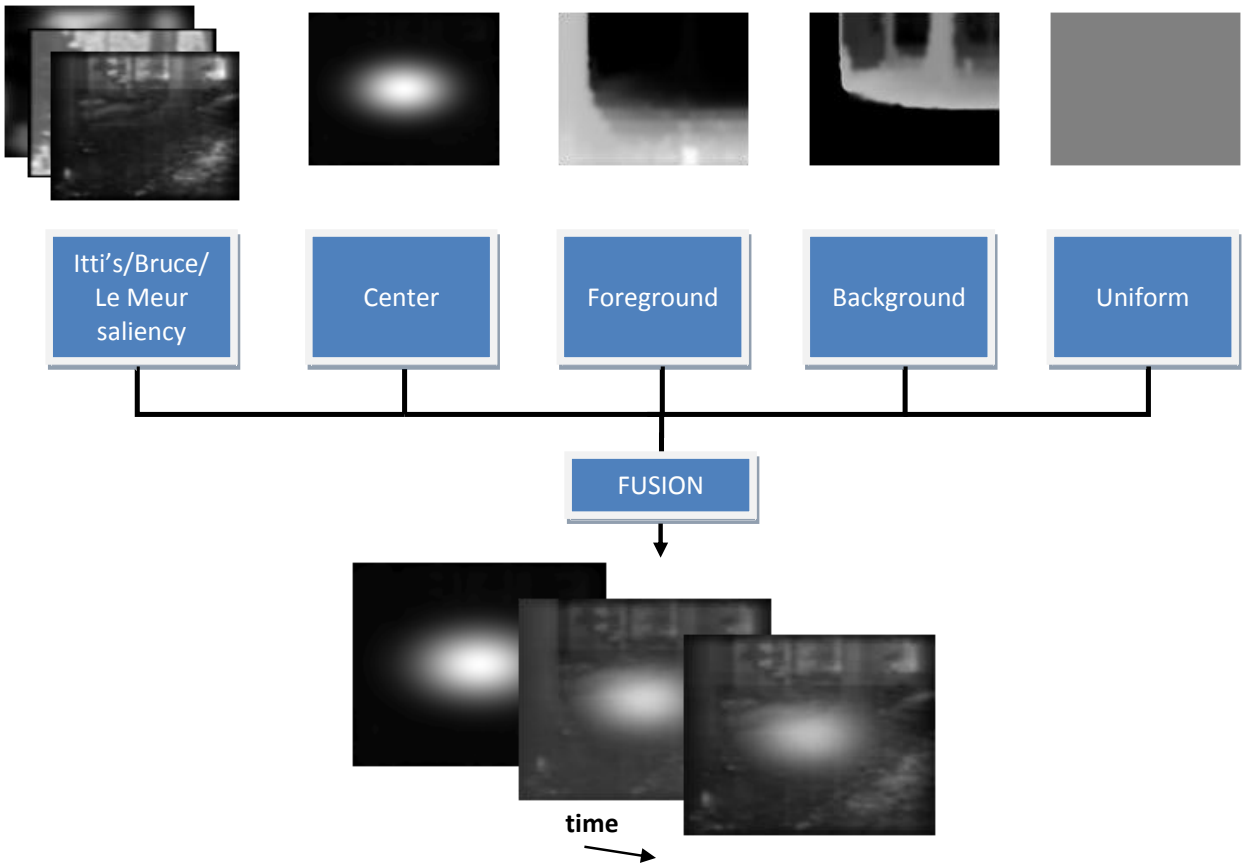

Figure 10 (a) Upper Row: Illustration of Itti’s saliency map obtained from image originally presented in Figure 8 (a), center bias in 2D condition, corresponding foreground and background feature maps.
(b) Middle row: Description of the proposed time-dependent model.
(c) Lower Row: Illustration of the 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).

Low-level saliency,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:

$$f(x,t) = p_{Sm}(t)\cdot\phi_{Sm}(x) + p_{Cb}(t)\cdot\phi_{Cb}(x) + p_{Fg}(t)\cdot\phi_{Fg}(x) + p_{Bg}(t)\cdot\phi_{Bg}(x) + p_{Un}(t)\cdot\phi_{Un}(x) \qquad (2)$$

with ϕ_{Sm} the saliency maps of one of the 3 state-of-the-art models, ϕ_{Cb} the central Gaussian function, ϕ_{Fg} and ϕ_{Bg} the foreground and background map respectively and ϕ_U the uniform density function. Each feature is homogeneous to a probability density function. $p_{Sm}(t)$, $p_{Cb}(t)$, $p_{Fg}(t)$, $p_{Bg}(t)$, and $p_{Un}(t)$ are the time-dependent weights to be estimated, their sum being equal to unity.

