Livrable D1.3 of the PERSEE project - Perceptual Modelling: Softwares results and final report.

Junle Wang, Josselin Gautier, Olivier Le Meur, Emilie Bosc, Vincent Ricordel

Projet PERSEE
SCHÉMAS PERCEPTUELS ET CODAGE VIDÉO 2D ET 3D
n° ANR-09-BLAN-0170

Livrable **D1.3** *01/07/2013*

Perceptual Modelling: Softwares results and final report.

Junle	WANG	IRCCyN
Josselin	GAUTIER	IRISA
Olivier	LE MEUR	IRISA
Emilie	BOSC	INSA
Vincent	RICORDEL	IRCCyN

Contents

1 Introduction

This document completes and finalises the list of perceptual models which have been designed over the project PERSEE.

In section II, a measurement of visual discomfort is presented. Section III is dedicated to the saliency detection for stereoscopic images. Section IV shows the existence of depth bias on natural images, and Section V presents a time-dependent visual attention model in stereoscopic condition combining center and depth bias.

2 Measurement of visual discomfort

The success of 3DTV applications depends on the ability of 3D systems to provide experiences with high quality. Despite the enhanced sensation of depth brought by stereoscopic 3D video applications, the Quality of Experience (QoE) may be reduced by new drawbacks such as that of visual discomfort. This is a new dimension to be considered when assessing the QoE of stereoscopic 3D systems, in addition to the conventional perceived image quality and the depth quantity.

In this section, our study and establishment of model of visual discomfort in stereoscopic 3D is presented. This work is based on previous studies [22, 23, 24] that already drew the main features of the visual discomfort model. In the following, the implementation is based on an internship results achieved during the project. First, the basic principles of the model will be described. Then the structure of the implemented model will be presented. Finally the validation process will show the performances of the model.

2.1 Visual discomfort model [22]: principles

An object's motion in stereoscopic video sequences can be categorized as in-depth or planar motion. Some studies showed that motion in depth, i.e., the magnitude of binocular disparity varying over time, could have great influence on visual discomfort, maybe even more important than the level of absolute disparity. In [22], Li et al. investigated the effects of relative disparity and planar motion velocity on visual discomfort through subjective assessment tests with non-expert observers. They showed that the relative disparity between the foreground and background in the stimulus might be more significant in visual discomfort perception than the binocular disparity of the foreground. Planar motion with faster velocity may result in more visual discomfort. The authors proposed two different models for visual discomfort that are expressed as

follows:

$$Q = a_1 \cdot v + a_2 \cdot d + a_3 \tag{1}$$

$$Q = b_1 \cdot d + b_2 \cdot v + b_3 \cdot d \cdot v + b_4 \tag{2}$$

where, Q represents visual discomfort,v is the velocity (degree/s) and d is the relative angular disparity (degree), the predicted coefficients for the two models were 0.0018, 0.2102, -0.0477 for a_1, a_2 and a_3, and 0.3110, 0.0026, -0.0006, -0.1888 for b_1, b_2, b_3 and b_4, respectively. The results presented in [22] showed that the models correlate quite well with the subjective perception of visual discomfort.

In [23] the authors conducted a multiple linear regression analysis, in order to refine the proposed model. The authors distinguish static stimuli and motion stimuli. The prediction of visual discomfort of motion stimuli considers four different factors: disparity amplitude, relative angular disparity offset, planar motion velocity and in-depth motion velocity. Relative disparity offset is predominant in the visual discomfort scores of static stimuli, while both relative angular disparity offset and planar velocity are key factors in the visual discomfort score of motion stimuli. In the following, the general model to predict visual discomfort relies on the following expression:

$$VD_s = -8.04 + 2.41 \times r_0 \tag{3}$$

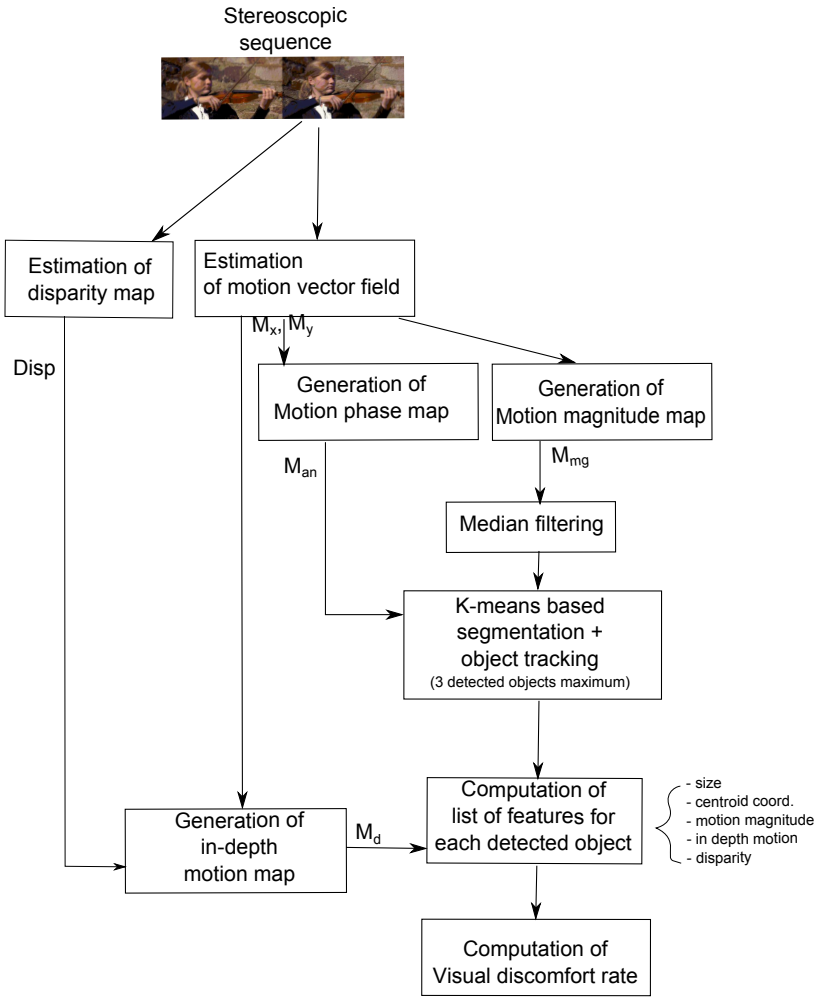
$$VD_p = -9.81 + 1.47 \times r_0 + 0.3 \times v_p - 0.07 \times r_0 \times v_p \tag{4}$$

$$VD_d = -6.04 + 1.23 \times r_0 + 0.31 \times v_d + (0.45 - 0.21 \times r_0) \times d_a \times v_p \tag{5}$$

whereVD_s, VD_p and VD_d stand for visual discomfort caused respectively by static, planar motion and in-depth conditions; r_0 is the relative angular disparity, d_a is the disparity amplitude, v_p is the planar motion velocity, v_d is the in-depth motion velocity. VD_s, VD_p and VD_d are negative values, so their range lie in $[-\infty - 0]$. The lower the absolute visual discomfort score value, the higher the perceived visual discomfort induced by the stereoscopic sequence, according to the model.

2.2 Structure of implemented model

The model of visual discomfort, defined by Eq. 3, 4 and 5, has been entirely implemented using MATLAB version 7.11.0 on Windows 7, through an Intel i5 processor, 2.67 GHz, with a 4 gb RAM. In particular, it has been designed with the assistance of MATLAB Signal Processing, Image Acquisition, Image Processing and Statistics Toolboxes. A general overview of the system is presented in Fig. 2.2. The computation of the visual discomfort score is based on the consideration of features of detecting moving objects. The steps depicted in this figure are discussed in the following.


Stereoscopic sequence
Estimation of disparity map
Estimation of motion vector field
M_x, M_y
Disp
Generation of Motion phase map
Generation of Motion magnitude map
M_{an}
M_{mg}
Median filtering
K-means based segmentation + object tracking
(3 detected objects maximum)
Computation of list of features for each detected object
- size
- centroid coord.
- motion magnitude
- in depth motion
- disparity
Generation of in-depth motion map
M_d
Computation of Visual discomfort rate

2.2.1 Estimation of motion vectors and disparity map

Motion vectors components (M_x and M_y along the horizontal axis and along the vertical axis respectively) and the disparity map are first computed. The in-depth motion map, noted M_d and is obtained through a pixel-by-pixel difference between the disparity map of the current frame and the disparity map of the previous one compensated with motion information of the current frame as follows:

$$M_d(x,y) = Disp(x,y) - Disp(x - M_x(x,y), y - M_y(x,y)) \quad (6)$$

where (x,y) is the pixel coordinates.