Figure 10 (a) gives an illustration of the involved features. The pseudo-code of the Figure 11 describes the EM algorithm. The weights $p_k^{(m)}(t)$ are the only parameters estimated for each iteration m. In practice, a fixed number M of 50 iterations is a good tradeoff between estimation quality and complexity.

With $t_k = \{t_{Sm}, t_{Cb}, t_{Fg}, t_{Bg}, t_U\}$ *the estimated missing probability for each feature*

Initialization of the weights $p_k^{(0)}(t) = 1/K \quad \forall k$

for each fixation rank from 1 to 25

for each iteration $m = 1..M$,

for each feature $k = 1..K$,

for each fixation $i = 1..N$,

Expectation step :

Given a current estimate of the parameters $p_k(t)$, t_k *is computed:*

$$t_{i,k}^m = P\{x_i \ comes \ from \ the \ feature \ k\}$$

$$t_{i,k}^m = \frac{p_k^{(m-1)}\phi_k(x_i)}{\sum_{l=1}^{K} p_l^{(m-1)}\phi_l(x_i)}$$

end

Maximization step:

The parameters $p_k^{(m)}(t)$ *are updated for the iteration* m *:*

$$p_k^{(m)}(t) = \frac{\sum_{i=1}^{N} t_{ik}^{(m)}}{N}$$

end

end

end

Figure 11 Pseudo-code of the EM algorithm.

The temporal contributions of the proposed features to visual attention are evaluated. The EM-based mixture model is run on half of the image dataset at each fixation rank (from the first to 25th 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 process is repeated at each fixation rank, and with fixations in 2D and 3D conditions

4.2 Results of the statistical analysis

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 resulting temporal contributions of all the visual guiding factors are illustrated on Figure 12.

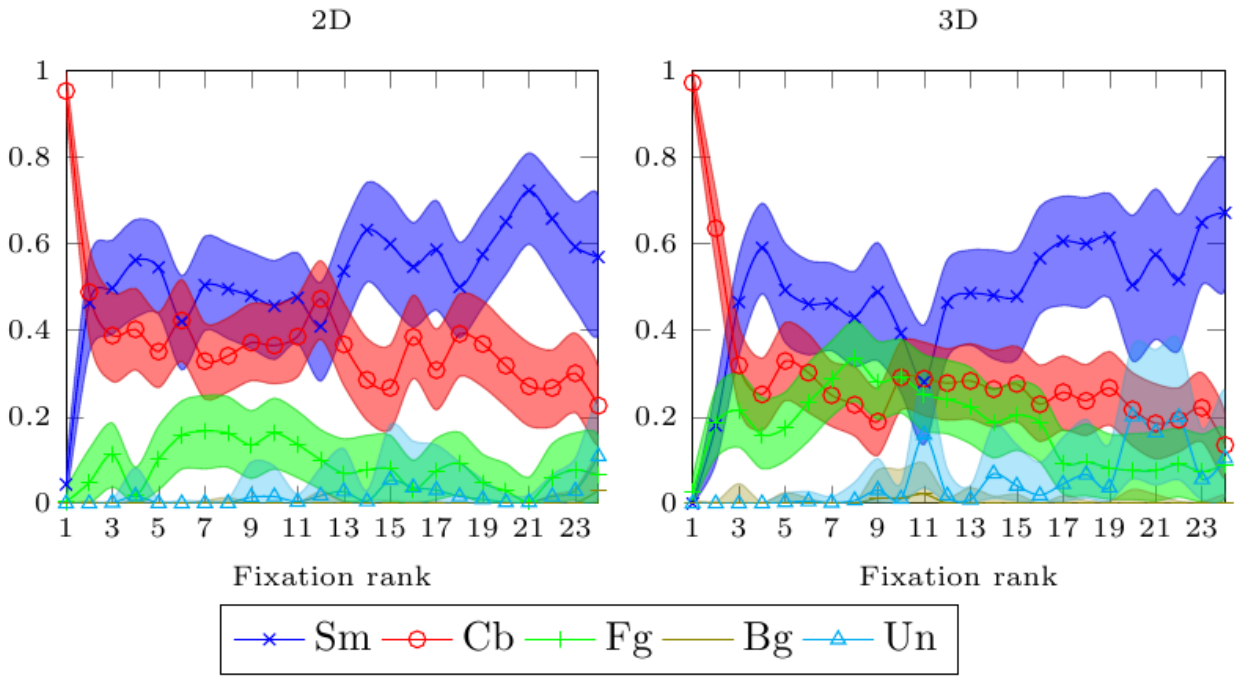

Figure 12 Temporal contributions (weights) of 5 features from 2D (left) and 3D (right) fixations as a function of the fixation rank. Low-level saliency feature ("Sm") here comes from Itti's model. The error areas at 95% are computed by a "bootstrap" estimate (1000 replications).