2.2.2 Estimation of motion magnitude and phase

The motion magnitude and the motion phase maps (M_mg and M_an respectively) are the absolute value and the angle respectively, of the estimated motion vectors in M_x and M_y. The motion phase map is useful for enhancing the robustness of the segmentation step that will be discussed later. M_mg is then filtered through a median filter in order to reduce the noise due to the inaccurate motion field estimation.

2.2.3 Segmentation for the detecting moving objects

After the median filtering step, a K-mean based segmentation is applied. The goal is to detect three main regions. The motion magnitude map (M_mg) of a frame is binarized. The threshold level used for binarization is obtained through K-means algorithm (Ridler and Calvard,1978, [32]): based on an initialization value, the algorithm finds the means above and belove the current value, and it updates the threshold to the average between these means. The algorithm iterates until it converges into a fixed value. Then, from this binary image the algorithm removes all connected components with less than 256 pixels. The connectivity required for the new binary image is eight. Then, the objects which have remained into the binary image are labeled (the pixels labeled with $'0'$ are those belonging to the background, those labeled with $'1'$ belong to the first region of the frame, and so on). At this stage, the number of detected regions can be larger than three. We assume that three is a reasonable number of detected objects for the estimation of visual discomfort. So if larger than three, the number of detected objects should be reduced and only the object with largest motion magnitude are considered in the estimation of visual discomfort. Thus, for each of these regions the algorithm evaluates the mean values of magnitude motion, angular motion and disparity, and so it creates three matrices of mean values. As a consequence, three thresholds for similarity criteria are needed: this is done through Otsu's method, which chooses the threshold to minimize the intraclass variance of the thresholded matrices

(one for magnitude motion, one for angular motion, one for disparity). As a result, the merging step of the segmentation can begin: if two adjacent regions are similar (difference of respective mean values below the selected thresholds) in terms of motion (both magnitude and angular) or disparity, then they are merged into a new single region. The algorithm iterates until no more merging can be done. The reason to have introduced a threshold for angular motion is that it can avoid the merging of two adjacent regions which are actually moving in different directions. Finally, just the three biggest objects of each frame remain, while the others are deleted since it is assumed that they may have an irrelevant influence on visual discomfort.

2.2.4 Object tracking

The object tracking step relies on features of the detected objects: size(in terms of pixels), centroid(in terms of 'x' and 'y' coordinates) and mean intensity of magnitude motion, in-depth motion and disparity. For every single frame, each object is compared with each object of the previous frame: if a similarity between two objects belonging to two consecutive frames is found, then they are considered as the same object. The comparison involves the five properties found in advance: all of those properties must be similar among the objects so as to consider them as the same object. Thus, there are five different thresholds (one per feature): difference of sizes, centroid positions, mean intensities of magnitude motion, in-depth motion and disparity. Previous studies related to timing of human visual perception [37, 1] suggest that objects appearing less than 150 ms in a video sequence are likely to be not perceived. Thus these objects should not influence the visual discomfort perception. This aspect has been considered in the implementation of the model, based on the frame rate of the input stereoscopic sequences.

2.2.5 Visual discomfort score

The final score should be based on the three visual discomfort components described by Eq. 3, 4 and 5. Many pooling methods have been explored in order to output the final visual discomfort score of the sequence. The different possible approaches can be considered:

- Frame-based: the three Visual Discomfort indicators (VD_s, VD_p, VD_d) are evaluated for every single object o_{ip} of each frame and then the final score is a combination of each frame visual discomfort score;

- Object-based: the three Visual Discomfort indicators (VD_s, VD_p, VD_d) are evaluated for each tracked object O_i into the sequence and then the final score

is a combination of the visual discomfort score of each detected object of the sequence;

- Sequence-based: the three Visual Discomfort indicators (VD_s, VD_p, VD_d) are evaluated for the entire sequence.

where O_i stands for the i-th object tracked into the sequence; o_{ip} represents the p-th occurrence of the i-th object. For each approach, the pooling method can then consist of considering the average, the median or the maximum value out of the three visual discomfort component output for the sequence. in the following, the performances of the different tested approaches will be presented. N, M and K stand respectively for the number of frames, the number of tracked objects and the number of occurrences of the i-th tracked object into the sequence

$$VD_{average} = \frac{1}{3}\sum(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_s(o_{ip}))+(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_p(o_{ip}))+(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_d(o_{ip})) \tag{7}$$

$$VD_{maximum} = \max[(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_s(o_{ip}))+(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_p(o_{ip}))+(\frac{1}{K \times M}\sum_{i=1,p=1}^{K,M} VD_d(o_{ip}))] \tag{8}$$

$$VD_{averagepost} = \frac{1}{3}\sum(\frac{1}{M}\sum_{i=1}^{M} VD_s(O_i)) + (\frac{1}{M}\sum_{i=1,p=1}^{M} VD_p(O_i)) + (\frac{1}{M}\sum_{i=1}^{M} VD_d(O_i))] \tag{9}$$

$$VD_{maximumpost} = \max[(\frac{1}{M}\sum_{i=1}^{M} VD_s(O_i))+(\frac{1}{M}\sum_{i=1,p=1}^{M} VD_p(O_i))+(\frac{1}{M}\sum_{i=1}^{M} VD_d(O_i)) \tag{10}$$

2.3 Validation of the model

This subsection presents the results of the tests conducted for estimating the performances of the proposed model. The first part will present the sequences used for estimating the visual discomfort scores. The second part discusses the results of the estimations.

2.3.1 Experimental material

An existing database made available from the Image and Video Communication (IVC) research group of IRCCyN composed of sixty four stereoscopic sequences were used. Subjective tests were previously been conducted on these 64 stereoscopic 3D sequences, assessing the quality of the stereoscopic contents regarding 3 different factors. There are no compression artifacts in this database (or they are not perceptible) and no view synthesis related distortions. The sequences have been post-processed just through a resize, since they were not natively in Full-HD. The test has included a room with a 3D display, an interface to let the viewer manage the test and give his grades to the sequences, and a Hard-Disk capable of storing all the videos involved. The test has been conducted following ITU-R BT.500-11 recommendations (i.e. lighting conditions and observation angle of the viewers). Furthermore, the 3D display that has been chosen is an Alienware Optx 23", with a resolution of 1920x1080 at 120Hz. "VideoQuality.exe" is the Remote 3D player software which has been used, while users have interacted with a Matlab voting interface displayed on a different screen. This consists of 3 different scales: depth sensation, quality of experience, and visual comfort, all with five different levels. Here below are some others specifications:

- Test room luminance: 51.49 cd/m^2
- Screen without glasses luminance: 0.38 / 342cd/m^2.
- Number of viewers : 25.
- Observation distance: 87cm (3 times the height of the screen).
- Test duration: 1 session of 55 min.
- Number of sequences: 64.

2.3.2 Model performances

The sequences considered for assessing the performances of the model included various cases (i.e. static, planar motion, in-depth motion) and a single video sequence contained various cases. Yet, the model proposed by Li et al. was based on experimentations including artificial stereoscopic conditions. In Li et al. experiments, one sequence included only one type of case (i.e. either static, or planar or in-depth motion). Real sequences such as those considered for the validation of the implemented model are likely to contain various cases. We assume that the most predominant case should express the best the final visual discomfort score. Thus, the sixty four sequences have been clustered into three groups: one containing just the sequences

with low levels of motion (both planar and in-depth), one containing sequences with a predominance of planar motion and one including sequences with presence of in-depth motion. Either VD_s, or VD_p or VD_d were calculated for the sixty four sequences, depending on the category they belong to. For example, VD_s is calculated for any sequence belonging to the sequences categorized in the "static" group; VD_p is calculated for any sequence belonging to the sequences categorized in the "planar motion" group; VD_d is calculated for any sequence belonging to the sequences categorized as "in-depth motion" group. Table 2.3.2 shows the Pearson linear correlation coefficients (PLCC) of the different visual discomfort methods with respect to the subjective scores. First of column of Table 2.3.2 refers to the PLCC scores when considering the whole set of sequences. Second column of Table 2.3.2 refers to the PLCC scores when considering the group categorezd as "static" only. Third column of Table 2.3.2 refers to the PLCC scores when considering the group categorezd as "planar motion" only. Fourth column of Table 2.3.2 refers to the PLCC scores when considering the group categorezd as "in-depth motion" only.

Name of the method	PLCC for all	PLCC for Static comp.	PLCC for Planar motion comp.	PLCC for in-depth motion comp.
$VD_{average}$	0.29	0.28	0.54	0.38
$VD_{maximum}$	0.21	0.42	0.26	0.38
$VD_{averagepost}$	0.29	0.28	0.54	0.38

The correlation coefficient increases when considering groups of sequences based on their features ("static", "planar motion", "in-depth motion"). Indeed, PLCC of $VD_{average}$ is 0.29 when considering the whole set of sequences. Its score is 0.54 when considering the "planar motion" class. PLCC score of $VD_{maximum}$ is 0.21 when considering the whole set of sequences. Its score is 0.42 when considering the "static" class. PLCC score of $VD_{averagepost}$ is 0.29 when considering the whole set of sequences. Its score is 0.54 when considering the "planar motion" class. This highlights the fact that the final visual discomfort score should be applied depending on the characteristics of the tested sequences. In other words, the pooling method for outputting the final visual discomfort score should rely on the detection of dominant features in the sequence. For future work, we plan to investigate this axis.

2.4 Conclusion and future work

This section presented the implemented model for objectively measuring visual discomfort in stereoscopic 3D. This work is based on previous studies [22, 23, 24, 23].

The basic principles of the model has been described. The structure of the implemented model, developed with Matlab, has been presented. The validation process showed encouraging results concerning the performances of the implemented model. In particular, the results showed the necessity to consider different pooling methods depending on the main features of the sequence, in terms of nature of the detected object motion.

3 Saliency detection for stereoscopic images

3.1 Introduction

Visual attention is an important characteristic in the Human Visual System (HVS) for visual information processing. With large amount of visual information, visual attention would selectively process the important visual information by filtering out others to reduce the complexity for scene analysis. These important visual information is also termed as salient regions or Regions of Interest (ROIs) in natural images. There are two different approaches for visual attention mechanism: bottom-up and top-down. Bottom-up approach, which is data-driven and task-independent, is a perception process for automatic salient region selection for natural scenes [16][21], while top-down approach is a task-dependent cognitive processing affected by the performed tasks, feature distribution of targets, and so on [38][9][45].

Over the past decades, many studies have tried to propose computational models of visual attention for various multimedia processing applications, such as visual retargeting [8], visual quality assessment [25], visual coding [12], etc. According to the Feature Integration Theory (FIT) [39], the early selective attention causes some image regions to be salient due to their different features (such as color, intensity, texture, depth, etc.) from their surrounding regions. Based on the FIT, many bottom-up saliency detection models have been proposed for 2D images/videos recently [16][21][3].

Note that the computational models of visual attention might focus on predicting sequences of gaze shifts and/or saliency maps. In our work, we limit ourselves to models that can compute saliency maps representing the level of bottom-up visual interest of each area in the visual scene (or each pixel in an image). Therefore, these models are also referred to as "visual saliency model".

Itti *et al.* proposed one of the earliest computational saliency detection model based on the neuronal architecture of the primates' early visual system
citeitti1998model. Bruce *et al.* designed a saliency detection algorithm based on information maximization [3]. Le Meur *et al.* proposed a computational model of visual attention based on characteristics of the HVS including contrast sensitivity

functions, perceptual decomposition, visual masking, and center-surround interactions [21]. Hou *et al.* proposed a saliency detection method by a concept of Spectral Residual [14]. The saliency map is computed by the log spectra representation of the image calculated from Fourier Transform. Based on Hou's model, Guo *et al.* designed a saliency detection algorithm based on phase spectrum, in which the saliency map is calculated by Inverse Fourier Transform on a constant amplitude spectrum and the original phase spectrum [12]. Recently, many saliency detection models have been proposed by patch-based contrast and obtain promising performance for salient region extraction [11][8][10]. In [10], a context-based saliency detection model is proposed based on patch-contrast from color and intensity features. Fang *et al.* introduced a saliency detection model in compressed domain for the application of image retargeting [8].

Recently, there are various emerging applications with the development of stereoscopic display [15]. Compared with saliency detection for 2D images, the depth cue has to to be taken into account in saliency detection for 3D images. Currently, there are several studies exploiting the 3D saliency detection [48]. Zhang *et al.* designed a stereoscopic visual attention algorithm for 3D video based on multiple perceptual stimuli [48]. Chamaret *et al.* built one Region of Interest (ROI) extraction method for adaptive 3D rendering [5]. Both studies [48] and [5] adopt depth map to weight the 2D saliency map to calculate the final saliency map for 3D images. Another method of 3D saliency detection model is built by incorporating depth saliency map into the traditional 2D saliency detection methods. In [27], Ouerhani *et al.* extended a 2D saliency detection model for 3D saliency detection by taking depth cues into account. Potapova *et al.* introduced a 3D saliency detection model for robotics tasks by incorporating the top-down cues into the bottom-up saliency detection [30]. Recently, Wang *et al.* proposed a computational model of visual attention for 3D images [46]. Apart from detecting salient areas based on 2D visual features, the model in [46] takes depth as an additional visual dimension. The measure of depth saliency is derived from the eye movement data obtained from an eye-tracking experiment using synthetic stimuli. Two different ways of integrating depth information in the modeling of 3D visual attention are proposed and examined in [46]. Being different from previous related studies in which no quantitative evaluation is performed by using eye-tracking ground-truth, the study [46] provided a public database with ground-truth of eye-tracking data for the performance evaluation. The results demonstrate a good performance of the model in [46], as compared to that of state-of-the-art 2D models on 2D images. The results also suggest that a better performance is obtained when depth information is taken into account through the creation of a depth saliency map rather than when it is integrated by a weighting method.

From the above description, the key of the 3D saliency detection model is how to adopt the depth cue besides the traditional 2D low-level features such as color, intensity, orientation, etc. Previous studies from neuroscience indicate that the depth feature would cause human beings' attention focusing on the salient regions as well as other low-level features such as color, intensity, motion, etc. [39][47]. Therefore, an accurate

3D saliency detection model should take depth contrast into account as well as contrast from other common low-level features.

In this work, we propose a novel saliency detection framework based on the feature contrast from color, luminance, texture, and depth. The proposed model is basically built on the energy contrast between image patches, which is used to represent the center-surround differences for image patches. It is well accepted that the DCT (Discrete Cosine Transform) is a superior representation for energy compaction and most of the signal information is concentrated on a few low-frequency components of the DCT [19]. Due to its energy compactness property, the DCT have been widely used in various signal processing applications in the past decades. In the proposed model, the input image and depth map are firstly divided into small image patches. Color, luminance and texture features are extracted based on DCT coefficients for each image patch in the original image, while depth feature is extracted based on DCT coefficients for each image patch in the depth map. The feature contrast is calculated based on the center-surround feature differences between image patches, weighted by a Gaussian model of spatial distances for the consideration of local and global contrast. Based on the compactness property of feature maps, a new fusion method is designed to fuse the feature maps to get the final saliency map for 3D images. Experimental results on the eye-tracking database demonstrate the much better performance of the proposed model compared with other existing ones.

The rest of this study is organized as follows. In Section 3.2, the proposed model is introduced in detail. Section 3.3 provides the experimental results between the proposed method with other existing ones. The final section gives the discussion and conclusion for the study.

3.2 The Proposed Model

In the proposed model, we calculate the saliency map based on the patch-based energy contrast from color, luminance, texture and depth features. In this section, we firstly introduce the feature extraction for the proposed model. Then the feature map calculation is described. In the final subsection, we present the new fusion method on how to combine feature maps to calculate the final saliency map for the 3D image.

3.2.1 Feature Extraction

In this study, the input image is divided into small image patches and then the DCT coefficients are adopted to represent the energy for each image patch. The input RGB image is firstly converted to YCbCr color space due to its perceptual property. In YCbCr color space, the Y component represents the luminance information, while

Cb and Cr are two color components. In the DCT, DC coefficients represent the average energy over all pixels in the image patch, while AC coefficients represent the detailed frequency properties for the image patch. Thus, we use the DC coefficient of Y component to represent the luminance feature for the image patch as $L = Y_{DC}$ (Y_{DC} is the DC coefficient of Y component), while the DC coefficients of Cb and Cr components are adopted to represent the color feature as $C_1 = Cb_{DC}$ and $C_2 = Cr_{DC}$ (Cb_{DC} and Cr_{DC} are the DC coefficients from Cb and Cr components respectively).