The best predictor for both viewing conditions is the predicted low-level saliency (from Itti's model and called Sm on Figure 12). 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 3rd to the 18th 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.

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 (11) and in accordance with (15), 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, this tendency exists even with a marker positioned randomly within a circle of 10° radius from screen center (11). 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 all the more in stereo condition where its contribution is much more important. This is coherent with our previous conclusions (cf. section 3.3). 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 22^{th} 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 do not 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 maps 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 19th fixations, a time slot for which the contribution of uniform distribution is stable and low in all conditions.

4.3 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 (2) 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 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 (2)) 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 weighting and summing all features.

Thereafter, we use 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 3.4.2. 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.

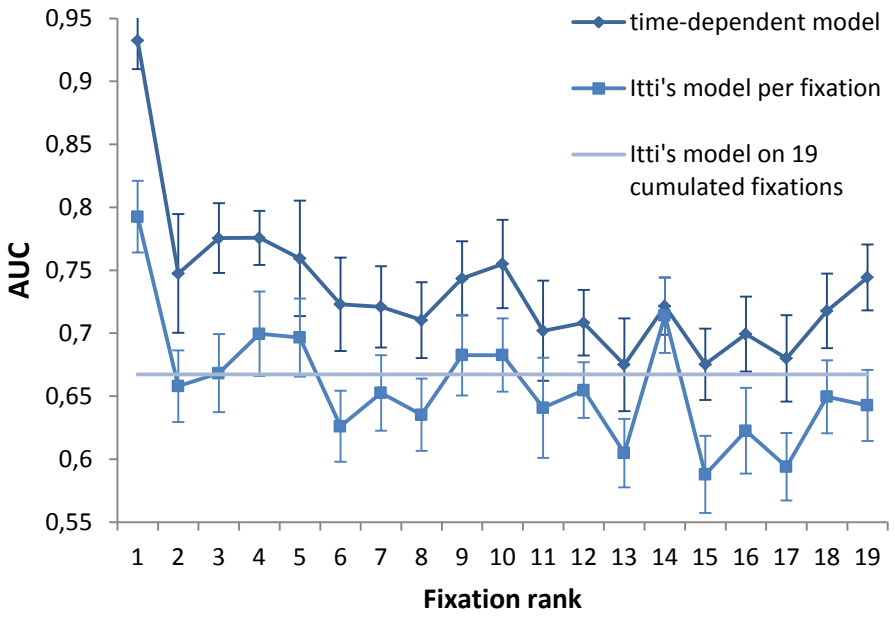

Figure 13 Temporal evolution of the performance of the time-dependent model based on Itti's, versus the Itti's model per fixation, and versus the Itti's model on 19 cumulated fixations

The AUC values of original Itti's model fixation per fixation are plotted in Figure 13 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.

To strengthen the analysis, the "Normalized Scanpath Saliency" (NSS) (31) 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/fixation rank". Results are also averaged over all participants and all images for each fixation rank.

The Figure 14 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.