Since the Cr and Cb components mainly include the color information, we use the AC coefficients from only the Y component to represent the texture feature for the image patch. The existing study in [36] has demonstrated that the first 9 low-frequency AC coefficients in zig-zag scanning can represent most energy for the detailed frequency information in one 8×8 image patch. Based on the study of [36], we use the first 9 low-frequency AC coefficients to represent the texture feature for each image patch as $T = \{Y_{AC1}, Y_{AC2}, ..., Y_{AC9}\}$.

For the depth feature, we firstly calculate the perceived depth information based on the disparity. The depth map M for the image pair can be calculated as introduced in our previous study [46]:

$$M = V/(1 + \frac{d \cdot H}{P \cdot W}) \tag{11}$$

where V represents the viewing distance of the observer; d denotes the interocular distance; P is the disparity between pixels; W and H represent the width (in cm) and horizontal resolution of the display screen, respectively.

Similar with feature extraction for color and luminance, we adopt the DC coefficient from image patches in depth map in Eq. (11) as $D = M_{DC}$ (M_{DC} represents the DC coefficient for the image patch in depth map M).

As described above, we can extract five features of color, luminance, texture and depth (L, C_1, C_2, T, D) for the input stereoscopic image. We will introduce how to calculate the feature map based on these extracted features in the next subsection.

3.2.2 Feature Map Calculation

According to the Feature Integration Theory [39], the salient regions in visual scenes pop out due to their feature contrast from their surrounding regions. Thus, the direct method to extract salient regions in visual scenes is to compute the feature contrast between image patches and their surrounding patches in the visual scene. In this study, we estimate the saliency value for each image patch based on the feature contrast between this image path and all the other patches in the image. Moreover, the distance between the image patch to each of its surrounding patches are also taken into account.

It is well accepted that the HVS is highly space-variant due to the different densities of cone photoreceptor cells in the retina [43]. The visual acuity decreases with the increasing eccentricity from the fixation region, which means that the HVS is more sensitive to the feature contrast from nearer neighborhood patches compared with that from farther neighborhood patches. Thus, we take this property of the HVS into consideration during the saliency estimation. Due to the generality of the Gaussian model, we use a Gaussian model of spatial distance between image patches to weight the feature contrast for feature map calculation. Therefore, the saliency value F_i^k of the image patch i in the feature k can be computed as:

$$F_i^k = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{l_{ij}^2}{2\sigma^2}} U_{ij}^k \tag{12}$$

where k represents the feature and $k \in \{L, C_1, C_2, T, D\}$; l_{ij} denotes the spatial distance between image patches i and j; U_{ij}^k represents the feature difference between image patches i and j from feature k; σ is the parameter for the Gaussian model and it determines the degree of local and global contrast for the saliency estimation. From this equation, we see that the saliency value of each image patch is calculated based on feature contrast from all other image patches. Due to the different weighting values for image patches from different spatial distances, the proposed model considers both local and global contrast for saliency estimation.

Since the color, luminance and depth features are represented by DC coefficients, the feature contrast between two image patches i and j can be calculated as the difference between DC coefficients as follows.

$$U_{ij}^m = \frac{B_i^m - B_j^m}{B_i^m + B_j^m} \tag{13}$$

where B^m represents the feature and $B^m \in \{L, C_1, C_2, D\}$;

The texture feature is represented as 9 low-frequency AC coefficients and we calculate the feature contrast from texture $U_{ij}^{'}$ between two image patches i and j as:

$$U_{ij}^{'} = \frac{\sqrt{\sum_t (B_i^{'t} - B_j^{'t})^2}}{\sum_t (B_i^{'t} + B_j^{'t})} \tag{14}$$

where t represents the AC coefficients and $t \in \{1, 2, ..., 9\}$; $B^{'}$ represents the texture feature.

3.2.3 Feature Map Fusion

After obtaining feature maps indicated in Eq. (12), we fuse these feature maps from color, luminance, texture and depth to compute the final saliency map. Most existing

studies of 3D saliency detection (*e.g.* [46]) use simple linear combination to fuse the feature maps to obtain the final saliency map. The weighting for the linear combination is set as constant values and is the same for all images. In this study, we propose a new method to assign adaptive weighting for the fusion of feature maps.

Generally, the salient regions in a good saliency map should be small and compact, since the HVS always focus on some specific interesting regions in images. Therefore, a good feature map should represent small and compact salient regions in the image. During the fusion of different feature maps, we can assign more weighting for those feature maps with small and compact salient regions and less weighting for others with more spread salient regions. Here, we define the measure of compactness by the spatial variance of feature maps. The spatial variance υ_k of feature map F_k can be computed as follows.

$$\upsilon_k = \frac{\sum_{(i,j)} \sqrt{(i - E_{i,k})^2 + (j - E_{j,k})^2} \cdot F_k(i,j)}{\sum_{(i,j)} F_k(i,j)} \tag{15}$$

where (i, j) is the spatial location in the feature map;k represents the feature channel and $k \in \{L, C_1, C_2, T, D\}$; $(E_{i,k}, E_{j,k})$ are the spatial expectation location of the salient regions which are calculated as:

$$E_{i,k} = \frac{\sum_{(i,j)} i \cdot F_k(i,j)}{\sum_{(i,j)} F_k(i,j)} \tag{16}$$

$$E_{j,k} = \frac{\sum_{(i,j)} j \cdot F_k(i,j)}{\sum_{(i,j)} F_k(i,j)} \tag{17}$$

We use the normalized υ_k values to represent the compactness property for feature maps. With larger spatial variance values, the feature map is supposed to be less compact. We calculate the compactness β_k of the feature map F_k as follows.

$$\beta_k = 1/(e^{\upsilon_k}) \tag{18}$$

where k represents the feature channel and $k \in \{L, C_1, C_2, T, D\}$.

Based on spatial variance values of feature maps calculated in Eq. (18), we fuse the feature maps for the final saliency map as follows.

$$S = \sum_k \beta_k \cdot F_k + \sum_{p \neq q} \beta_p \cdot \beta_q \cdot F_p \cdot F_q \tag{19}$$

The first term in Eq. (19) represents the linear combination of feature maps weighted by corresponding compactness; while the second term is adopted to enhance the common salient regions which can be detected by any two different feature maps. Different

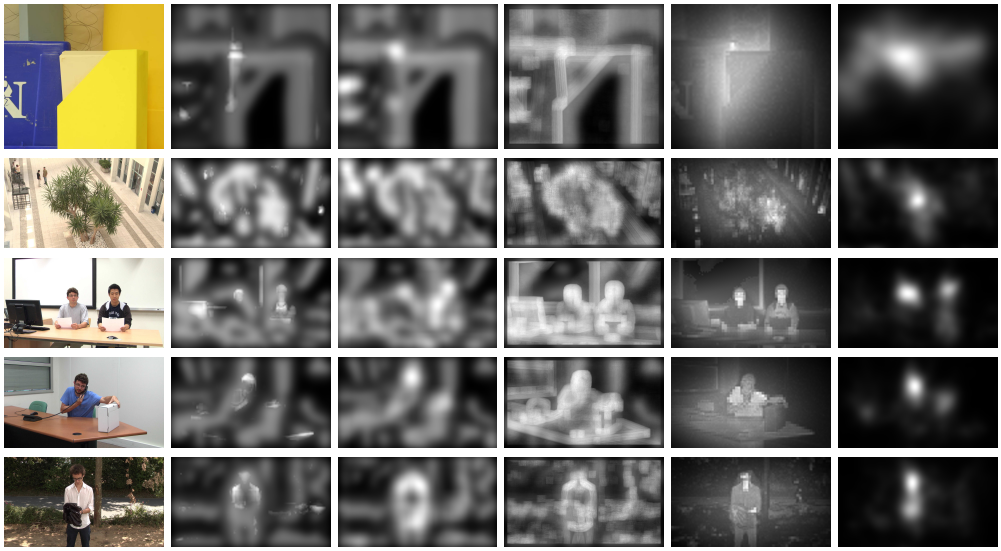


Figure 1: Visual comparison of stereoscopic saliency detection models. Column 1: original left images; Columns 2 - 6: saliency maps by Model 1 in [46], Model 2 in [46], Model 3 in [46], the proposed model and the ground truth, respectively.

from existing studies using the constant weighting values for different images, the proposed fusion method assign different weighting values for different images based on their compactness property.