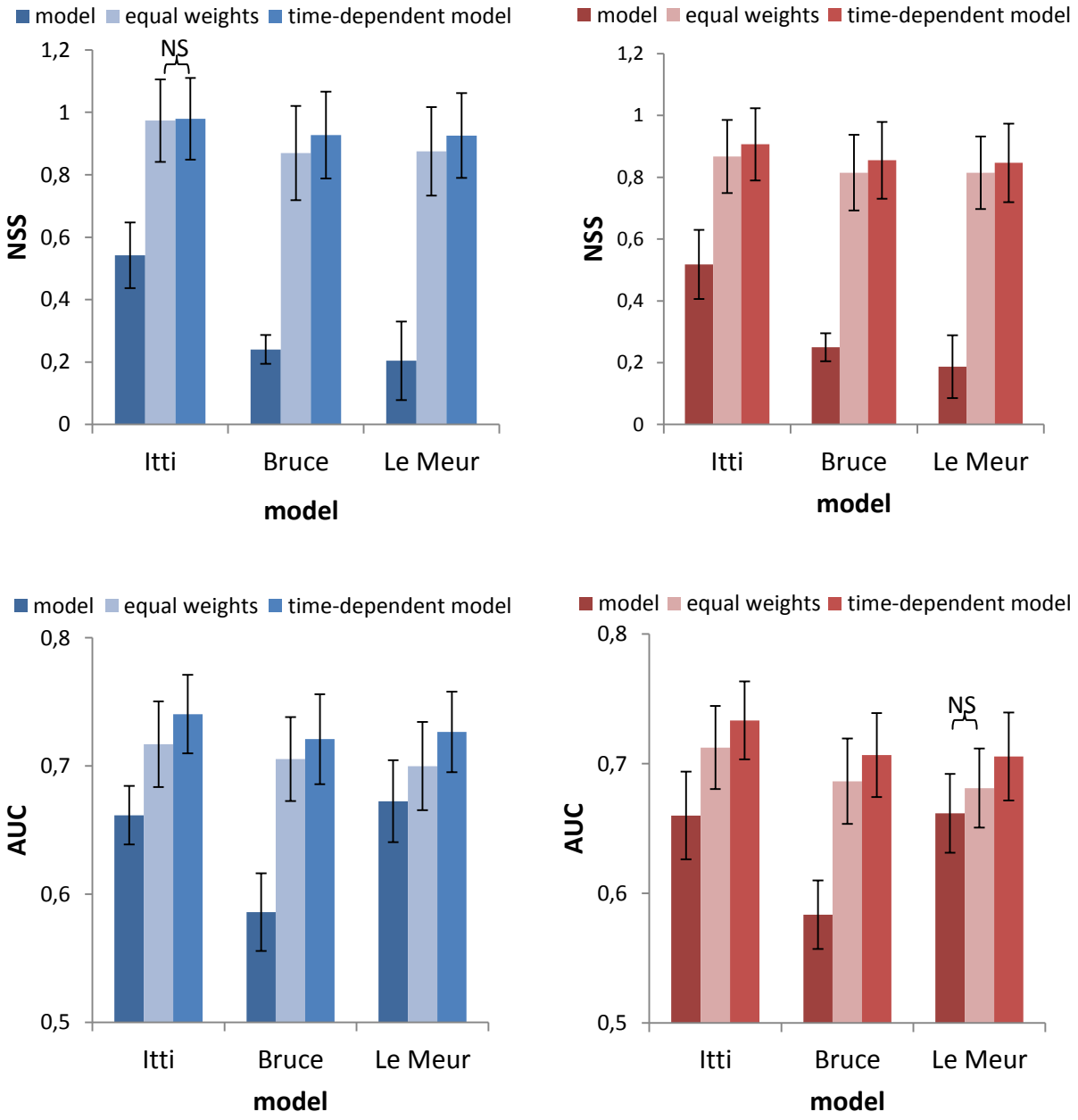

Figure 14 Comparison between 6 saliency models in 2D (left) and 3D conditions (right). Upper row: the NSS criterion, lower row the AUC criterion. The error bars correspond to the SEM. NS corresponds to Non-Significant. When the term NS is not indicated, results are significantly different ($p<0.05$).

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 5 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 14).

The time-dependent model based on Itti's low-level saliency 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 saliency models, both 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.

Discussion

The proposed approach based on time-dependent weighting improves the performance of existing attention models. 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 based on low-level visual features, 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 improve their predictability (32). 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 (33). 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 do not 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, observers do not only 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 figure-ground organization in the bottom-up visual attention.

5 Conclusion

The purpose of this study was to assess the differences in the visual deployment in monoscopic and stereoscopic viewing conditions, and to evaluate the contributions of relevant features that might participate to the visual attention. In addition, we propose a new saliency model in which a time-dependent pooling of relevant features is used to predict where we look at on natural scenes.

Behavioral observations first underline that visual exploration in a depth layer detection task significantly differs with the introduction of binocular disparity. This lasts up to the 30^{th} fixation. If the influence of viewing time on center bias is already demonstrated, our result suggests that this central

tendency significantly differs between 2D and 3D conditions. This is particularly true in the middle and late time. Moreover a depth bias is also observed: participants tend to look at closer areas with the introduction of binocular disparity, and significantly in the early time.

Following these observations on external center and depth biases, 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 (30). Systematic recentering tendency followed by 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 (15) in the case of low-level feature combination, this hypothesis is very simple with regards to the complexity of visual attention construction (34), and with regards to other computational proposals of fusion (35). 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.

6 Acknowledgment

Authors would like to thank Lina Jansen, Selim Önat and Peter König for providing us the eye tracking database and giving us helpful information and comments for this study. We also would like to thank Tien Ho-Phuoc for the support on bootstrap estimate.