3.3 Experimental Evaluation

In this section, we evaluate the performance of the proposed model based on the eye tracking database [44] proposed in our previous study [46]. To the best of our knowledge, this is the only available eye tracking database for 3D images in the research community. The ground-truth maps in this database are represented as fixation density maps generated from the data recorded by a SMI RED remote eye-tracker. This database includes various types of stereoscopic images such as outdoor scenes, indoor scenes, scenes including objects, scenes without any various object, etc. Some samples of the left images and corresponding ground-truth maps are shown in the first and last columns of Fig. 1, respectively.

In this experiment, we use the similar measure methods as the study [46] to evaluate the performance of the proposed method. The performance of the proposed model is measured by comparing the ground-truth map and the saliency map from the saliency detection model. As there are left and right images for any stereoscopic image pair, we

Table 1: Comparison results of PLCC, KLD and AUC values from different stereoscopic saliency detection models.

Models	PLCC	KLD	AUC
Model 1 in [46]	0.356	0.704	0.656
Model 2 in [46]	0.424	0.617	0.675
Model 3 in [46]	0.410	0.605	0.670
The Proposed Model	0.5499	0.3589	0.7032

use the saliency result of the left image to do the comparison, similar with the study [46]. The PLCC (Pearson Linear Correlation Coefficient), KLD (Kullback-Leibler Divergence), and AUC (Area Under the Receiver Operating Characteristics Curve) are used to evaluate the quantitative performance of the proposed stereoscopic saliency detection model. Among these measures, PLCC and KLD are calculated directly from the comparison between the fixation density map and the predicted saliency map, while AUC is computed from the comparison between the actual gaze points and the predicted saliency map. With larger PLCC and AUC values, the saliency detection model can predict more accurate salient regions for 3D images. In contrast, the performance of the saliency detection model is better with the smaller KLD value between the fixation map and saliency map.

The quantitative comparison results are given in Table 1. In Table 1, Model 1 in [46] represents the fusion method from 2D saliency detection model in [16] and depth model in [46]; Model 2 in [46] represents the fusion method from 2D saliency detection model in [3] and depth model in [46]; Model 3 represents the fusion method from 2D saliency detection model in [14] and depth model in [46]. From this table, we can see that the PLCC and AUC values from the proposed model is larger than those from models in [46], while KLD value from the proposed model is lower than those from models in [46]. This means that the proposed model can estimate more accurate saliency maps compared with other models in [46].

To better demonstrate the advantages of the proposed model, we provide some visual comparison samples from different models in Fig. 1. From the second column of this figure, we can see that the stereoscopic saliency maps from the fusion model by combining Itti's model [16] and depth saliency [46] mainly detect the contour of salient regions in images. The reason for this is that the 2D saliency detection model in [16] calculates saliency map mainly by local contrast. Similarly, there is the same drawback for the saliency maps from the third column of Fig. 1. For the saliency results from the fusion model by combing 2D saliency model in [3] and depth saliency in [46], some background regions are detected as salient regions in images, as shown in saliency maps from the fourth column of Fig. 1. In contrast, the saliency results from the proposed stereoscopic saliency detection model can estimate much more accurate salient regions with regard to the ground truth map from eye tracking data, as shown in Fig. 1.

3.4 Discussion and Conclusion

As demonstrated in the experimental part, the proposed model can obtain much better performance than other existing ones in saliency estimation for 3D images. The superior performance might be caused by top-down cue besides bottom-up mechanism. The ground-truth maps used in this study were collected based on the fixation data during 15 seconds, and they include the fixations resulting from both bottom-up and top-down mechanisms [45]. Since the proposed algorithm is a patch-based saliency detection method and it can detect the ROIs including the complete salient objects in 3D images (as shown in the experimental results), the top-down mechanism might be included in the proposed method. In contrast, the existing models in [46] which incorporate the 2D saliency methods [16, 3, 14] are designed for only bottom-up mechanism. Therefore, the proposed method can obtain much better performance than the ones in [46] for saliency estimation of 3D images.

Overall, we propose a new stereoscopic saliency detection model for 3D images in this study. The features of color, luminance, texture and depth are extracted from DCT coefficients to represent the energy for small image patches. The saliency is estimated based on the energy contrast weighted by a Gaussian model of spatial distances between image patches for the consideration of both local and global contrast. A new fusion method is designed to combine the feature maps for the final saliency map. Experimental results show the promising performance of the proposed saliency detection model for stereoscopic images based on a recent eye tracking database.

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 [17]. 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 [26]. 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.[17])

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 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 2

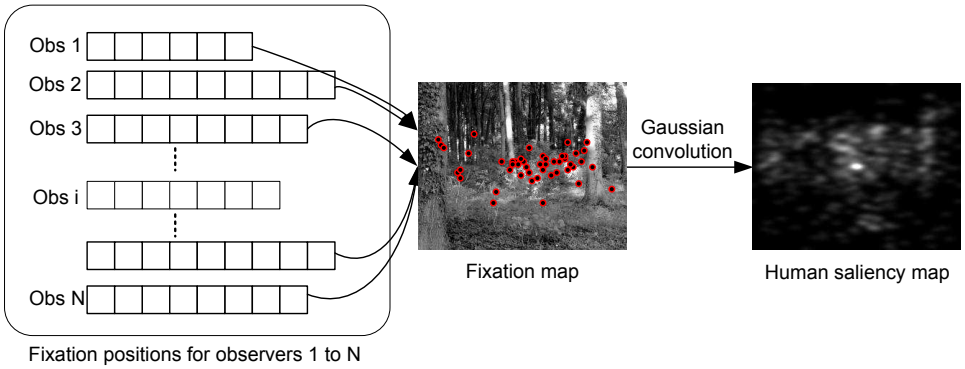


Figure 2: Illustration of the human saliency map computation from N observers

4.2 Behavioral and computational studies

Jansen et al. [17] 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 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 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 3 shows AUC values in function of the fixation rank. Over the whole viewing time (called “All” on the right-hand side of Figure 3), 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

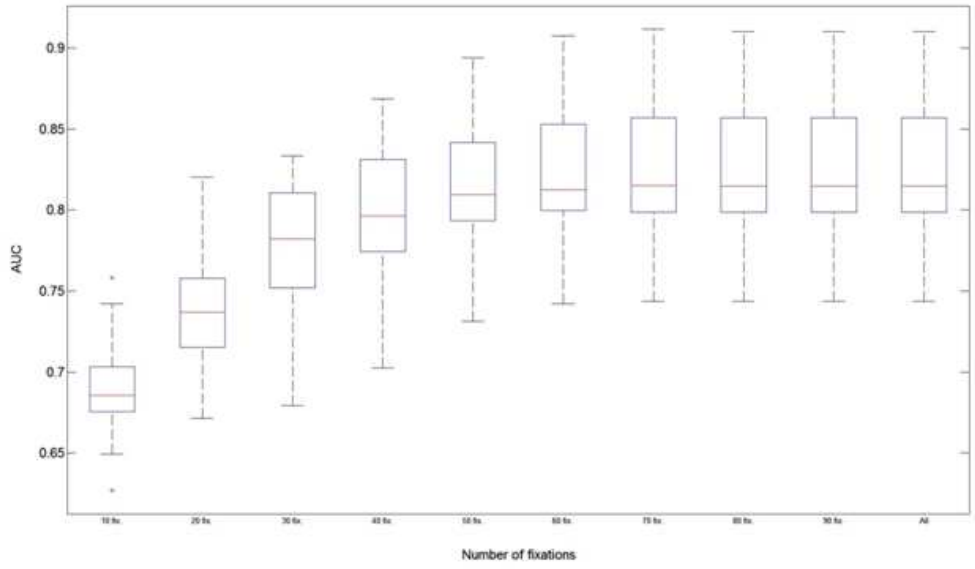


Figure 3: Boxplot of the AUC values between 2D and 3D human (experimental) saliency maps as a function of the number of fixations (the top 20% 2D salient areas are kept).