7 References

1. Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol. 1985;4(4):219–27.
2. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20(11):1254–9.
3. Yarbus AL. Eye movements and vision. Plenum press: New York; 1967.
4. Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search.

Psychological review. 2006;113(4):766–86.
5. Cutting JE, Vishton PM. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. Perception of space and motion. 1995;5:69–117.
6. Maki A, Nordlund P, Eklundh JO. A computational model of depth-based attention. Proceedings of the 13th International Conference on Pattern Recognition, 1996. p. 734–9.
7. Maki A, Nordlund P, Eklundh JO. Attentional scene segmentation: integrating depth and motion. Computer Vision and Image Understanding, 2000;78(3):351–73.
8. Ouerhani N, Hugli H. Computing visual attention from scene depth. Proceedings of 15th International Conference on Pattern Recognition, 2000. p. 375–8.
9. Zhang Y, Jiang G, Yu M, Chen K. Stereoscopic visual attention model for 3D video. Advances in Multimedia Modeling. 2010;314–24.
10. Bruce NDB, Tsotsos JK. An attentional framework for stereo vision. Proceedings of The 2nd Canadian Conference on Computer and Robot Vision, 2005. p. 88–95.
11. Tatler BW. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision. 2007;7(14).
12. Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. IEEE 12th International Conference on Computer Vision. 2009. p. 2106–13.
13. Zhao Q, Koch C. Learning a saliency map using fixated locations in natural scenes. Journal of vision. 2011;11(3).
14. Vincent BT, Baddeley R, Correani A, Troscianko T, Leonards U. Do we look at lights? Using mixture modelling to distinguish between low-and high-level factors in natural image viewing. Visual Cognition. 2009;17(6):856–79.
15. Ho-Phuoc T, Guyader N, Guerin-Dugue A. A Functional and Statistical Bottom-Up Saliency Model to Reveal the Relative Contributions of Low-Level Visual Guiding Factors. Cognitive Computation. 2010;2(4):344–59.
16. Jansen L, Onat S, König P. Influence of disparity on fixation and saccades in free viewing of natural scenes. Journal of Vision. 2009;9(1).
17. Steger JM, König P. Fusion of 3D Laser Scans and Stereo Images for Disparity Maps of Natural Scenes. Institute of Cognitive Science, Osnabruck; 2010.
18. Le Meur O, Baccino T, Roumy A, others. Prediction of the Inter-Observer Visual Congruency (IOVC) and application to image ranking. 2011;
19. Bindemann M. Scene and screen center bias early eye movements in scene viewing. Vision research. 2010;
20. Bruce ND. Features that draw visual attention: an information theoretic perspective. Neurocomputing. 2005;65:125–33.
21. Le Meur O, Le Callet P, Barba D, Thoreau D. A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006;28(5):802–17.
22. Zhaoping L, Guyader N, Lewis A. 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. 2009;9(11).
23. Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. Vision Research. 2002;42(1):107–23.
24. Tatler BW, Baddeley RJ, Gilchrist ID. Visual correlates of fixation selection: Effects of scale and time. Vision Research. 2005;45(5):643–59.

25. Underwood G. Cognitive processes in eye guidance: algorithms for attention in image processing. Cognitive Computation. 2009;1(1):64–76.
26. Cutsuridis V. A cognitive model of saliency, attention, and picture scanning. Cognitive Computation. 2009;1(4):292–9.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1977;1–38.
28. Rubin E. Visuell wahrgenommene figuren: Studien in psychologischer analyse. Gyldendalske boghandel; 1921.
29. Palmer S. Vision: From photons to phenomenology. Cambridge, MA: MIT Press; 2000.
30. Qiu FT, Sugihara T, von der Heydt R. Figure-ground mechanisms provide structure for selective attention. Nature neuroscience. 2007;10(11):1492–9.
31. Peters RJ, Itti L. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. IEEE Conference on Computer Vision and Pattern Recognition. 2007. p. 1–8.
32. Jost T, Ouerhani N, Wartburg R, M\üri R, H\ügli H. Assessing the contribution of color in visual attention. Computer Vision and Image Understanding. 2005;100(1-2):107–23.
33. Le Meur O, Chevet JC. Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. IEEE Transactions on Image Processing, 2010;19(11):2801–13.
34. VanRullen R. Visual saliency and spike timing in the ventral visual pathway. Journal of Physiology-Paris. 2003;97(2-3):365–77.
35. Chamaret C, Chevet JC, Le Meur O. Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. 17th IEEE International Conference on Image Processing (ICIP), 2010. p. 1077–80.