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. [17], 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 [34]). Recently, Bindemann [2] 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.

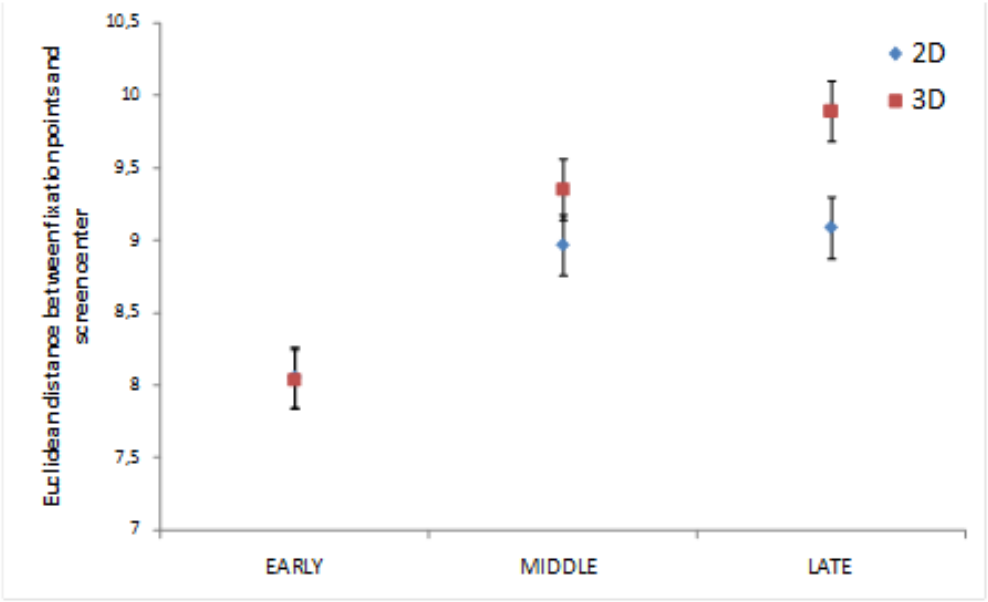


Figure 4: Average Euclidean distance between the screen center and fixation points. The error bars correspond to SEM (Standard Error of the Mean).

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 5 and Figure 6). 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«0.001, 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.

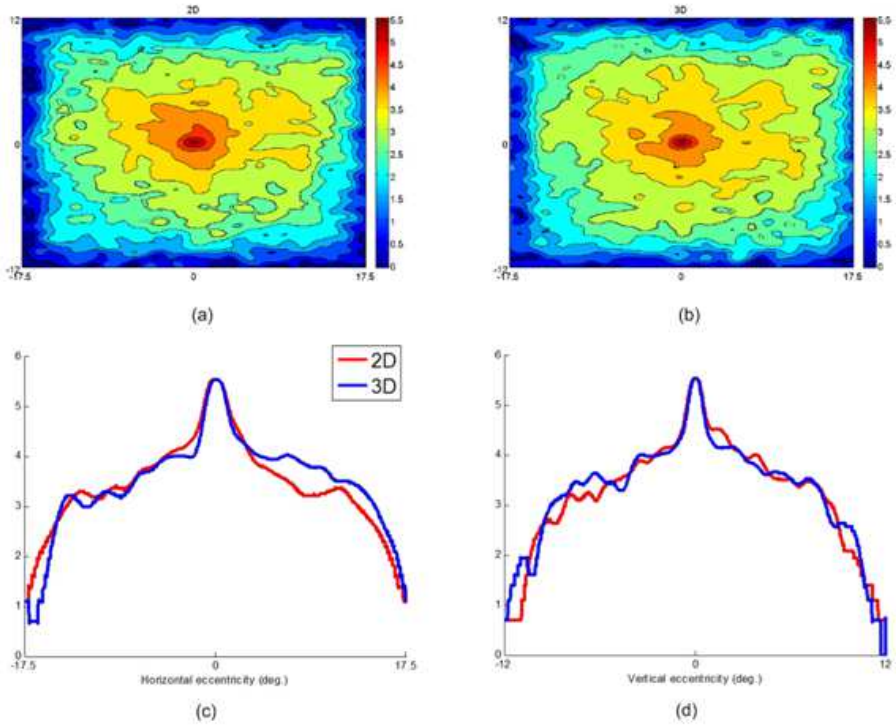


Figure 5: (a) and (b) are the distributions of fixations for 2D and 3D condition, respectively. (c) and (d) represent the horizontal and vertical cross sections through the distribution shown in (a) and (b). All the visual fixations are used to compute the distribution.

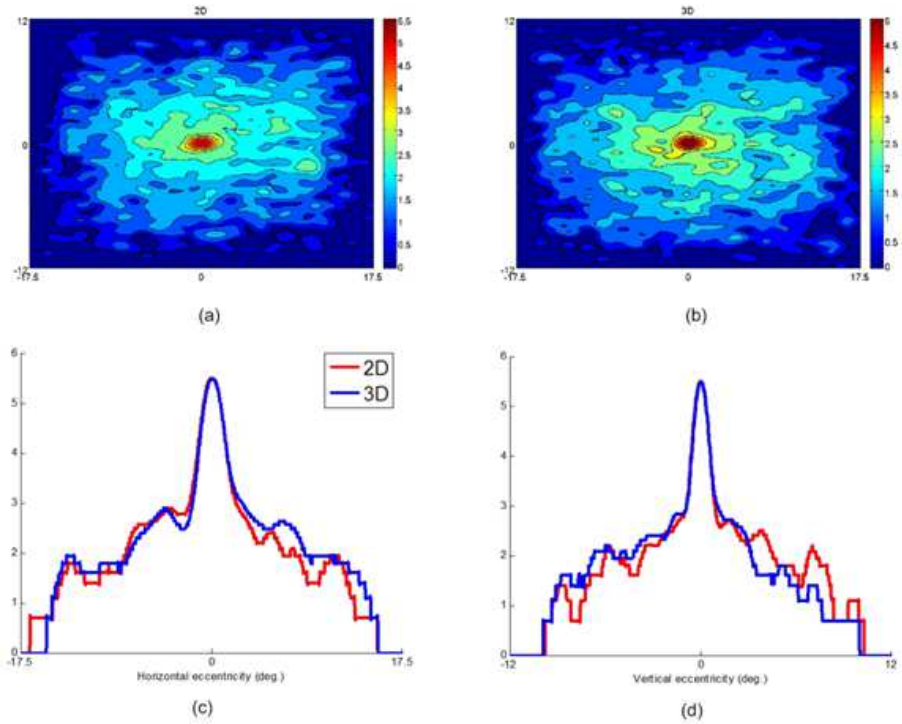


Figure 6: (a) and (b) are the distributions of fixations for 2D and 3D condition, respectively. (c) and (d) represent the horizontal and vertical cross sections through the distribution shown in (a) and (b). All the visual fixations are used to compute the distribution.

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

In [17], 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 7 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 7: 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 8.

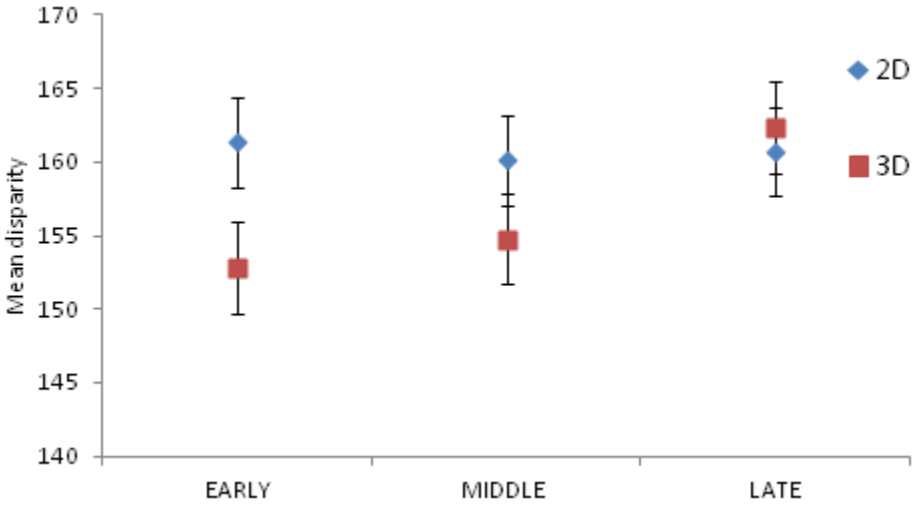


Figure 8: Mean disparity (in pixels) in function of the viewing time (early, middle and late). The error bars correspond to SEM (Standard Error of the Mean).

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. [17], and support the existence of a depth bias.

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

5.1 Introduction

Recent studies [34], [50] 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 [29],[35],[40],[6]. But considering their interactions and impacts over time might be crucial to improve the predictability of existing saliency models [34], [35].

5.2 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 [42]. 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:

$$f(x,t) = \sum_{k=1}^{K} p_k(t)\phi_k(x) \tag{20}$$

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 [7]. Before explaining this method, we describe the center and depth modeling.

5.3 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 [2]. However, the extent to which this potential laboratory artifact is an inherent feature of strategy of human vision remains an open subject. Tatler [34] 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 [34], 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 [49]. As proposed in [13], 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 σ_x^2 and σ_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 ($\sigma_{x2D} = 4.7°, \sigma_{y2D} = 2.5°, \sigma_{x3D} = 4.3°, \sigma_{y3D} = 2.3°$)

5.4 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 [33], that can be understood as an element of the edge interpretation depth cue [28], drives the visual attention pre-attentively [31]. 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 9 (a).

5.5 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 9(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:

$$f(x,t) = p_{sm}(t)\phi_{sm}(x) + p_{cb}(t)\phi_{cb}(x) + p_{fg}(t)\phi_{fg}(x) + p_{bg}(t)\phi_{bg}(x) + p_{un}(t)\phi_{un}(x) \quad (21)$$

with ϕ_{sm} the saliency maps of one of the 3 models, ϕ_{cb} the central Gaussian function, ϕ_{fg} and ϕ_{bg} the foreground and background map respectively and ϕ_{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 9(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.

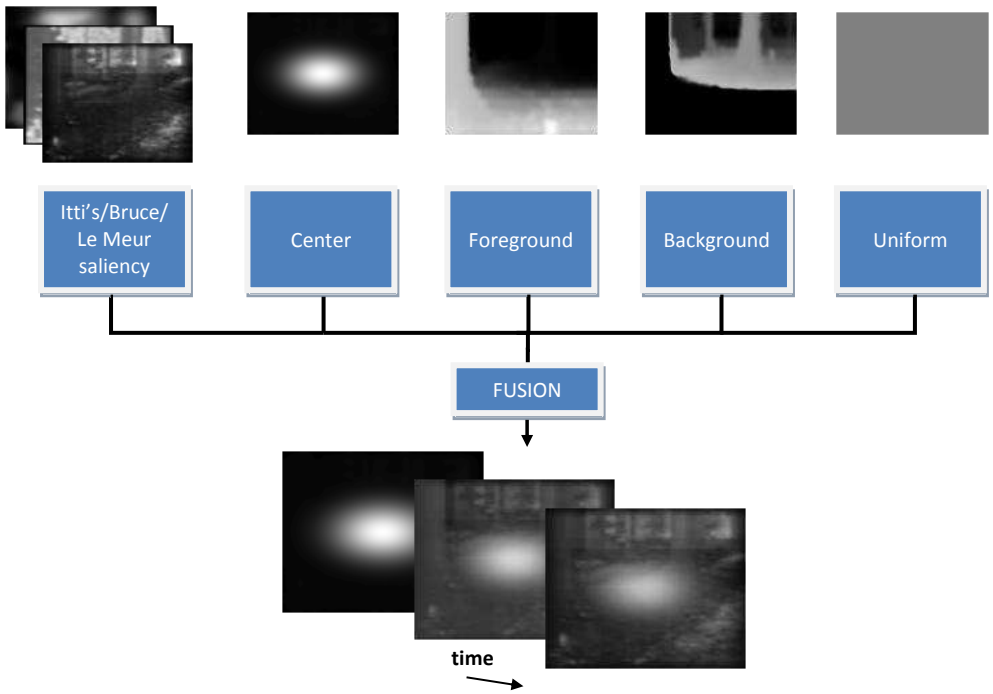


Figure 9: (a) Upper Row: Illustration of Itti's saliency map obtained from an image, the center bias in 2D condition,the corresponding foreground and background feature maps. (b) Middle row: Description of the proposed time-dependent model. (c) Lower Row: Illustration of the resulting time-dependent 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:
      $p_k^{(m)}(t) = \frac{\sum_{i=1}^{N} t_{i,k}^{(m)}}{N}$

5.6 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 5.6:

The best predictor for both viewing conditions is the predicted saliency map (from Itti's model and called Sm on Figure 5.6). 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.

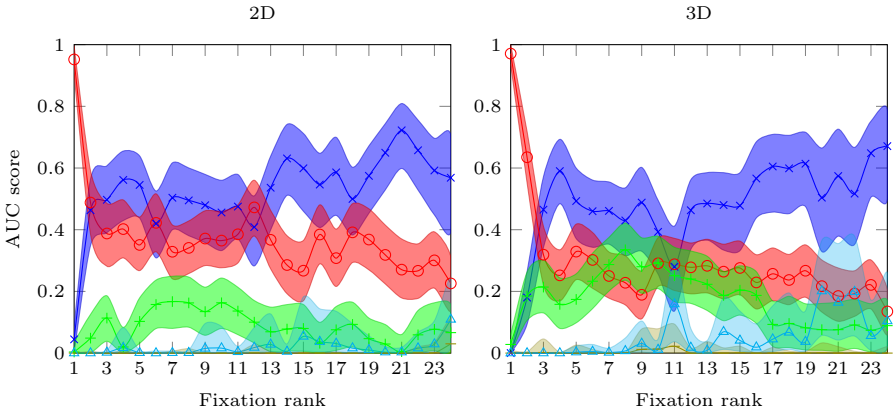


Figure 10: Temporal contributions (weights) of 5 features on 2D (left) and 3D (right) fixations to eye movements as a function of the fixation rank. Low-level saliency feature ("Sm") here comes from Itti's model. The error areas at 95% are computed by a "bootstrap" estimate (1000 replications).

5.7 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 [34] and in accordance with [13], 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 [34]. 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.

5.8 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 (21) 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 (21)) 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 11 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.

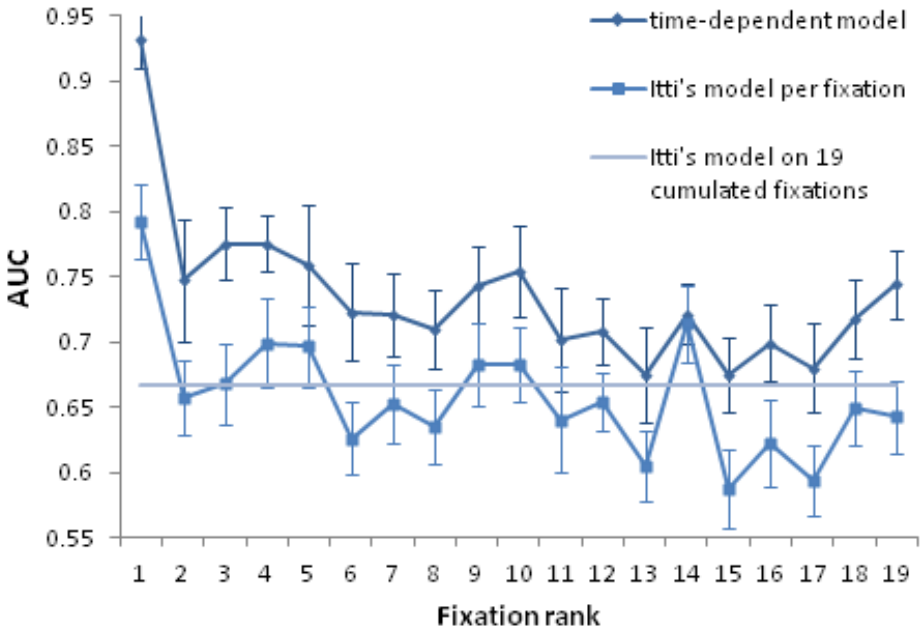


Figure 11: Temporal evolution of the performance of the time-dependent model based on Itti's, versus the Itti's model per fixation, and versus the Itti's model on 19 cumulated fixations.

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 12 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.

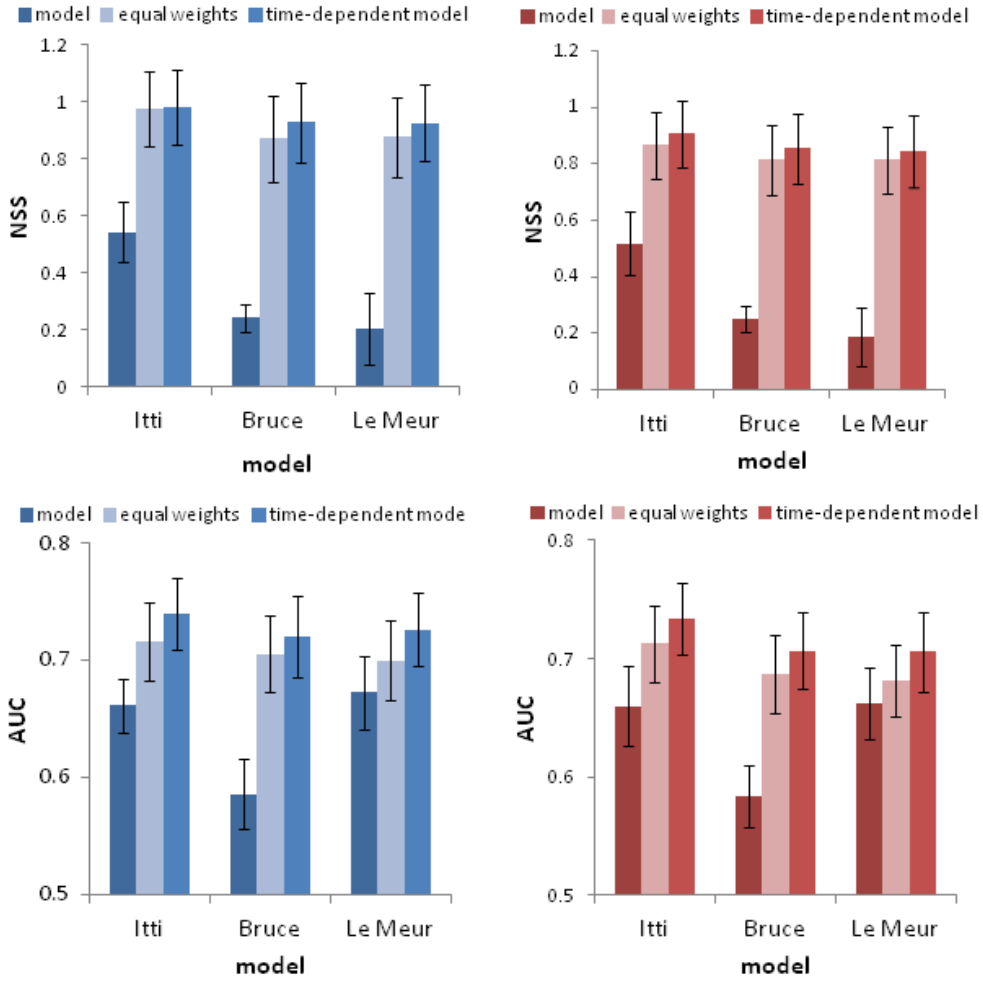


Figure 12: Comparison between 6 saliency models in 2D (left) and 3D conditions (right). Upper row: the NSS criterion, lower row the AUC criterion. The error bars correspond to the SEM. NS corresponds to Non-Significant. When the term NS is not indicated, results are significantly different ($p<0.05$).

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 12). 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.

5.8.1 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 [18]. 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 [20].

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.

5.9 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 [31]. 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 [13] in the case of low-level feature combination, this hypothesis is very simple with regards to the complexity of visual attention construction[41], and with regards to other computational proposals of fusion [4]. 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] Kaoru Amano, Naokazu Goda, Shin'ya Nishida, Yoshimichi Ejima, Tsunehiro Takeda, and Yoshio Ohtani. Estimation of the timing of human visual perception from magnetoencephalography. *The Journal of neuroscience*, 26(15):3981–3991, 2006.

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

[3] N.D.B. Bruce and J.K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009.

[4] 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.

[5] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur. Adaptive 3d rendering based on region-of-interest. In *Proceedings of SPIE*, volume 7524, page 75240V, 2010.

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

[7] 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.

[8] Yuming Fang, Zhenzhong Chen, Weisi Lin, and C-W Lin. Saliency detection in the compressed domain for adaptive image retargeting. *Image Processing, IEEE Transactions on*, 21(9):3888–3901, 2012.

[9] Yuming Fang, Weisi Lin, Chiew Tong Lau, and Bu-Sung Lee. A visual attention model combining top-down and bottom-up mechanisms for salient object detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1293–1296. IEEE, 2011.

[10] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012.

[11] Viswanath Gopalakrishnan, Yiqun Hu, and Deepu Rajan. Salient region detection by modeling distributions of color and orientation. *Multimedia, IEEE Transactions on*, 11(5):892–905, 2009.

[12] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Image Processing, IEEE Transactions on*, 19(1):185–198, 2010.

[13] 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.

[14] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. Ieee, 2007.

[15] Q. Huynh-Thu, M. Barkowsky, P. Le Callet, et al. The importance of visual attention in improving the 3d-tv viewing experience: Overview and new perspectives. *IEEE Transactions on Broadcasting*, 57(2):421–431, 2011.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[17] 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.

[18] 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.

[19] Syed Ali Khayam. The discrete cosine transform (dct): theory and application. *Michigan State University*, 2003.

[20] 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.

[21] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):802–817, 2006.

[22] Jing Li, Marcus Barkowsky, and Patrick Le Callet. The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos. In *Proceeding of the International Workshop on Quality of Multimedia Experience QoMEX*, pages pp.1–6, Mechelen, Belgique, September 2011.

[23] Jing Li, Marcus Barkowsky, and Patrick Le Callet. VISUAL DISCOMFORT IS NOT ALWAYS PROPORTIONAL TO EYE BLINKING RATE: EXPLORING SOME EFFECTS OF PLANAR AND IN-DEPTH MOTION ON 3DTV QOE. In *Proceedings of VPQM 2013*, pages pp.1–6, Scottsdale, USA, 2013.

[24] Jing Li, Marcus Barkowsky, Junle Wang, and Patrick Le Callet. Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses. In *Proceeding of 17th International Conference on Digital Signal Processing 2011*, page 1, Corfu, Greece, 2011.

[25] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[26] 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.

[27] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375–378. IEEE, 2000.

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

[29] 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.

[30] E. Potapova, M. Zillich, and M. Vincze. Learning what matters: combining probabilistic models of 2d and 3d saliency cues. *Computer Vision Systems*, pages 132–142, 2011.

[31] 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.

[32] TW Ridler and S Calvard. Picture thresholding using an iterative selection method. *IEEE transactions on Systems, Man and Cybernetics*, 8(8):630,Äì632, 1978.

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

[34] 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.

[35] 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.

[36] Ch Theoharatos, Vasileios K Pothos, Nikolaos A Laskaris, George Economou, and Spiros Fotopoulos. Multivariate image similarity in the compressed domain using statistical graph matching. *Pattern Recognition*, 39(10):1892–1904, 2006.

[37] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520,Äì522, 1996.

[38] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[39] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

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

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

[42] 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.

[43] B.A. Wandell. *Foundations of vision*, volume 21. Sinauer Associates, 1995.

[44] J. Wang, M. Perreira Da Silva, P. Le Callet, and V. Ricordel. IRCCyN/IVC 3DGaze database. http://www.irccyn.ec-nantes.fr/spip.php?article1102&lang=en, 2011.

[45] Junle Wang, Damon M. Chandler, and Patrick Le Callet. Quantifying the relationship between visual salience and visual importance. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging XV - part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 18-21, 2010, Proceedings*, volume 7527 of *SPIE Proceedings*, page 75270. SPIE, 2010.

[46] Junle Wang, M.P. DaSilva, P. LeCallet, and V. Ricordel. A computational model of stereoscopic 3d visual saliency. *Image Processing, IEEE Transactions on*, 22(6):2151–2165, 2013.

[47] J.M. Wolfe and T.S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[48] Y. Zhang, G. Jiang, M. Yu, and K. Chen. Stereoscopic visual attention model for 3d video. *Advances in Multimedia Modeling*, pages 314–324, 2010.

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

[50] 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.